

---

# **Forest Type Maps Generated by Machine Learning on Remotely-Sensed Data**

## **Report to NRCS Oregon**

**In partial fulfillment of Grant No. 69-0436-17-036  
NRCS Oregon - USDA**

**Authors:**  
David Diaz  
Sara Loreno

July 2019

---

For questions or comments, please contact

**Sara Loreno**  
**Natural Resources Data Scientist**  
**[sloreno@ecotrust.org](mailto:sloreno@ecotrust.org)**  
**(503)467-0784**

Ecotrust  
721 NW 9<sup>th</sup> Avenue, Suite 200  
Portland, OR 97209

# Ecotrust

## TABLE OF CONTENTS

1. BACKGROUND .....	3
2. OVERVIEW OF FOREST TYPE MAPS .....	5
3. NEXT STEPS TOWARDS AUTOMATED STAND DELINEATION .....	8

## LIST OF TABLES

Table 1: Inventory Plots used for Model Training .....	4
--	---

## LIST OF FIGURES

Figure 1: Areas Mapped (Lidar Coverage) .....	3
Figure 2: Locations of Inventory Plots.....	4
Figure 3: First Generation of Machine Learning Forest Prediction - Size Class.....	5
Figure 4: First Generation of Machine Learning Forest Prediction - Cover Class .....	6
Figure 5: First Generation of Machine Learning Forest Prediction - Species Composition.....	7
Figure 6: Region-merging segmentation algorithm demonstration with canopy height.....	8

# 1. BACKGROUND

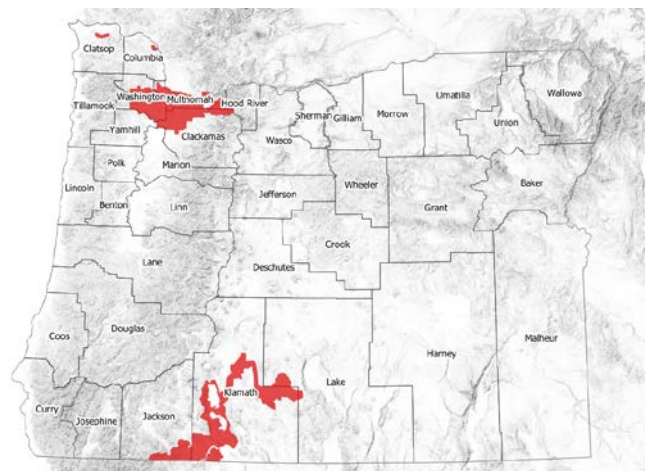
In this report, we summarize work to date generating forest maps using data from satellite and airborne laser scanning (lidar) data sources which complement several thousand forest inventory plots and stand delineations recorded by state and federal agencies across Oregon and Washington. Down-scaled climate data and climate-derived indicators as well as topographic and other environmental layers are also incorporated to inform estimates of forest conditions across the landscape.

In this project, Machine Learning (ML) algorithms are trained to perform several predictive modeling tasks. In ML terminology, we predict “labels” (i.e., forest type attributes) based on numerous “features” (i.e., predictor variables derived from remotely-sensed data). Most of the landscape is covered by publicly-available remote sensing data from which our features for predictive modeling are derived. However, only a limited portion of the landscape has corresponding ground-truth, or “labels” that describe current forest conditions using the terms (dominant species, stocking level, size class) commonly required for completion of Forest Management Plans (FMPs) in the State of Oregon.

We train the predictive models using data from a portion of the landscape where both features and labels are available, holding out some of these data from model training to allow validation of model performance before applying these algorithms more broadly across the landscape.

This report presents the forest type maps produced by our first iteration of ML models. These models have been trained without significant fine-tuning to improve performance and accuracy, which we will be performing in coming months along with evaluation of ground-truth data collected by landowners collaborating in this CIG project.

**Figure 1: Areas Mapped (Lidar Coverage)**



## 1.1. Machine Learning Models

All predictive modeling is performed using open-source tools and software. Documentation of the computing environment as well as a series of Jupyter Notebooks which illustrate the data processing and predictive modeling pipeline are published on GitHub<sup>1</sup>.

