**PROJECT ID:** CAFS.20.79

**YEAR:**  **3** 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, [kasey.legaard@maine.edu](mailto: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 will 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:**  We have continued to develop fully and semi-automated machine learning (ML) and forest mapping workflows based on multispectral and 3D vegetation measurement data, using FIA plot data for model training and validation. Our methods leverage multi-objective ML algorithms to control or minimize systematic prediction error that commonly arises when associating remote sensing data with FIA plot measurements using established ML or statistical modeling techniques. Developments over the past year have focused on improved data handling algorithms and on more computationally efficient workflows to better support large-area projects, most of which currently rely on either Sentinel-2 imagery processed at 10 m resolution or point cloud data processed to 10 m resolution gridded metrics. Recent changes to our satellite image processing workflows integrate improvements to automated local image coregistration, haze and cirrus correction, cloud/shadow detection and masking, topographic illumination correction, forest road detection, and unsupervised disturbance detection to improve ML models. Haze and cirrus correction incorporates significant improvements over published algorithms, which tended to perform poorly with Sentinel-2 data. We have developed an alternative cloud/shadow detection algorithm based on gradient boosted regression trees to improve detection accuracy over standard approaches (i.e., FMask or similar algorithms). Our cloud/shadow masking methods have been implemented by modifying the widely used Python FMask application, and we are currently working in collaboration with students from the Monroe Community College Geospatial Information Science and Technology program to develop and distribute QGIS plugins to facilitate wider application of our methods (as well as Python FMask itself). Once clouds and shadows have been eliminated from image data, we use a multi-stage ML process to reduce the impact of missing data on forest predictions.  We have developed and implemented several procedures for improving species and forest type predictions adjacent to abrupt forest edges. To develop robust relationships between 10 m resolution imagery and FIA plots, we average pixel values within neighborhoods that match the nominal area sampled by FIA subplots. This improves predictions overall, but causes blurring and significant error adjacent to abrupt edges. We currently identify problematic edges through a combination of a novel forest road mapping algorithm and traditional edge detection algorithms. We thereafter eliminate or minimize effects on forest predictions through adaptive neighborhood filtering procedures.  Species and forest type predictions are based on image and plot data collected over a full FIA measurement cycle. Forest cover change during that cycles may therefore dissociate image and measurement data, negatively impacting model training and prediction accuracy. We have therefore integrated automated, unsupervised methods to identify recent change over a specified observation period based on multispectral image transformations that highlight changes in canopy conditions. We have further integrated changes in our ML that minimize the impact of outliers in training data, which often arise from canopy change in the vicinity of FIA plots. We have also made significant progress developing efficient workflows for mapping change using a semi-automated approach based on multi-objective ML that simultaneously minimizes omission and commission error. This approach requires extensive training data, but produces disturbance maps that are both highly accurate (>90% disturbance class accuracy) and unbiased (equal omission and commission error rates). We have developed semi-supervised methods and software to accelerate the collection of reference data to support this disturbance mapping approach. We have merged 10 m disturbance maps into species mapping workflows to mask areas of erroneous or suspect predictions. We are also making steady progress in developing efficient workflows and software to produce unbiased annual time series of disturbance data from either Sentinel or Landsat imagery, mapping areas of approximately 2-10 million acres at a time.  Much of this year’s work has been supported by statewide mapping projects funded over short data production timelines. We are currently developing data for a next-generation, high-resolution land cover and forest type map for the state of Maine, in collaboration with the Maine GeoLibrary and the NOAA Coastal-Change Analysis Program. Under this collaboration, the state will receive a 1 m land cover map conforming to NOAA C-CAP specifications and a 10 m land cover and forest type map that includes C-CAP non-forest classes plus 15 forest type classes derived from our ML and forest mapping workflows. Advances in ML and automated image processing made this project affordable and feasible, and further development over the course of the project will ensure that future updates are cost-effective for the state. Combined outcomes will provide land cover and forest composition data at previously unprecedented thematic, spatial, and temporal resolutions. Statewide land cover and forest type mapping will conclude in the fall of 2023.  