Forest Type Prediction: Accuracy Assessment

Report to NRCS Oregon

In partial fulfillment of Grant No. NR19436XXXXG012
NRCS Oregon - USDA

Authors:
David Diaz
Sara Loreno

Submitted April 2020

For questions or comments, please contact

Sara Loreno
Natural Resources Data Scientist
sloreno@ecotrust.org
(503)467-0784

Ecotrust
721 NW 9th Avenue, Suite 200
Portland, OR 97209

ecotrust

TABLE OF CONTENTS

LIST OF FIGURES

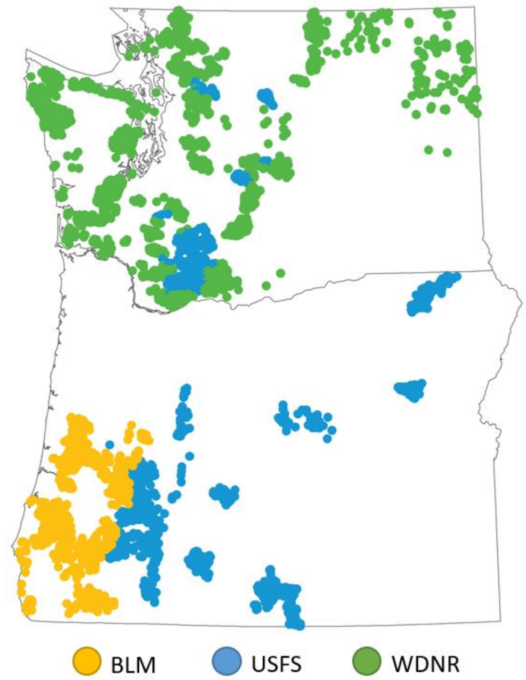
LIST OF TABLES

1. BACKGROUND

In this report, we provide an update on our work processing raw data from publicly-available satellite and airborne laser scanning (lidar) data sources to generate maps for forest conditions for use by non-industrial forest owners in Oregon for the purposes of completing Forest Management Plans.

This project has applied Machine Learning algorithms to develop maps with the predicted canopy cover, diameter size class, dominant canopy height, and dominant species composition Earlier reports have described and illustrated the data sources and attributes that went into our modeling. In this report, we focus on providing information about how accurately our models performed for the prediction of several basic forest type attributes.

Figure 1: Forest Inventory Plot Locations



2. OUTPUTS GENERATED

2.1. Reproducible Open-Source Code

The code demonstrating the entire process from processing raw data through generating predictive maps has been published online with permissive open-source licenses. Code for processing lidar data into raster products including canopy height maps, digital elevation models, topographic hillshading, etc. is published here: https://github.com/Ecotrust/pyFIRS. With the support of a Microsoft AI for Earth Grant, this lidar-processing pipeline has now also been implemented using virtual machines through Microsoft Azure cloud computing services. Template-style Jupyter Notebooks which are employed to execute the pipeline are published here: https://github.com/Ecotrust/pyFIRS/tree/master/pyFIRS/examples

The work to train and apply new models to predict forest conditions was executed in open-source Python code and is published online here: https://github.com/Ecotrust/ForestMapping. The work ranging from extracting plot-level data from thousands of plots measured across Oregon and Washington (**Figure 1: Forest Inventory Plot Locations**) through the training, tuning, and application of machine learning models is illustrated through a series of Jupyter Notebooks here: https://github.com/Ecotrust/ForestMapping/tree/master/notebooks.

2.2. Lidar-Derived Rasters

A variety of directly-useable data products are generated directly from lidar point clouds. All lidar-derived raster products are formatted into 1000 x 1000m tiles in GeoTiff format. All lidar-derived vector products are formatted as ESRI shapefiles. Lidar points clouds are formatted in LAZ, a compressed version of the LAS point cloud format.

High-resolution raster products have been generated for each lidar acquisition we have processed including a 1m-resolution “bare earth” Digital Elevation Model (DEM); a 1m-resolution topographic hillshade image, a 0.5m-resolution raster for the intensity of lidar returns; and a 0.5m-resolution Canopy Height Model (CHM). Vector products include the footprint of the lidar acquisition, a tile index following a 1000m-x-1000m tiling scheme, and building footprints based on points estimated to belong to buildings through the use of the lasclassify tool from LAStools. In addition, the processed (and classified) point clouds are also retained.