The h2o Python package was used to build predictive models. Both Random Forests and Gradient Boosting Machines algorithms have been trained and the current best-fit models have also been uploaded to GitHub and can be downloaded and re-used if the array of lidar, imagery, climate, and other predictor variables are available. The maps shown here were produced from predictions by the Gradient Boosting Machines algorithm, which performed slightly better than Random Forests across several different parameterizations we experimented with. Over the coming months, we will conduct more rigorous model tuning and cross-validation, as well as incorporation of more imagery to help improve predictions of species composition.

<sup>1</sup> <https://github.com/ECOTRUST/ForestMapping>

## 1.2. Modeling Task Formulated

These ML models have been trained to predict forest attributes at the plot-scale, using 20m-x-20m pixels which correspond to an approximately 1/10<sup>th</sup>-acre resolution.

Data from over 5,000 plots collected by state and federal agencies have been gathered and mapped. These plots cover most regions of Oregon and Washington, although northwestern Oregon is poorly represented. Nevertheless, the sampling of both west-side and east-side forest conditions is relatively robust. As far as we are aware, this is the largest combined dataset of plots in our region that have been explicitly established for the purpose of training forest mapping models using remote sensing data. Each agency seems poised to use only their own plot data for their respective forest mapping efforts, and no other efforts that we are aware of have been made to integrate these data into a larger and more comprehensive dataset for agency or other use.

The ML models are currently trained on these plots to predict three attributes: species composition; tree diameter size class; and canopy cover class. Species composition is captured as a combination of up to two different species groups to distinguish a forest type (e.g., Douglas-fir + western hemlock is a distinct class from Douglas-fir alone or Douglas-fir + western redcedar). A total of 70 different categories include different combinations of up to two tree species groups (e.g., True Firs count as a single species group). Tree diameter size class reflect the estimated Quadratic Mean Diameter (QMD) of trees binned into the following categories: nonstocked (<1"); seedling/sapling (1-5"); small (5-10" QMD); medium (10-15"); large (15-20") and very large (20"+). Canopy cover is captured in four classes: sparse (<10%); open (10-40%); moderate (40-70%); and closed (70%+). Three separate models have been trained to predict each of these variables.

Based on initial model exploration, the predictions for canopy cover and tree diameter size class utilize only lidar-derived and satellite imagery-derived predictor variables. Species composition classification utilizes these features as well as additional attributes including distance to the nearest stream or water body, soil attributes reflecting bulk density and texture, potential vegetation type, and recent down-scaled climate data from a single point in time.

In addition to the most likely class for each target label, these models generate estimates of the probability that a pixel belong to any of the available classes. For example, a typical prediction might indicate a pixel has a 75% probability of being in the 5-10" QMD class, a 20% probability of being in the 10-15" QMD class, and 5% for the 15-20" QMD class. These probability estimates help communicate how well the model generalizes what it has learned and how well it can distinguish between classes. As described below, these probabilities also offer an opportunity to guide subsequent clustering of pixels into larger management units or forest stands covering several acres or more.

