Features Derived from Remotely-Sensed Data for Forest Type Prediction

Report to NRCS Oregon

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

Authors:
David Diaz
Sara Loreno

January 2019

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 TABLES

LIST OF FIGURES

1. BACKGROUND

In this report, we summarize work to date processing raw data from publicly-available satellite and airborne laser scanning (lidar) data sources. These data are being processed along with down-scaled climate and climate-derived indicators for use in mapping forest conditions across the landscape.

This project intends to utilize modern Machine Learning methods to perform this predictive modeling task. In Machine Learning terminology, we will be developing models for predicting “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 can be 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 will train models using data from a portion of the landscape where both features and labels are available, holding out a portion of those areas for validation of the predictive model before broader use. Once the model has been trained and validated, we will then apply the model to predict forest conditions across the rest of the landscape where remotely sensed data have been processed.

This report describes our data sources and processing methods along with summary statistics and graphics indicating how the various forest conditions for which we currently have inventory data correspond to the “features” derived from remotely-sensed data.

2. DATA SOURCES AND PROCESSING

2.1. Forest Inventory

At the outset of these project, we had planned to utilize forest inventory from the US Forest Service Forest Inventory & Analysis (FIA) program as the ground-truth data for training our predictive models. Following extended consultation with FIA coordinators, we learned that satisfying FIA data confidentiality (privacy concerns associated with locations of FIA plots on private land) would introduce a significant additional effort in order to access the data and to likely preclude the publication of remote-sensing metrics associated with these plots. To resolve this issue, we identified new sources for ground-truth data that do not involve the data confidentiality or privacy concerns.

We acquired stand delineations and corresponding forest inventory data from the Oregon Department of Forestry (**Figure 1: ODF Forest Stand Data Coverage**). Of nearly 15,000 unique forest stands delineated by ODF, just over 4,000 stands have field-collected measurements (from 100,000 plots). The remainder of delineated stands have forest type attributes assigned by ODF forest managers.

Figure 1: ODF Forest Stand Data Coverage

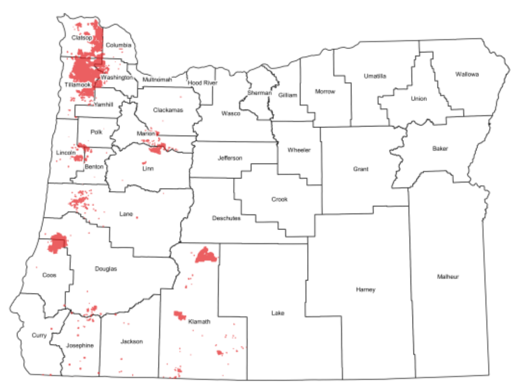
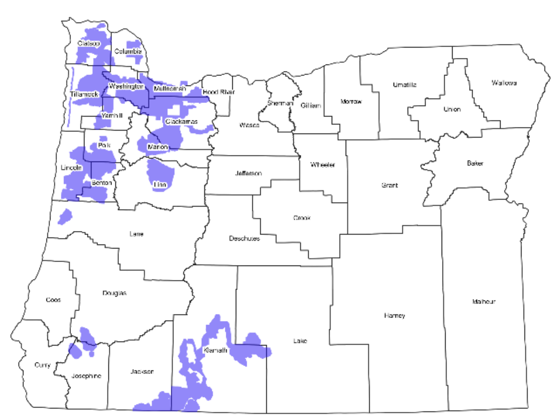


Figure 2: Coverage of Processed Lidar Data



Stands with inventory data were loaded into the Forest Vegetation Simulator (Crookston and Dixon, 2005; Dixon, 2017) to generate fifteen different forest condition indicators at the time of their measurement (see **Table 1: Forest Condition Indicators**).