In addition to these products which are expected to be directly useful for forest owners and service providers (including through the Land Mapper web app we are developing to serve them to the public) , we have also produced dozens of lidar-derived rasters at 10m resolution that are expected to be useful for forest mappers and modelers. These include numerous metrics generated using the FUSION GridMetrics and GridSurfaceStats tools which describe, among other things, the horizontal and vertical distributions of lidar points within each grid cell. These are typically the metrics that are employed for building predictive models.

We are currently in the process of migrating these data products to Microsoft Azure cloud storage where they will be made publicly-available.

2.3. Predicted Forest Type Attributes

We tested numerous predictive models and chose the best-fitting ones to generate the following attributes across lidar acquisitions at 10m spatial resolution:

- **Dominant Height –** also known as “Lorey Height” represents the average height of trees that are in dominant canopy positions, traditionally measured from the 40 tallest trees per acre.
- **Quadratic Mean Diameter (QMD) –** average diameter at breast height of trees, weighted by trees of larger size.
- **Canopy Cover –** the percentage of ground area that is directly covered by tree crowns. This attribute is largely derived from directly-measured canopy cover from lidar, but has adapted to reproduce canopy cover estimates produced by the Forest Vegetation Simulator from measured field plots.
- **Growing Stock Trees per Acre (TPA)** – the number of trees per acre that are at least 5” DBH or greater. This attribute was chosen because it is less sensitive to large numbers of seedlings and saplings which are commonly observed from natural regeneration.
- **Basal Area (BA) –** the cross-sectional area of tree stems measured at breast height and reported on a per acre basis.
- **Total Cubic Volume –** the aboveground volume of live trees as estimated from allometric equations for each species using measured stem diameters and tree heights.
- **Species Composition –** a classification/rating indicating the absence, presence, abundance, or dominance of a tree species group. Predictions were made for species groups including Douglas-fir, Hemlock (western + mountain), Cedar (western red + incense), True Firs, Maples, Alders, Ponderosa + Jeffrey Pine, Spruce, Tanoak, Juniper, Larch, Other Softwoods, and Other Hardwoods

2.4. Accuracy Measures

2.4.1. Regression Models

Dominant Height, QMD, Canopy Cover, Growing Stock TPA, and BA are recorded as integer or real-valued (i.e., floating point) data. Prediction models for these attributes were built to solve a regression (as opposed to classification) task using the scikit-learn Python package. Models were trained and tuned using 80% of the available plot-level data, while 20% of the plot-level data were set aside to test fitted models and quantify their accuracy. We evaluated a variety of models including Random Forests (RF), Support Vector Machines (SVM), k-Nearest Neighbors (kNN), and Gradient Boosting (GB) Machines. We also fit several linear models using all or a subset of available predictor variables using LASSO and ElasticNet approaches which regularize (or penalize) the inclusion of many predictor variables.

Our regression models were evaluated using several metrics, including:

- **Mean Absolute Error (MAE)** – average magnitude of prediction errors without regard for the direction of error, calculated as mean of the absolute value of differences between predictions and observations. Ranges from 0 to infinity, with 0 indicating perfect prediction. Unlike RMSE, this scoring method does not assign heavier weight to larger errors.
- **Mean Bias Error (Bias)** – average prediction error considering direction of error, calculated as the mean of differences between predictions and observations and reported in the same untis of the variable being predicted. Ranges from negative infinity to positive infinity, with zero indicating an unbiased model.

- **Root Mean Squared Error (RMSE)** – average magnitude of prediction errors weighted quadratically by the size of the error, calculated as the square root of the mean of the sum of squared differences between predictions and observations and reported in the same units of the variable being predicted. Ranges from 0 to infinity, with 0 indicating perfect prediction.
- **Normalized RMSE (NRMSE)** – RMSE divided by the mean value of the predicted attribute, reported as a percentage.

A recent study in British Columbia (Tompalski et al., 2019) investigating the transferability of machine learning models to predict forest attributes between different lidar acquisitions suggested that training models only on a single region's data (as opposed to training them on all available data across regions) could lead to better predictive performance. To investigate this, for all regression tasks, we fit "regional" models that were trained exclusively on plots that were recorded within a single EPA Level III Ecoregion as well as "global" models that were trained on all plot data cross all ecoregions. In contrast to Tompalski et al. (2019), we found that global models commonly outperformed regional models.

2.4.2. Classification Models

The target variable to be predicted for species composition was derived from proportion of plot-level basal area occupied by a species group. The proportion of basal area was binned and converted to an ordinal variable (a categorical variable with ordered values).