Figure 2: Locations of Inventory Plots

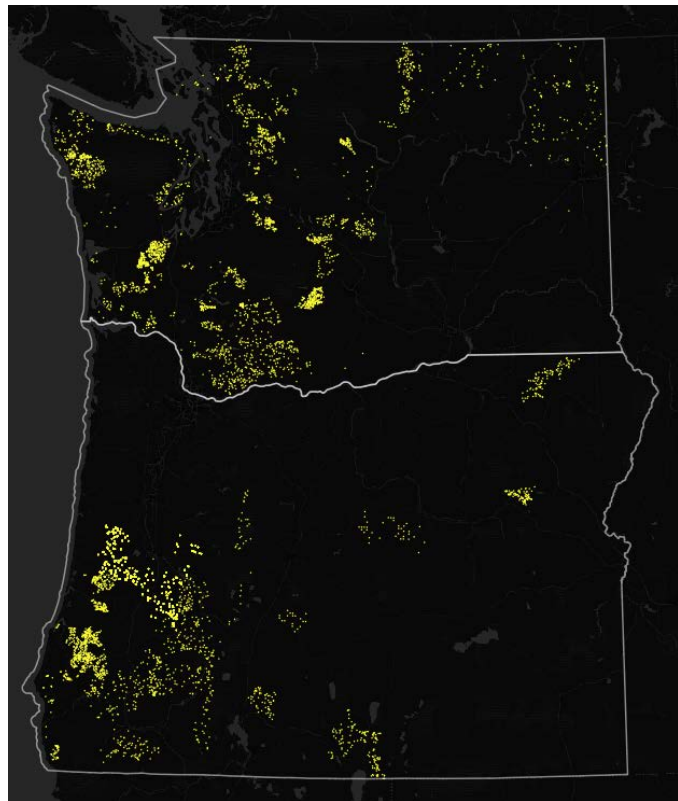


Table 1: Inventory Plots used for Model Training

Agency	Plot Size	# Plots
BLM	1/8-acre (42' radius)	1,860
WA DNR	1/10-acre (37' radius)	3,500
USFS	1/4-acre (59' radius)	800



## 2. OVERVIEW OF FOREST TYPE MAPS

Forest type maps have now been generated covering four distinct lidar acquisitions: one from the Portland Metro area acquired in 2014, and three across south-central Oregon collected from 2010-2017. The maps below illustrate the predictions currently produced by our best-fitting models. In general, these models appear to predict QMD within 3", on average, of the field-measured QMD, and within 10%, on average, of the canopy cover estimated from the plot measurements.

Quantifying accuracy for species composition remains a work in progress. We currently have estimates from the model in the form of a confusion matrix which show the number of occurrences of predicted-vs.-actual classes in table form. However, because many of the forest types are closely related (e.g., misclassifying a Douglas-fir plot as a Douglas-fir + western hemlock plot should not be treated as serious an error as misclassifying a Douglas-fir plot as an aspen plot would be). As we embark upon more detailed model tuning, we expect to report more robust model performance metrics in subsequent grant reports for this project.