Table 1: Forest Condition Indicators/Labels

Used in Oregon's FMP Template
Forest Type (species), Diameter Size Class / Growth Stage, Stocking Class, Structure / Successional Class, Number of Canopy Layers

Additional Indicators
Trees per Acre, Basal Area per Acre, Stand Density Index, Crown Competition Factor, Dominant Height, Quadratic Mean Diameter, Total Cubic Volume, Merchantable Cubic Volume, Merchantable Boardfoot Volume, Canopy Cover (%)

2.2. Satellite Imagery

Using Google Earth Engine (Gorelick et al., 2017), we accessed cloud-free 8-band satellite imagery from the Landsat 7 satellite for image tiles that included any forest stands for which we have inventory data. We gathered imagery from the 2007-2018 to include the full range of dates when lidar data were acquired for these stands.

Table 2: Satellite Image Features

Red, Green, Blue, Near Infrared, Shortwave Infrared 1, Shortwave Infrared 2, Brightness, Greenness, Wetness, Normalized Difference Vegetation Index (NDVI), Enhanced NDVI (ENDVI), Soil-Adjusted Vegetation Index (SAVI)

For each forest stand delineated by ODF, we calculated the distributions of twelve different metrics across the stand, generally following the approach of Hudak et al., (2014). We expanded on this approach by recording values at each decile (minimum, 10th percentile, … 90th percentile, maximum) observed across the satellite image pixels within each stand (see Table 2: Satellite Image Features), producing a distribution of values for each metric in each stand. For example, we calculated the minimum, 10^{th} percentile, …, 90^{th} percentile, and maximum value of the red band observed in each stand.

Table 3: Lidar Features

Canopy Height Distribution *(5^{th}, 25^{th}, 50^{th}, 75^{th} 95^{th} and 100^{th} percentile),* Proportion and Median Intensity of Returns by Height (<0.15m, 1.37m, 5m, 10m, 20m, >30m), Canopy Cover (%), Elevation, Slope, Planiform and Profile Curvature, and Solar Radiation Index

All computation to filter, clip, and summarize this satellite imagery was performed in the cloud using Google Earth Engine. We downloaded the derived data, including the stand ID and associated image-derived metrics as a text file.

2.3. Lidar Point Clouds

To date, we have acquired and processed data from 11 lidar acquisitions in Oregon. The footprints of these lidar acquisitions are displayed in **Figure 2: Coverage of Processed Lidar Data**. and their attributes are summarized in **Table 4**. All lidar point clouds had ground-classified points, and these points were used to construct terrain models. Further processing of the point clouds was performed to remove noise, classify points into vegetation and buildings, and generate a variety of derived products including a canopy height model, footprints of buildings, and 24 different terrain- and canopy-related metrics (see Table 3: Lidar Features). These 24 metrics were generated as a 10m resolution raster. For each stand, the deciles of each of these metrics across the pixels inside that stand were calculated.

Table 4: Lidar Data Processed

Lidar Acquisition	# Points (Billions)	Acres	Compressed Size (GB)
OLC Big Windy (2015)	8.2	140,187	63
OLC Clackamas (2013)	19.6	462,929	110
OLC Green Peter (2012)	10.4	226,301	60
OLC Keno (2012)	7.1	204,667	41
OLC Klamath (2010)	26.6	714,165	105
ODF (2007)	16.0	358,455	69
ODF Northwest (2015)	40.0	432,342	142
OLC Central Coast (2012)	43.8	827,315	227
OLC Metro (2014)	62.9	789,369	312
OLC Siskiyou (2017)	11.1	137,808	74
OLC Yamhill (2012)	9.3	205,779	53
Total	**255.1**	**4,499,318**	**1,256**

2.4. Down-Scaled Climatic Variables

We utilized ClimateNA software (Wang et al., 2016) to extract historical down-scaled climate data including monthly, seasonal, and annual climatic attributes, as well as derived metrics such as

Growing Degree Days, Potential Evapotranspiration, etc. that are useful for vegetation modeling. These data have been used in various peer-reviewed studies for mapping of species distributions, ecoregions, seed zones, bioclimate envelopes, etc. We computed the down-scaled climate attributes for every ODF forest stand using the location and elevation of the centroid of each forest stand. A total of 248 climate and climate-derived metrics are available for modeling. They include, for example, monthly, seasonal, and annual maximum temperature, relative humidity, total precipitation, etc.