Species abundance was rated using the following classes:

- Absent: 0% of plot basal area (BA)
- Present: 0-33% of BA
- Abundant: 33-66% of BA
- Dominant: 66+% of BA

In addition to the predictive models fitted for each species group, a model to predict the proportion of hardwoods and softwoods was also developed. Hardwood-vs-Softwood abundance using the following classes:

- Nonstocked (0% BA by softwoods and hardwoods)
- Dominant Hardwood, No Softwood (100% hardwood)
- Dominant Hardwood, Minor Softwood (>66% hardwood, <33% softwood)
- Hardwood > Softwood, both Abundant (>33% hardwood, >33% softwood)
- Evenly mixed (% hardwood = % hardwood)
- Softwood > Hardwood, both Abundant (>33% softwood, >33% hardwood)
- Dominant Softwood, Minor Hardwood (>66% softwood, <33% hardwood)
- Dominant Softwood, No Hardwood (100% softwood)

Classification accuracy is quantified using the following metrics:

- **Precision** – The number of true positives divided by the number of true positives plus false positives. Precision quantifies the ability of a classifier not to label as positive a sample that is negative. Ranges between 0 and 1.
- **Recall –** The number of true positives divided by the number of true positives plus false negatives. Recall quantifies the ability of the classifier to find all the positive samples. Ranges between 0 and 1, with 1 indicating perfect recall.
- **Specificity –** The number of true negatives divided by the number of true negatives plus false positives. Specificity quantifies the ability of the classifier to find all the negative samples. Ranges between 0 and 1, with 1 indicating perfect recall.

- **F1 Score –** The harmonic mean of precision and recall, calculated as 2 times precision times recall divided by precision plus recall. Ranges between 0 and 1, with 1 indicating perfect prediction.
- **Cohen's Kappa** – A statistic that characterizes how well two classifications agree adjusted by the amount of agreement that could be expected due to chance. Given the ordinal nature of the species abundance data, a linear-weighted version of Cohen's Kappa is used which more heavily penalizes confusion of far-apart categories. Ranges between -1 and 1, with zero indicating the level agreement expected solely to chance, and 1 indicating perfect agreement. The scale proposed by Landis and Koch (1977) for interpreting Cohen's Kappa values is in common use:

Kappa Statistic	Interpretation
< 0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost perfect

3. RESULTS

In tables of results for regression models, highlighted values indicate the best-performing model type for that region's subset of data. These scores reflect when the global model, trained on data from all regions, are used to predict a forest attribute on held-out data from each region (or across all regions). The results for the models trained only on a single region's samples are not shown in this report. In each table, LASSO-5 refers to LASSO linear regression constrained to using only five predictor variables.

For the sake of brevity, only MAE and Bias results for regression models are displayed in this report. Tables containing all metrics, as well as the comparisons of regional versus global models are accessible via the links provided beneath each section.

3.1. Dominant Height

Full set of accuracy ratings as well as graphs available at:
https://github.com/Ecotrust/ForestMapping/blob/master/notebooks/15A_Fitting%20sklearn%20Models%20for%20Top%20Height.ipynb

Table 1: Dominant Height, Mean Absolute Error (MAE), feet

	Model Type							Sample Size	Avg. Height
	Elastic Net	**LASSO**	**LASSO-5**	**kNN**	**RF**	**GB**	**SVM**		
REGIONAL RESULTS									
Blue Mountains	10.0	9.9	9.6	10.3	8.9	9.5	9.9	104	61
Coast Range	12.3	12.2	12.7	12.6	11.4	11.1	11.7	1,580	109
North Cascades	11.5	11.3	11.0	9.3	8.9	9.3	9.7	443	85
Cascades	11.0	10.8	12.1	10.8	9.9	9.8	9.5	613	101
Klamath Mountains	12.2	12.4	14.0	11.7	10.1	10.3	11.0	275	95
Eastern Cascades	10.3	10.0	10.9	8.6	7.4	7.1	8.3	203	79
Northern Rockies	8.8	8.6	8.9	9.7	8.1	8.4	9.7	83	73
Puget Lowland	5.8	5.9	6.9	6.1	5.7	5.3	6.5	122	94
Willamette Valley	6.1	5.3	7.0	9.7	7.6	7.5	9.1	41	112
GLOBAL RESULTS	11.4	11.3	11.9	11.1	10.0	10.0	10.5	3,473	101