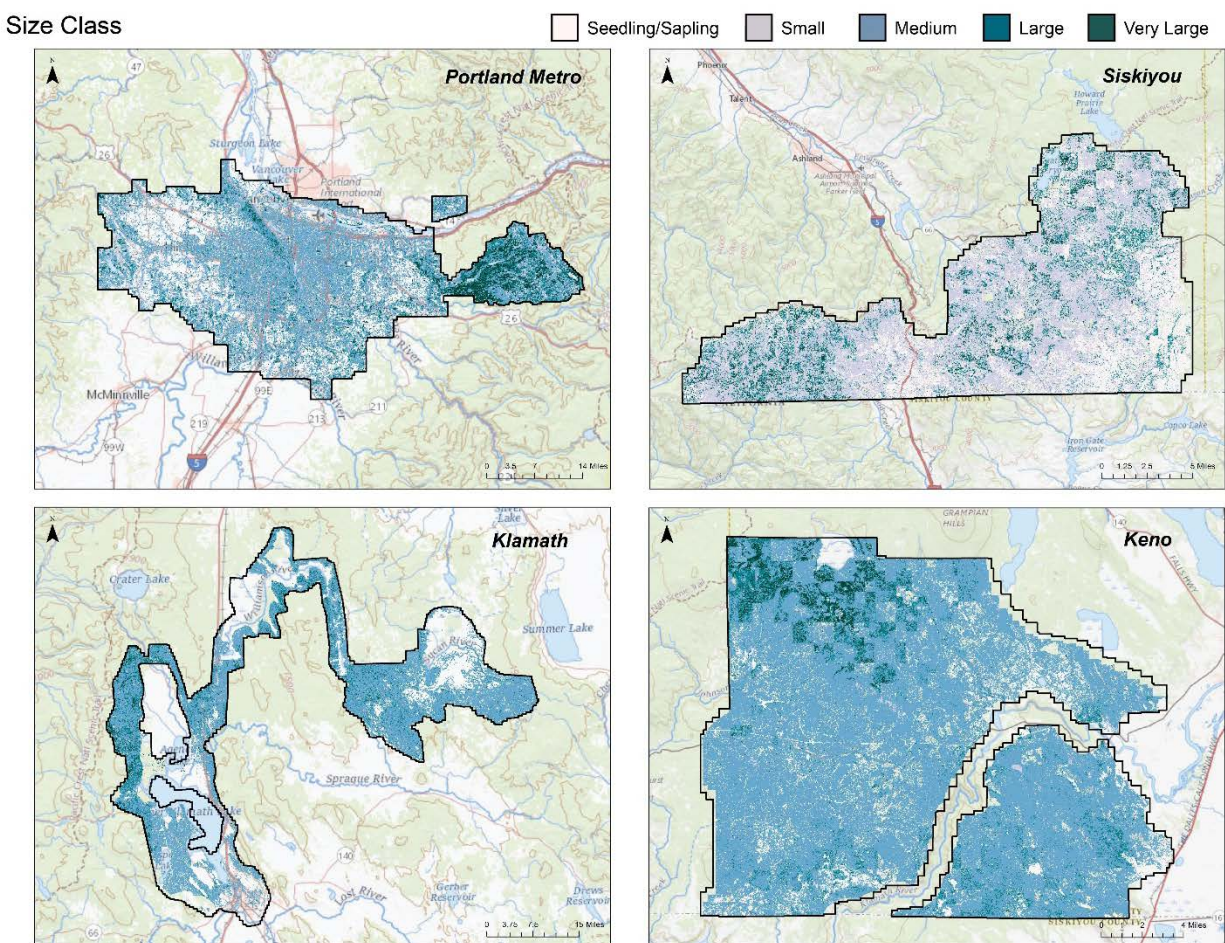


Figure 3: First Generation of Machine Learning Forest Prediction - Size Class



## Cover Class

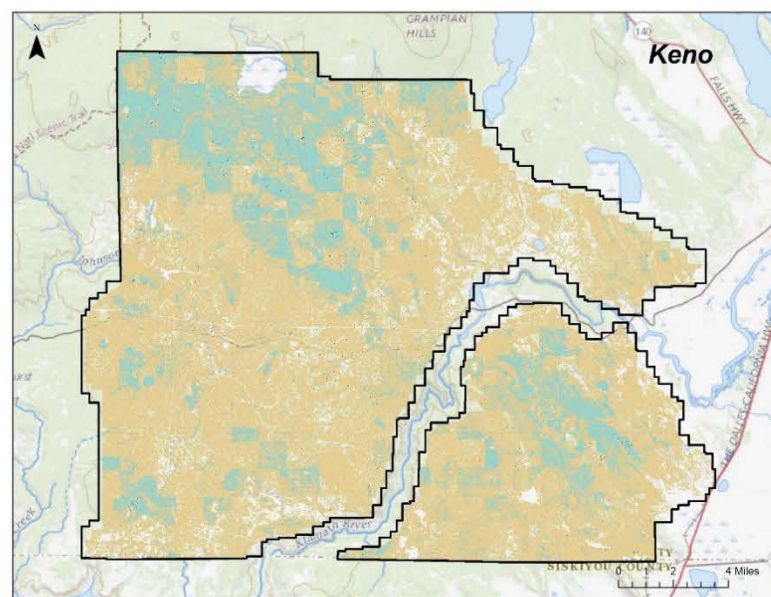