3. ILLUSTRATION OF INVENTORY AND REMOTE SENSING DATA

8,127 out of 14,756 stands delineated by ODF are covered by lidar data that we have now processed. 1,904 out of 4,176 stands with inventory data are also covered by lidar data we have processed. The inventory and remotely sensed data from these 1,904 stands (with more stands to be added as we process additional lidar acquisitions) will be used as the training data for our Machine Learning models.

In the graphs below, we present some initial visualizations illustrating how remotely-observed data vary by dominant species (forest type), tree diameter class, and stand structure class. These are intended to reflect the diversity of data we now have available to fit the Machine Learning models rather than an indication of the best variables for discriminating between different forest conditions. As we proceed into predictive modeling, we will be conducting further analysis to determine methods for transforming these indicators (e.g., normalizing, dimensionality reduction, etc.) to identify the features that will ultimate be used as predictor variables in our Machine Learning predictive models.

3.1. Forest Type / Composition

Figure 3: Climate Data by Forest Type

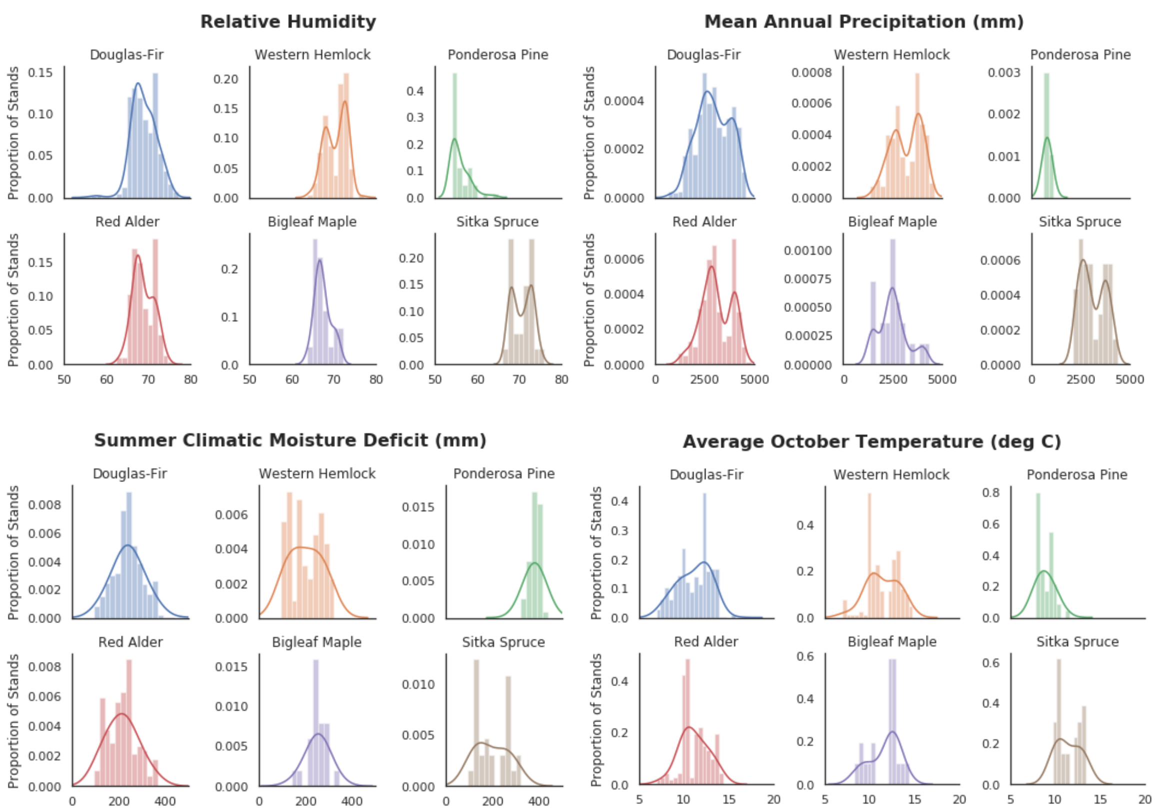


Figure 4: Distributions of Satellite Metrics by Forest Type

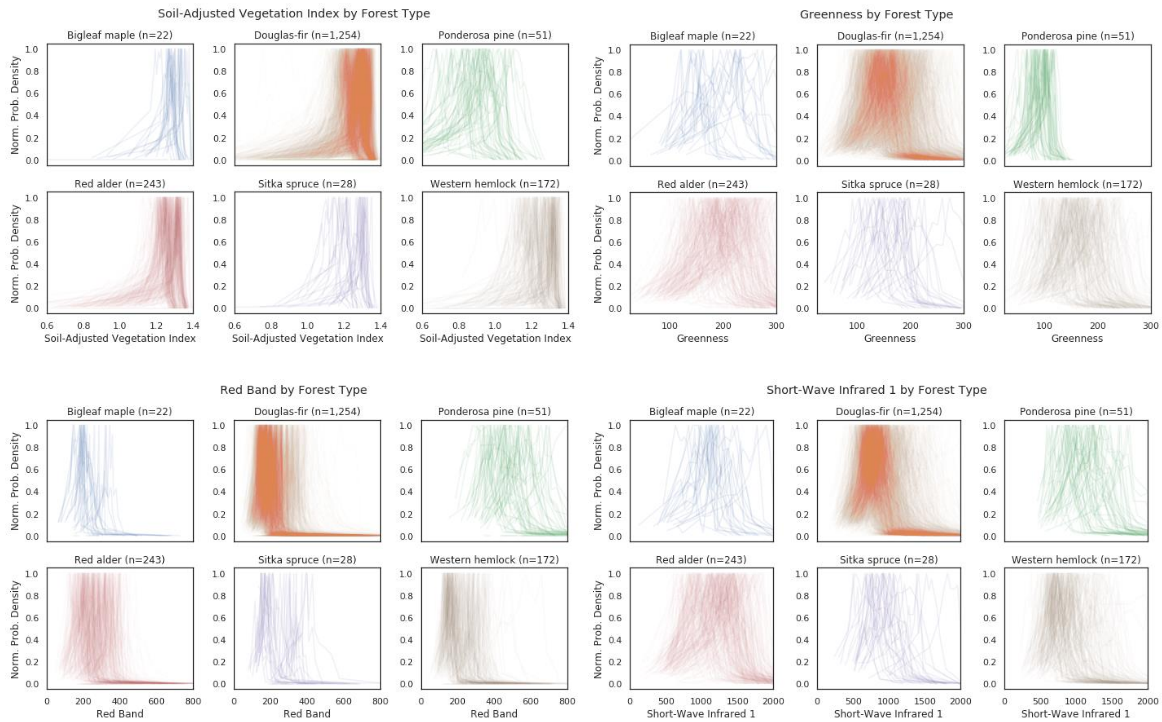
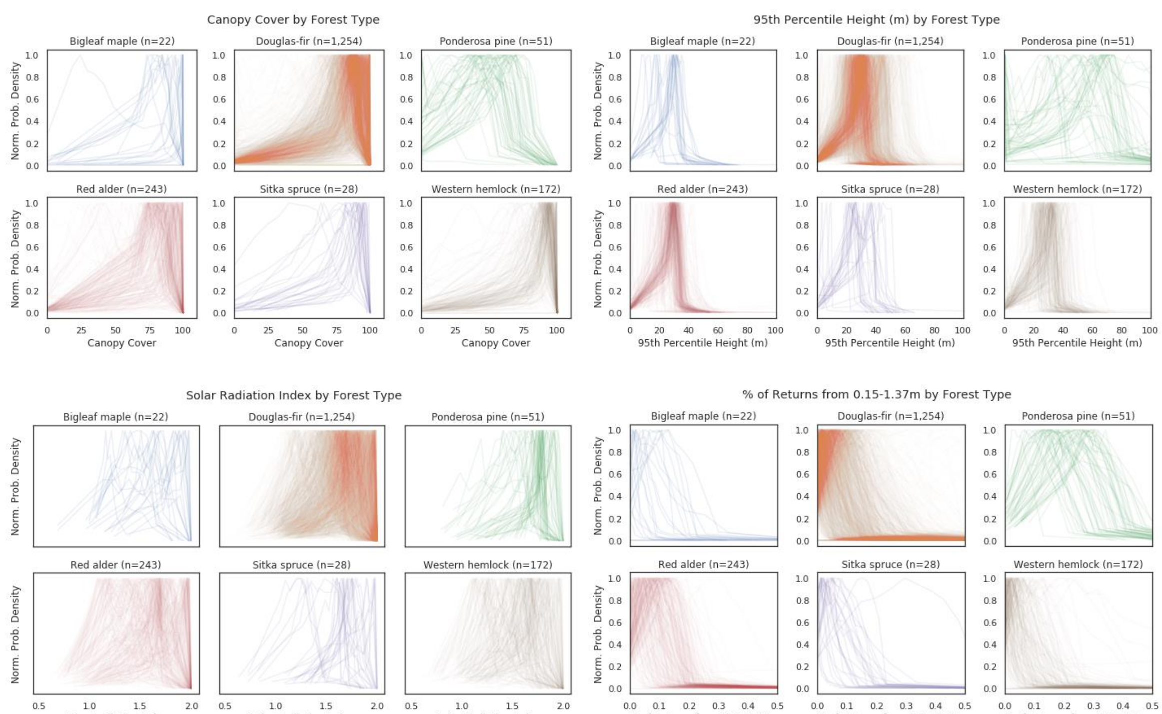