Table 2: Dominant Height, Mean Bias Error (Bias), feet

	Model Type							Sample Size	Avg. Height
	Elastic Net	**LASSO**	**LASSO-5**	**kNN**	**RF**	**GB**	**SVM**		
REGIONAL RESULTS									
Blue Mountains	1.1	1.5	-2.4	2.3	0.3	-0.1	0.8	104	61
Coast Range	-1.8	-1.9	-0.8	-1.9	-0.8	-1.1	-1.1	1,580	109
North Cascades	-1.1	-0.7	-1.9	-0.2	0.1	0.2	-1.4	443	85
Cascades	-3.1	-2.8	-4.3	-3.3	-4.0	-3.7	-2.9	613	101
Klamath Mountains	5.2	5.8	6.2	3.0	4.8	4.6	5.6	275	95
Eastern Cascades	-0.9	-0.6	-2.9	-1.1	-1.6	-1.7	0.1	203	79
Northern Rockies	1.3	1.5	-0.7	1.1	1.1	1.7	1.3	83	73
Puget Lowland	1.5	1.2	2.1	2.4	2.9	1.8	3.5	122	94
Willamette Valley	-3.8	-3.3	-3.0	-7.7	-7.6	-7.5	-8.4	41	112
GLOBAL RESULTS	-1.0	-0.9	-1.1	-1.2	-0.7	-0.8	-0.6	3,473	101

3.2. Quadratic Mean Diameter (QMD)

Full set of accuracy ratings as well as graphs available at:
https://github.com/Ecotrust/ForestMapping/blob/master/notebooks/15B_Fitting%20sklearn%20Models%20for%20QMD.ipynb

Table 3: QMD, Mean Absolute Error (MAE), inches

	Model Type							Sample Size	Avg. QMD
	Elastic Net	**LASSO**	**LASSO-5**	**kNN**	**RF**	**GB**	**SVM**		
REGIONAL RESULTS									
Blue Mountains	1.7	1.6	1.7	1.5	1.5	1.4	1.3	104	12.4
Coast Range	3.8	3.8	3.8	3.9	3.7	3.8	3.9	1,580	14.9
North Cascades	2.7	2.7	2.8	2.7	2.7	2.7	2.7	443	11.7
Cascades	2.8	2.9	2.9	2.7	2.5	2.6	2.7	613	13.0
Klamath Mountains	5.7	5.7	5.9	5.1	5.0	5.1	5.2	275	11.1
Eastern Cascades	2.8	2.7	2.8	2.4	2.4	2.5	2.5	203	11.9
Northern Rockies	2.4	2.3	2.4	2.2	2.3	2.1	1.7	83	11.0
Puget Lowland	2.1	2.0	1.9	2.5	2.3	2.3	2.5	122	12.5
Willamette Valley	5.8	5.5	5.4	6.8	5.8	5.6	5.2	41	17.5
GLOBAL RESULTS	3.5	3.5	3.5	3.4	3.3	3.3	3.4	3,473	13.6

Table 4: QMD, Mean Bias Error (Bias), inches

	Model Type							Sample Size	Avg. QMD
	Elastic Net	**LASSO**	**LASSO-5**	**kNN**	**RF**	**GB**	**SVM**		
REGIONAL RESULTS									
Blue Mountains	-1.6	-1.4	-1.5	0.6	-0.4	-0.5	0.2	104	12.4
Coast Range	-1.2	-1.2	-1.2	-1.2	-1.0	-1.0	-0.7	1,580	14.9
North Cascades	-0.5	-0.4	-0.5	0.1	-0.1	-0.2	-0.0	443	11.7
Cascades	-0.4	-0.3	-0.5	-0.5	-0.5	-0.4	0.0	613	13.0
Klamath Mountains	2.4	2.5	2.7	1.6	1.8	1.8	2.4	275	11.1
Eastern Cascades	-1.1	-0.9	-1.1	-0.8	-1.2	-1.1	-0.5	203	11.9
Northern Rockies	0.6	0.8	1.0	0.8	0.6	0.5	0.1	83	11.0
Puget Lowland	0.4	0.3	-0.0	1.5	1.0	1.0	1.3	122	12.5
Willamette Valley	-5.8	-5.5	-5.4	-6.8	-5.8	-5.6	-4.8	41	17.5
GLOBAL RESULTS	-0.7	-0.6	-0.7	-0.5	-0.5	-0.5	-0.2	3,473	13.6