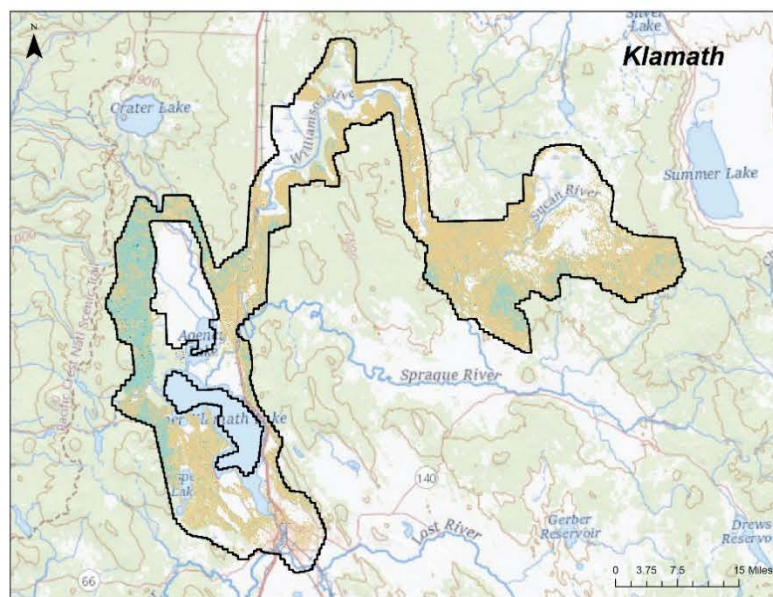
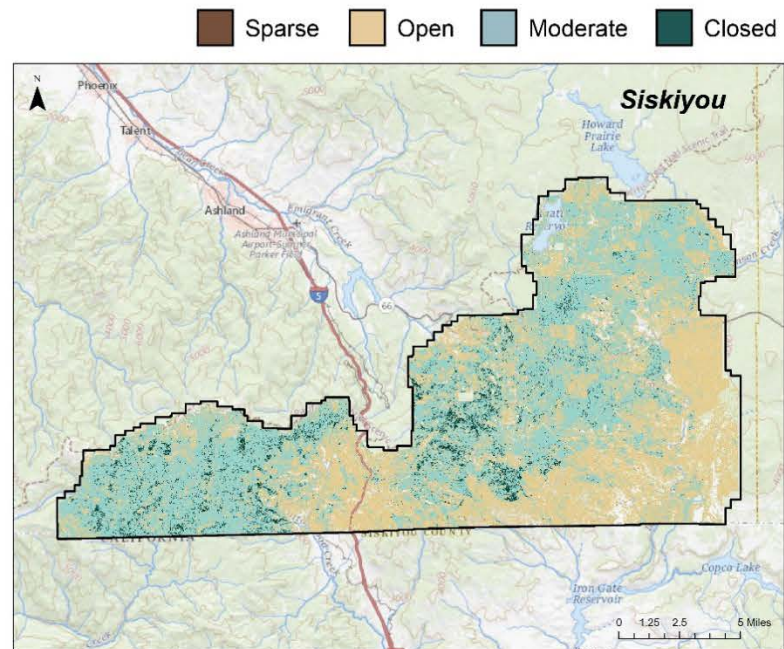
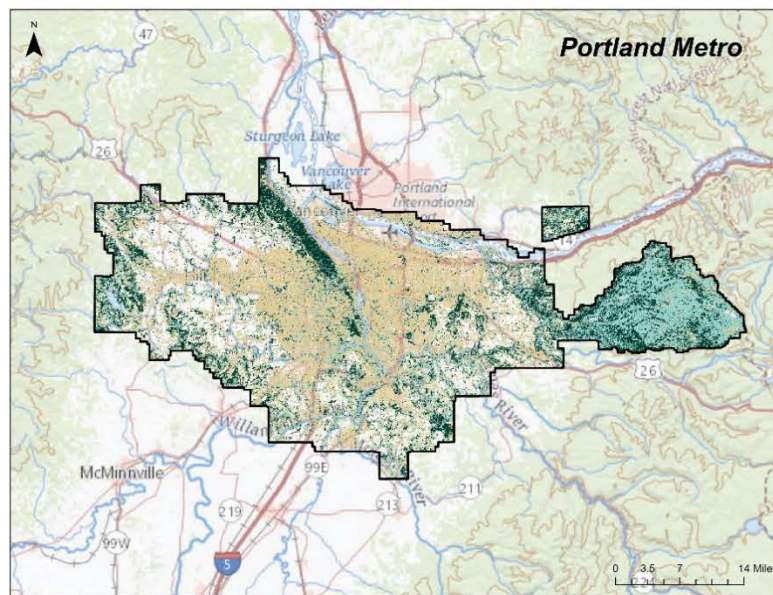
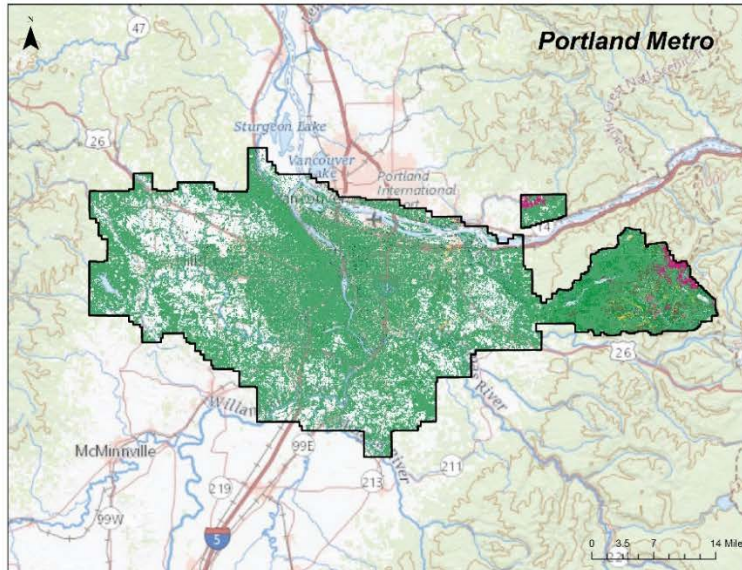