Figure 5: Distributions of Lidar Metrics by Forest Type



3.2. Tree Size and Structure

Figure 6: Distributions of Satellite Metrics by Tree Size / Growth Stage

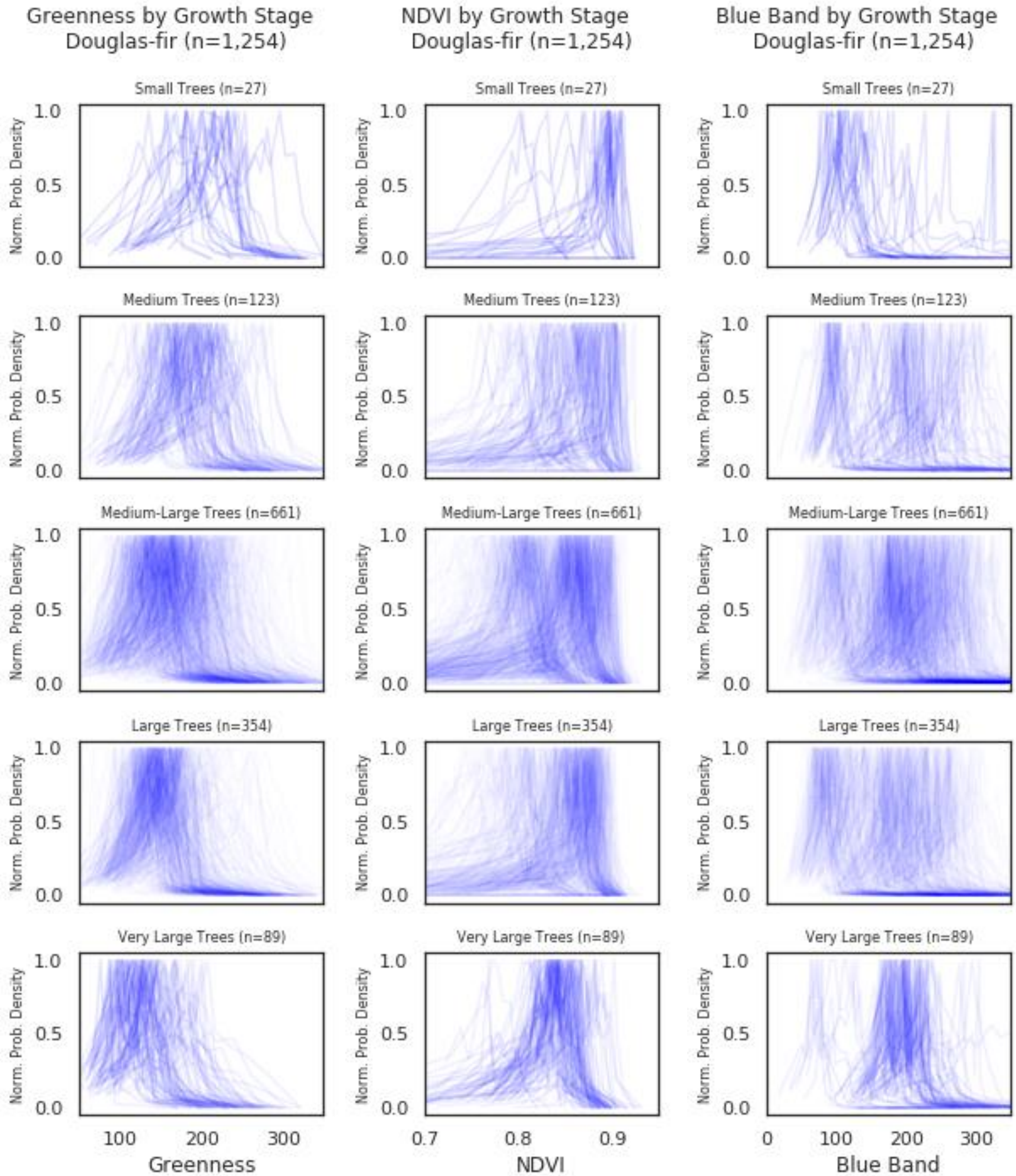
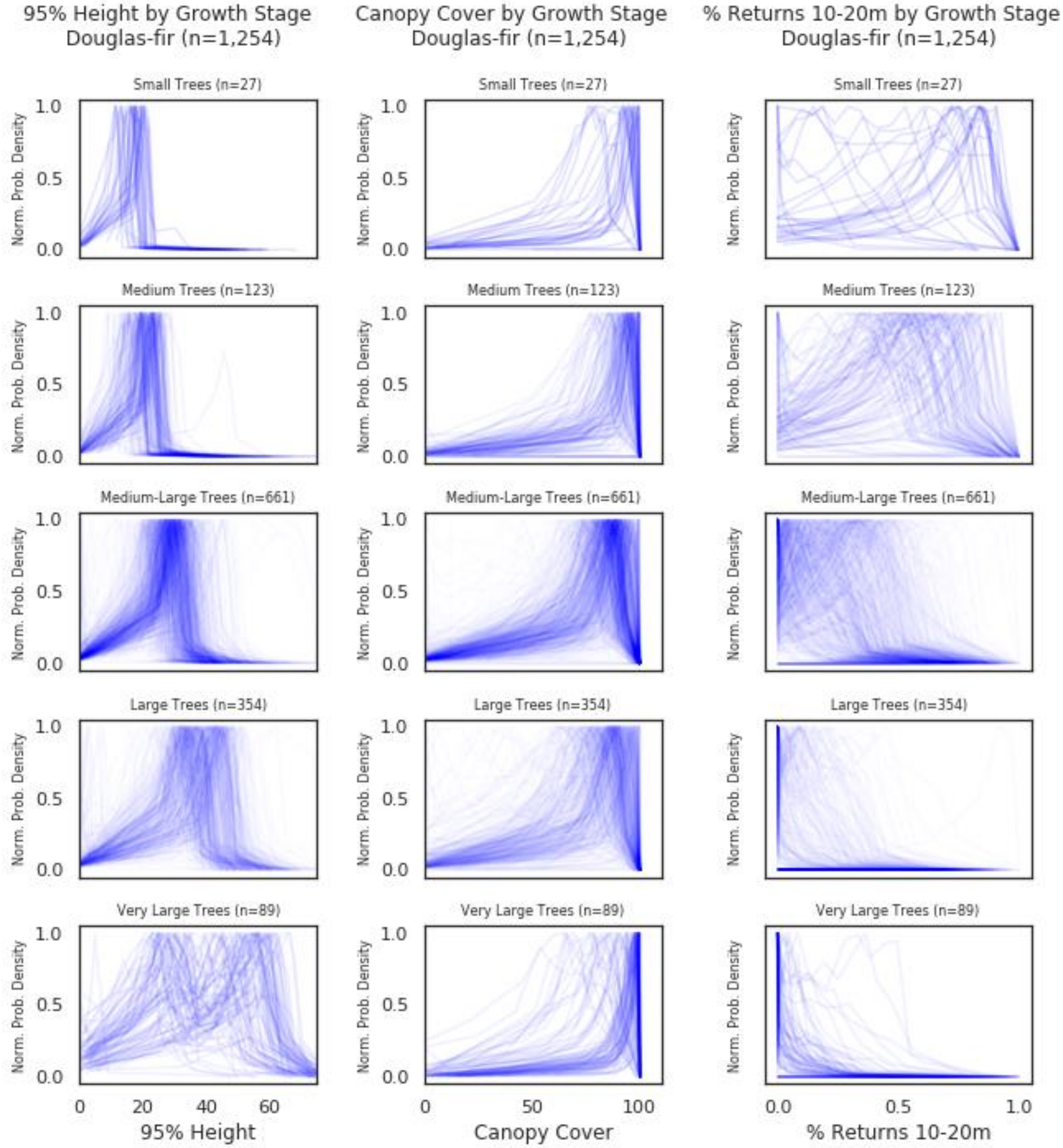