3.3. Canopy Cover

Full set of accuracy ratings as well as graphs available at:
https://github.com/Ecotrust/ForestMapping/blob/master/notebooks/15D_Fitting%20sklearn%20Models%20for%20Canopy%20Cover.ipynb

Table 5: Canopy Cover, Mean Absolute Error (MAE), percent

	Model Type							Sample Size	Avg. Cover
	Elastic Net	**LASSO**	**LASSO-5**	**kNN**	**RF**	**GB**	**SVM**		
REGIONAL RESULTS									
Blue Mountains	8.6	8.5	8.6	7.4	7.2	7.3	6.8	104	25
Coast Range	11.4	11.4	11.7	11.3	11.1	11.1	11.4	1,580	69
North Cascades	11.1	11.0	10.8	10.0	10.4	10.2	10.0	443	62
Cascades	12.0	12.1	12.2	11.1	11.6	11.4	11.2	613	64
Klamath Mountains	12.7	12.6	12.6	12.5	12.6	12.7	12.0	275	58
Eastern Cascades	11.4	11.3	10.8	10.6	10.6	10.6	10.0	203	48
Northern Rockies	8.4	8.3	9.0	10.0	9.6	8.5	7.1	83	43
Puget Lowland	7.5	7.5	7.8	9.0	8.7	8.6	9.1	122	68
Willamette Valley	8.3	8.2	7.7	7.9	8.5	7.8	7.2	41	65
GLOBAL RESULTS	11.4	11.3	11.4	11.0	11.0	10.9	10.8	3,473	64

Table 6: Canopy Cover, Mean Bias Error (Bias), percent

	Model Type							Sample Size	Avg. Cover
	Elastic Net	**LASSO**	**LASSO-5**	**kNN**	**RF**	**GB**	**SVM**		
REGIONAL RESULTS									
Blue Mountains	1.7	1.6	1.9	-1.0	0.8	0.6	-0.9	104	25
Coast Range	-0.4	-0.5	-0.7	-0.2	-0.2	-0.3	-2.0	1,580	69
North Cascades	0.8	0.8	0.8	-0.6	0.6	0.5	-2.2	443	62
Cascades	2.6	2.6	2.6	1.4	2.0	1.9	-0.1	613	64
Klamath Mountains	5.5	5.3	5.4	4.7	3.9	3.5	0.1	275	58
Eastern Cascades	5.0	5.0	4.2	3.9	3.4	3.3	1.3	203	48
Northern Rockies	5.7	5.5	6.2	4.7	6.3	5.2	1.5	83	43
Puget Lowland	-3.3	-3.3	-4.5	-5.1	-4.6	-4.3	-4.8	122	68
Willamette Valley	3.6	3.5	2.7	2.9	3.5	2.8	-0.8	41	65
GLOBAL RESULTS	1.4	1.3	1.2	0.8	1.0	0.9	-1.1	3,473	64

3.4. Growing Stock Trees per Acre (TPA)

Full set of accuracy ratings as well as graphs available at:
https://github.com/Ecotrust/ForestMapping/blob/master/notebooks/15E_Fitting%20sklearn%20Models%20for%20Growing%20Stock%20TPA.ipynb

Table 7: Growing Stock TPA, Mean Absolute Error (MAE), trees per acre

	Model Type							Sample Size	Avg. TPA
	Elastic Net	**LASSO**	**LASSO-5**	**kNN**	**RF**	**GB**	**SVM**		
REGIONAL RESULTS									
Blue Mountains	24.1	24.2	23.8	29.3	29.3	30.5	22.2	104	72
Coast Range	57.4	57.4	59.8	57.2	56.0	57.0	57.2	1,580	156
North Cascades	61.3	61.2	67.5	63.1	62.2	59.7	59.6	443	180
Cascades	58.0	58.0	61.5	56.9	55.9	54.2	55.7	613	177
Klamath Mountains	50.4	50.4	49.8	48.8	47.9	49.0	47.3	275	149
Eastern Cascades	43.6	43.6	46.4	41.2	41.7	42.3	40.4	203	123
Northern Rockies	53.9	53.9	58.3	55.6	57.9	52.7	53.0	83	134
Puget Lowland	66.6	66.6	65.7	60.7	64.8	65.1	59.6	122	174
Willamette Valley	47.0	47.1	45.6	43.8	39.1	40.0	38.4	41	150
GLOBAL RESULTS	55.7	55.7	58.5	55.4	54.6	54.4	54.1	3,473	160