Figure 4: First Generation of Machine Learning Forest Prediction - Cover Class



## Species Composition



## 70 Species group combinations (see Apendix)

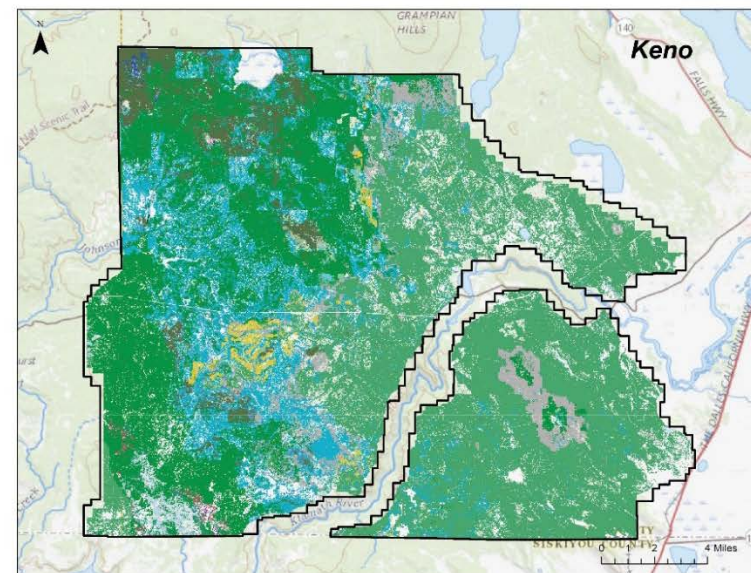
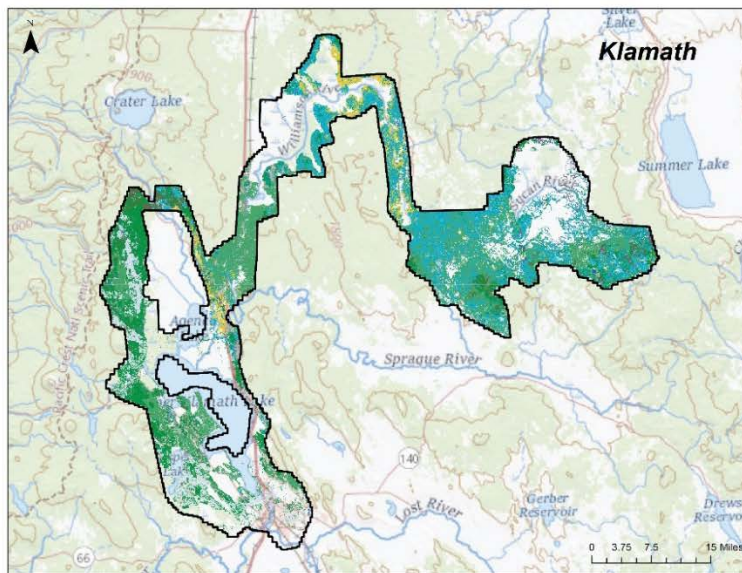
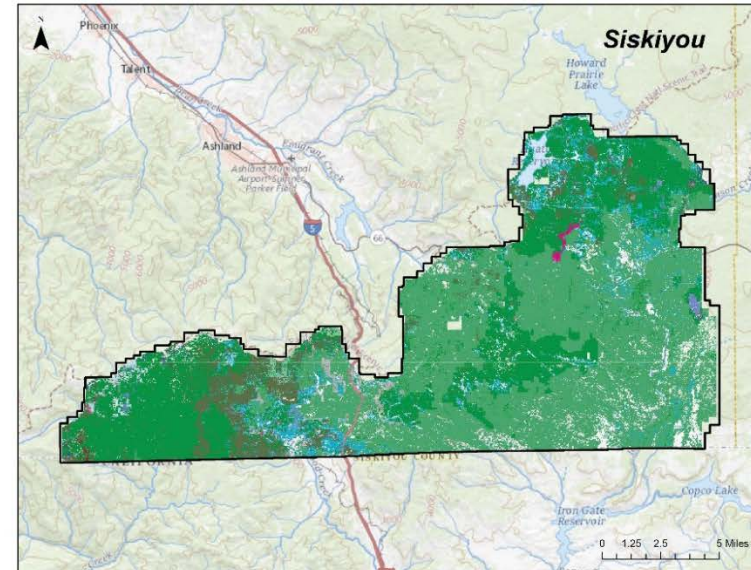
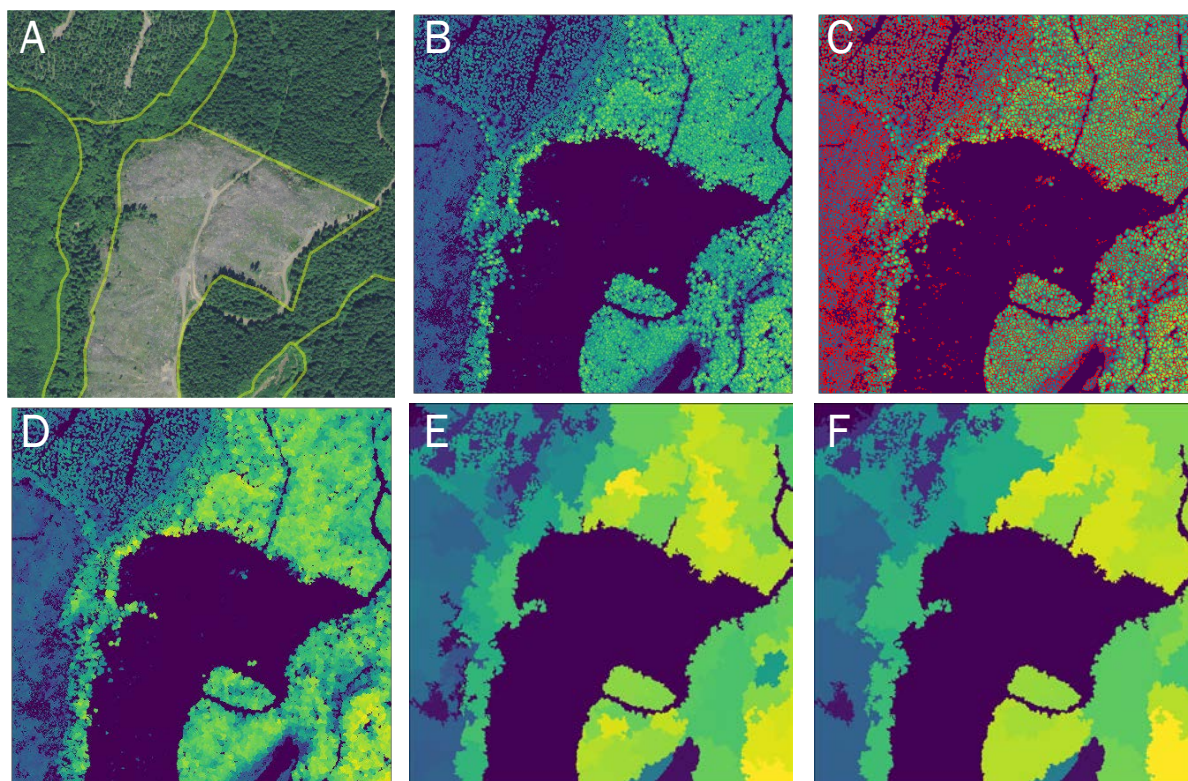


Figure 5: First Generation of Machine Learning Forest Prediction - Species Compositio



### 3. NEXT STEPS TOWARDS AUTOMATED STAND DELINEATION

Although most researchers generally focus their efforts on producing confident pixel-based estimates of forest conditions, our project has set our focus on generating stand delineation for the purpose of integration with widely-used Forest Management Plan templates in Oregon and Washington. To accomplish this objective, we must delineate larger management units such as patches or stands which correspond to coherent forest conditions of reasonable operational scale (e.g., 5-20+ acres each).