In addition to statewide forest type mapping, we have been funded by the Maine Department of Environmental Protection through the Maine GeoLibrary to map aboveground forest biomass and carbon density at 10 m resolution, coordinated with the production of the new 10 m land cover and forest type data. Biomass and carbon mapping will utilize 3D NAIP data, Sentinel-2 imagery, and disturbance data developed from Landsat disturbance time series produced using our multi-objective ML approach. To meet a very compressed timeline, we have developed highly efficient methods for processing large-area point clouds into gridded metrics suitable for biomass modeling. Our approach is based on advanced database technologies that enable extremely fast computations. Our software application can run on either office hardware or on the cloud, where processing can be scaled to an arbitrary degree. We have fully tested our approach and are currently processing data statewide on UMaine’s computing cluster. Statewide biomass and carbon mapping will conclude in the fall of 2023.  We continue to use USFS FIA plot data as a primary source of model validation due to lack of any other sufficiently representative and publically available independent reference data. We have therefore developed software to support a rigorous cross-validation of species models, a process that is complicated by our use of 2-stage ML models, where prediction of species occurrence is followed prediction of species relative abundance. Cross-validation must therefore simultaneously encompass both modeling stages. To benchmark results, we have additionally invested significant effort in the full integration of the Scikit-learn open-source ML library, enabling side-by-side comparisons between our own multi-objective ML algorithms and more established algorithms like random forest, boosted regression trees, or nearest neighbor methods. We are also working to establish pilot studies with both public and private organizations within Maine to evaluate species predictions and derivative forest type or composition maps (e.g., raster or vector maps of the top 3-5 species). Evaluation of certain derivative forest type products, particularly more heavily processed vector products, is difficult to do without the expertise of local landowners and foresters. Finally, our methods and software are in principle fully generalizable to other forest regions. We would like to pursue application and testing outside the state of Maine pending access to either FIA or other plot measurement data sufficient for model training and testing. |
| **DELIVERABLES:**   * Workflows and software to scale species, forest type, disturbance, and biomass mapping at low cost:   + Sentinel-based species and forest type modeling and mapping at 10 m pixel resolution   + 2-stage species modeling (species occurrence followed by species relative abundance) using multi-objective machine learning to control error in both stages   + Multi-stage, multi-model cross-validation techniques   + Benchmarking multi-objective ML against established methods distributed in widely used open-source libraries   + Sentinel- and Landsat-based disturbance and harvest mapping at high accuracy and low bias   + New, extremely efficient methods for large-area point cloud processing * New state of Maine high-resolution land cover and forest type data, scheduled for public release fall 2023 * New state of Maine high-resolution aboveground biomass and carbon density data, scheduled for public release fall 2023 * Monroe Community College (MCC) summer student internships and capstone projects directed at the development of open-source QGIS plugins for cloud and shadow masking * NSF Research Experience for Undergraduates award supporting undergraduate development of advanced spatial data management technologies (with student recruitment and mentoring facilitated through participation in the NSF CAREERS Cyberteam program, project entitled “Remote sensing data management for large-area forest mapping projects”). |
| **MEMBER COMPANY BENEFITS:**  Key member benefits from project activities conducted to date include end-to-end development and large-area testing of multi-objective machine learning and automated geospatial data processing methods developed at the University of Maine. Year 3 outcomes included substantial improvements to established workflows, culminating in the completion and preliminary testing of production-ready software for large-area, high-resolution species and forest type mapping. We have integrated open-source ML libraries to support side-by-side comparisons with other methods, and will benchmark our multi-objective ML methods against multiple methods as we continue to compile results in Maine. Work within Maine has been driven by the opportunity to develop and distribute next generation land cover, forest type, and forest biomass/carbon maps, with support provided by numerous organization including forest industry. State mapping projects on short production timelines have necessitated additional software development to accelerate data production, which will ultimately lead to reduced time and cost associated with future projects. Development has included extremely efficient methods for processing large-area point cloud data derived from either 3D NAIP or LiDAR. All methods and supporting software are generalizable to forests outside of the northeast. Although year 3 work has been dominated by statewide mapping projects and continued process development, we wish to extend processing to test sites in other regions. Our intention is that members ultimately benefit from the development and implementation of methods that reduce the time and cost associated with inventory and mapping. |