Table 8: Growing Stock TPA, Mean Bias Error (Bias), trees per acre

	Model Type							Sample Size	Avg. TPA
	Elastic Net	**LASSO**	**LASSO-5**	**kNN**	**RF**	**GB**	**SVM**		
REGIONAL RESULTS									
Blue Mountains	-15.2	-15.3	-12.2	-18.0	-18.1	-12.8	-11.4	104	72
Coast Range	-8.5	-8.5	-12.7	-9.0	-8.1	-9.7	-9.5	1,580	156
North Cascades	-26.7	-26.7	-26.1	-32.7	-27.7	-24.8	-26.3	443	180
Cascades	-19.5	-19.4	-31.2	-26.7	-22.4	-18.9	-16.6	613	177
Klamath Mountains	-13.4	-13.4	-14.2	-17.6	-16.7	-15.1	-21.6	275	149
Eastern Cascades	-8.8	-8.9	-3.7	-6.4	-8.5	-4.5	-3.2	203	123
Northern Rockies	-14.2	-14.2	-11.0	-13.8	-13.9	-8.7	-2.3	83	134
Puget Lowland	-46.7	-46.7	-41.7	-45.2	-47.3	-50.7	-46.4	122	174
Willamette Valley	28.4	28.5	33.5	22.5	22.3	19.0	15.7	41	150
GLOBAL RESULTS	-14.2	-14.2	-17.3	-16.8	-15.1	-14.3	-14.2	3,473	160

3.5. Basal Area

Full set of accuracy ratings as well as graphs available at:
https://github.com/Ecotrust/ForestMapping/blob/master/notebooks/15F_Fitting%20sklearn%20Models%20for%20BA.ipynb

Table 9: Basal Area, Mean Absolute Error (MAE), square feet per acre

	Model Type							Sample Size	Avg. BA
	Elastic Net	**LASSO**	**LASSO-5**	**kNN**	**RF**	**GB**	**SVM**		
REGIONAL RESULTS									
Blue Mountains	16.6	16.7	19.3	16.7	15.8	17.4	15.4	104	58
Coast Range	69.3	69.3	73.0	71.6	71.4	70.0	69.5	1,580	230
North Cascades	53.1	52.9	61.9	51.8	50.5	49.5	53.3	443	170
Cascades	49.1	48.9	56.1	52.8	50.3	49.9	49.5	613	203
Klamath Mountains	50.4	50.5	61.0	48.8	48.1	46.9	50.7	275	205
Eastern Cascades	27.8	27.9	38.2	28.7	30.2	29.4	27.4	203	134
Northern Rockies	25.3	25.4	26.7	23.5	21.8	21.9	24.3	83	98
Puget Lowland	33.3	33.4	37.6	36.6	36.3	36.3	34.7	122	180
Willamette Valley	76.5	76.4	69.2	80.2	78.0	80.9	78.1	41	232
GLOBAL RESULTS	55.7	55.6	61.5	57.1	56.3	55.4	55.8	3,473	204

Table 10: Basal Area, Mean Bias Error (Bias), square feet per acre

	Model Type							Sample Size	Avg. BA
	Elastic Net	**LASSO**	**LASSO-5**	**kNN**	**RF**	**GB**	**SVM**		
REGIONAL RESULTS									
Blue Mountains	6.5	6.6	13.5	2.2	-0.2	2.0	4.8	104	58
Coast Range	-20.6	-20.7	-19.4	-20.1	-17.1	-19.2	-23.0	1,580	230
North Cascades	-9.4	-9.2	-13.0	-15.3	-13.1	-13.5	-11.5	443	170
Cascades	-9.8	-9.5	-20.3	-17.5	-15.5	-13.3	-11.1	613	203
Klamath Mountains	-12.7	-12.2	-3.4	-23.0	-18.1	-15.9	-16.2	275	205
Eastern Cascades	-3.1	-3.0	-0.3	-3.1	-5.4	-3.3	-4.0	203	134
Northern Rockies	6.7	6.6	4.2	5.2	4.5	6.0	6.0	83	98
Puget Lowland	-10.3	-10.2	-7.5	-7.0	-8.4	-13.3	-11.1	122	180
Willamette Valley	-58.8	-58.5	-45.6	-64.8	-62.7	-65.0	-62.0	41	232
GLOBAL RESULTS	-13.8	-13.7	-14.2	-16.7	-14.6	-15.0	-15.9	3,473	204