**Figure 6: Region-merging segmentation algorithm demonstration with canopy height**

A) Natural color image of 1000x1000m forest area with stand boundaries as drawn by Oregon Department of Forestry. B) Lidar-derived canopy height model at 0.5m resolution; C) TAO boundaries determined through watershed segmentation of canopy height model; D) Mean height assigned to each TAO; E) Region-merging of TAOs using graph-based segmentation algorithm until regions are at least 0.5 acres; F) Region-merging continued until regions are at least 1.0 acres.













































We currently have a working pipeline for executing stand delineation using a region-merging algorithm implemented following an initial segmentation of a canopy height model into “tree clumps” or tree-associated objects, which is illustrated in the figure above. The region-merging algorithm decides which clusters should be merged into successively larger units based on a statistical distance measure between adjacent clusters. In the case above, the distance is calculated solely based on the average canopy height within each region. In traditional image analysis, this segmentation is guided by distance in red, green, and blue bands of an image. Generalizing this approach further, we will integrate a variety of indicators that help to maximize the separation between different forest types. To identify those indicators (and transformation of them to maximize separability of forest types), we will expand upon the canopy height segmentation template by adding fields which are observed to distinguish between existing stand delineations including hundreds of thousands of polygons drawn by Oregon Department of Forestry and Washington Department of Natural Resources.



To accomplish this work, we will continue processing lidar data that coincides with these stand delineations, allowing us not only to expand forest type predictions to these new areas, but also to extract more information about how to recognize forest types from them as well. This combination of plot-level and stand-level inventory measurements is one of the unique aspects of our approach in this project.

## 4. APPENDIX

Legend for Species Composition maps.

	DOUGLAS FIR
	DOUGLAS FIR / ENGELMANN AND OTHER SPRUCES
	DOUGLAS FIR / INCENSE CEDAR
	DOUGLAS FIR / LODGEPOLE PINE
	DOUGLAS FIR / OTHER WESTERN SOFTWOODS
	DOUGLAS FIR / PONDEROSA AND JEFFREY PINE
	DOUGLAS FIR / WESTERN HEMLOCK
	DOUGLAS FIR / WESTERN LARCH
	DOUGLAS FIR / WESTERN WHITE PINE
	ENGELMANN AND OTHER SPRUCE
	LODGEPOLE PINE
	NONSTOCKED
	OTHER WESTERN HARDWOODS
	OTHER WESTERN HARDWOODS / DOGULAS FIR
	OTHER WESTERN HARDWOODS / PONDEROSA PINE
	OTHER WESTERN HARDWOODS / RED ALDER
	OTHER WESTERN SOFTWOODS
	PONDEROSA PINE AND JEFFREY PINES
	PONDEROSA PINE AND JEFFREY PINES / LODGEPOLE PINE
	PONDEROSA PINE AND JEFFREY PINES / OTHER PINES
	RED ALDER
	RED ALDER / WESTERN HEMLOCK
	RED ALDER / WESTERN REDCEDAR
	TRUE FIR
	TRUE FIR / DOUGLAS FIR
	TRUE FIR / INCENSE CEDAR
	TRUE FIR / LODGEPOLE PINE
	TRUE FIR / OTHER HARDWOODS
	TRUE FIR / PONDEROSA PINE
	TRUE FIR / WESTERN HEMLOCK
	WESTERN RED CEDAR
	WESTERN RED CEDAR / WESTERN HEMLOCK
	WESTERN RED CEDAR / WESTERN LARCH
	INCENSE CEDAR
	COTTONWOOD AND ASPEN
	COTTONWOOD AND ASPEN / LODGEPOLE PINE
	OAK
	OAK DOUGLAS FIR
	SITKA SPRUCE
	SUGAR PINE
	WESTERN HEMLOCK
	WESTERN LARCH
	WESTERN LARCH / WESTERN WHITE PINE
	WESTERN LARCH / WHITE PINE