Figure 7: Distributions of Lidar Metrics by Tree Size/Growth Stage



4. DISCUSSION OF SHIFT FROM FIA TO STAND LEVEL INVENTORY DATA

As mentioned earlier in our description of the forest inventory data used for this work, we had originally proposed to use plot-level measurements collected on the US Forest Service's Forest Inventory and Analysis (FIA) program. We have shifted our approach to use stand delineations and inventory data that for which there are many more samples to use for modeling, and for which there are no confidentially or privacy restrictions as these data are in the public domain.

We believe this shift from an FIA to a stand-level inventory focus offers several advantages.

First, there is a larger number of training samples for our predictive models to learn from. For example, in the footprint of lidar data we have already processed, there are roughly 8,300 forest stands in the ODF dataset (with 43,320 plots measured), compared to an estimated 1,514 FIA plots. There are also additional stand-level inventory datasets we can use for this modeling task. We have already acquired a comparable stand-level inventory dataset from the Washington Department of Natural Resources (WA DNR) through a public records request that much larger than the data from ODF. For example, the total number of plots measured in the ODF dataset (not limited to lidar footprints we have processed) is 98,000, while the number of plots measured in the WA DNR dataset is 450,000. Similarly, there are 15,000 stands delineated by ODF, while there are 60,000 forest stands delineated by WA DNR. We are also in the process of gathering US Forest Service stand-level inventory data from National Forests through the FSVeg database.

Second, the plots describing the forest inventory data from these stand-level inventory datasets are clustered together geographically. These plots are accompanied by stand delineations, and there are multiple plots per stand which help depict variability of forest conditions in areas that are considered a single forest type by state and federal land managers. Together, the combination of higher spatial density and existence of stand delineation boundaries offer clear opportunities for the segmentation phase of our predictive modeling work.

Considering the larger number of samples and the geographic concentration and boundary information available from these stand-level datasets, we believe we will be able to develop more robust forest condition predictive model, and be able to publish the training data, fitted models, and forest type maps without any privacy/confidentiality constraints that would've been involved with the use of FIA data.

5. REFERENCES CITED

Crookston, N.L., Dixon, G.E., 2005. The forest vegetation simulator: A review of its structure, content, and applications. Comput. Electron. Agric. 49, 60–80. https://doi.org/10.1016/j.compag.2005.02.003

Dixon, G.E., 2017. Essential FVS: A user's guide to the Forest Vegetation Simulator (Internal Report). USDA Forest Service, Forest Management Service Center, Fort Collins, CO.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ., Big Remotely Sensed Data: tools, applications and experiences 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Hudak, A.T., Haren, A.T., Crookston, N.L., Liebermann, R.J., Ohmann, J.L., 2014. Imputing Forest Structure Attributes from Stand Inventory and Remotely Sensed Data in Western Oregon, USA. For. Sci. 60, 253–269. https://doi.org/10.5849/forsci.12-101

Wang, T., Hamann, A., Spittlehouse, D., Carroll, C., 2016. Locally Downscaled and Spatially Customizable Climate Data for Historical and Future Periods for North America. PLOS ONE 11, e0156720. https://doi.org/10.1371/journal.pone.0156720