3.6. Total Cubic Volume

Full set of accuracy ratings as well as graphs available at:
https://github.com/Ecotrust/ForestMapping/blob/master/notebooks/15C_Fitting%20sklearn%20Models%20for%20Cubic%20Volume.ipynb

Table 11: Cubic Volume, Mean Absolute Error (MAE), hundreds of cubic feet per acre

	Model Type							Sample Size	Avg. Cubic Volume
	Elastic Net	**LASSO**	**LASSO-5**	**kNN**	**RF**	**GB**	**SVM**		
REGIONAL RESULTS									
Blue Mountains	5.2	5.5	5.5	5.8	5.8	6.3	6.0	104	14.7
Coast Range	30.1	30.0	30.2	32.1	30.7	30.9	32.0	1,580	95.4
North Cascades	20.9	20.6	21.7	20.6	18.5	17.2	18.4	443	57.0
Cascades	19.5	19.4	19.6	20.8	19.3	19.0	19.1	613	75.8
Klamath Mountains	19.3	19.1	19.8	16.9	17.7	16.4	18.2	275	74.5
Eastern Cascades	8.8	8.6	8.7	10.0	9.6	9.9	9.0	203	43.0
Northern Rockies	4.9	5.1	5.7	8.0	5.2	4.9	4.8	83	27.3
Puget Lowland	14.4	14.2	15.0	17.4	14.8	15.2	16.4	122	63.6
Willamette Valley	23.2	21.1	23.8	28.7	30.6	29.9	32.3	41	95.2
GLOBAL RESULTS	22.8	22.7	23.1	24.0	22.7	22.5	23.3	3,473	79.5

Table 12: Cubic Volume, Mean Bias Error (Bias), hundreds of cubic feet per acre

	Model Type							Sample Size	Avg. Cubic Volume
	Elastic Net	**LASSO**	**LASSO-5**	**kNN**	**RF**	**GB**	**SVM**		
REGIONAL RESULTS									
Blue Mountains	1.5	2.0	1.6	0.9	0.7	1.6	2.4	104	14.7
Coast Range	-11.9	-11.8	-11.5	-11.3	-10.2	-10.8	-12.7	1,580	95.4
North Cascades	-5.6	-4.9	-6.9	-4.6	-4.3	-4.1	-6.9	443	57.0
Cascades	-7.2	-6.4	-7.1	-7.7	-8.0	-6.0	-5.4	613	75.8
Klamath Mountains	-0.4	0.4	1.6	-6.0	-3.3	-2.0	-3.8	275	74.5
Eastern Cascades	-2.7	-2.1	-2.5	-0.4	-1.9	-0.3	-0.9	203	43.0
Northern Rockies	2.2	2.5	2.0	1.9	1.0	1.3	1.9	83	27.3
Puget Lowland	-3.8	-3.8	-3.4	0.5	0.1	-1.7	-4.3	122	63.6
Willamette Valley	-20.1	-17.8	-19.5	-24.3	-27.6	-27.2	-32.1	41	95.2
GLOBAL RESULTS	-7.7	-7.3	-7.5	-7.7	-7.1	-6.8	-8.2	3,473	79.5

3.7. Species Composition

Full set of accuracy ratings available at:
https://github.com/Ecotrust/ForestMapping/blob/master/notebooks/15H_Fitting%20and%20Tuning%20sklearn%20Models%20for%20Species%20Presence%20and%20Abundance.ipynb

For species classification, Random Forest was commonly observed to be the best predictive model in preliminary model runs. Classification accuracy is shown here for Random Forest models only, and was fitted to data from all ecoregions.

It is worth noting that species abundance is highly imbalanced. In particular, a large majority of the training data are from Douglas-fir dominant plots. Class imbalances this large often present major challenges to classification models due to the lack of many examples from which less common observations can be confidently learned.

Table 13: Sample Count by Species Group

SPECIES GROUP	ABSENT	PRESENT	ABUNDANT	DOMINANT
Douglas-fir	2,330	1,362	2,042	5,159
Hemlock	6,224	3,068	970	631
Red Alder	8,023	2,068	408	394
Other Hardwood	8,325	2,136	259	173
Cedar	8,510	1,882	374	127
True Fir	8,889	1,309	410	285
Maple	9,175	1,545	135	38
Ponderosa Pine	10,151	286	178	278
Spruce	10,324	383	109	77
Other Softwood	10,436	425	7	25
Tanoak	10,482	301	82	28
Larch	10,503	244	84	62
Oak	10,528	334	27	4
Lodgepole Pine	10,583	211	46	53
Juniper	10,849	32	5	7

Table 14: Sample Count at Different Levels of Softwood vs. Hardwood Abundance

COMPOSITION CLASSIFICATION	NUMBER OF SAMPLES
Nonstocked	115
Dominant Hardwood, No Softwood	362
Dominant Hardwood, Minor Softwood	408
Hardwood > Softwood, both Abundant	290
Evenly mixed	16
Softwood > Hardwood, both Abundant	576
Dominant Softwood, Minor Hardwood	4,333
Dominant Softwood, No Hardwood	4,793

Table 15: Classification Scores for Tree Species Groups

Species Group	Precision	Recall	Specificity	F1	Kappa
Douglas-fir	0.56	0.58	0.81	0.57	0.52
Hemlock	0.67	0.66	0.81	0.66	0.50
Red Alder	0.73	0.69	0.72	0.71	0.42
Cedar	0.74	0.74	0.54	0.74	0.27
True Fir	0.81	0.81	0.74	0.81	0.49
Maple	0.90	0.88	0.88	0.88	0.51
Ponderosa Pine	0.93	0.92	0.76	0.93	0.60
Spruce	0.93	0.92	0.56	0.92	0.39
Tanoak	0.98	0.96	0.99	0.96	0.49
Larch	0.95	0.94	0.70	0.95	0.39
Oak	0.97	0.97	0.70	0.97	0.45
Lodgepole Pine	0.96	0.93	0.70	0.94	0.46
Juniper	1.00	0.99	1.00	1.00	0.57
Other Hardwood	0.79	0.72	0.73	0.75	0.33
Other Softwood	0.94	0.90	0.41	0.92	0.26
Softwood-Hardwood	0.62	0.61	0.82	0.60	0.50

4. SUMMARY OF RESULTS

In general, the regression models performed very well, predicting basic forest attributes like canopy cover, dominant height, QMD, BA, growing stock TPA, and total cubic volume with very limited bias. The accuracies of these models are comparable or better than published studies that have typically involved much smaller datasets than the one used here.

It should be expected that lidar-derived predictive models perform particularly well in areas of canopy height and canopy cover given the distribution of laser pulses fired from an airplane are commonly returning from the upper canopy. For these metrics, we saw predictions were typically very accurate.

For other attributes that are known to be correlated with canopy height, such as QMD, BA, and total cubic volume, we saw more variance in model predictions and relatively lower accuracy compared to canopy height and cover. Nevertheless, these predictive models are competitive with accuracies reported in the scientific literature. Furthermore, the prediction accuracy is typically finer than the scale of categories used to classify forest conditions into major types. For example, canopy cover is typically binned into classes of 0-10%, 10-40%, 40-70%, and 70-100%, indicating that errors in predictive modeling averaging 10% with near-zero bias are likely to be reliable for cover-level categorization. Similarly, the size class of a forest stand is typically binned by QMD into classes of 0-1”, 1-5”, 5-10”, 10-15”, 15-20” and >20”. With predictions among regions typically with errors averaging in the 2-3.5” range, we believe these models will provide reasonable estimates of average tree size.

In general, we are confident in the performance of these models to predict forest structural attributes.

The predictions of species abundance, however, was much more mixed. The imbalance in the dataset (an overwhelming majority of samples were from Douglas-fir-dominated forest plots) left models with relatively few examples from which to learn to recognize abundance or dominance for many less common species. Thus, many models appeared to default to predicting species absence or presence, and rarely predicting species abundance or dominance, commonly leading to false negatives when less common species were in fact abundant or dominant. This lack of sensitivity to detecting less common species may also be due to the use of a single satellite image and single year of climate data. As we move forward into future work on forest type mapping, we will explore the integration of multiple satellite images over the course of a year to help reduce errors in species composition estimates at the plot level. We are also optimistic that the variation between species mixtures at an aggregated (e.g., stand level of 5-10+ acres) scale will be more likely to be distinguishable in time-series of satellite imagery than predictions at the plot-level will be.

5. REFERENCES CITED

Landis, J.R., Koch, G.G., 1977. The Measurement of Observer Agreement for Categorical Data. Biometrics 33, 159–174. https://doi.org/10.2307/2529310

Tompalski, P., White, J.C., Coops, N.C., Wulder, M.A., 2019. Demonstrating the transferability of forest inventory attribute models derived using airborne laser scanning data. Remote Sens. Environ. 227, 110–124. https://doi.org/10.1016/j.rse.2019.04.006