Forest Ecology and Management 448 (2019) 509–527

Contents lists available at ScienceDirect

Forest Ecology and Management

journal homepage: www.elsevier.com/locate/foreco

Predicting understory vegetation structure in selected western forests of the United States using FIA inventory data

Michael A. Krebs[a,*], Matthew C. Reeves[b], L. Scott Baggett[c]

[a] USDA Forest Service, Rocky Mountain Research Station, 800 E. Beckwith Ave., Missoula, MT 59801, United States
[b] Human Dimensions Program, USDA Forest Service, Rocky Mountain Research Station, 800 E. Beckwith Ave., Missoula, MT 59801, United States
[c] USDA Forest Service, Rocky Mountain Research Station, 240 West Prospect, Fort Collins, CO 80526, United States

ARTICLE INFO

Keywords:
FIA
Shrub
Herb
Understory vegetation
Random Forests
Overstory

ABSTRACT

Understory vegetation structure and its relationship with forest canopies and site conditions are important determinants of carbon stocks, wildlife habitat, and fuel loading for wildland fire assessments. Comprehensive studies are needed to assess these relationships through the use of consistently collected field-based data. One approach to achieve this is to make use of preexisting forest inventory data to estimate understory vegetation height and cover from site and overstory attributes. In this study, overstory, understory, and abiotic data describing site conditions were obtained from over 6700 Forest Inventory and Analysis (FIA) fixed radius plots collected between 2000 and 2012 to assess how understory vegetation cover and height vary with overstory attributes and site characteristics. The focus was restricted to four common forest types including lodgepole pine (*Pinus contorta* var. *latifolia*), Douglas-fir (*Pseudotsuga menziesii*), ponderosa pine (*Pinus ponderosa*), and grand fir (*Abies grandis*) found on approximately 43 million hectares in the western United States. Random Forest regression classification trees were developed for cover and height of shrub and herb understories as a function of field-measured predictor variables. Separate analyses were undertaken for the Pacific Northwest (PNW) and the Interior West (IW) Forest Inventory and Analysis (FIA) regions. Models developed from the IW data generally performed better and the OOB (out-of-bag) percent variance explained varied from 8.08% for forb height to 39.24% for shrub height. For the PNW data, percent variance explained ranged from 13.82% for forb height to 27.4% for shrub height. Percent variance explained values were higher in all corresponding models for the IW than PNW, except for forb and grass height. Differences in model performance were smallest in the case of forb cover (27.17% vs. 26.15%) and greatest in the case of percent shrub cover (30.92% vs. 15.53%) for IW and PNW models, respectively. Cover models within each dataset performed better, on average, than their associated height models. The most influential variables for predicting understory cover and height were ones representing overstory conditions and conform to ecological expectation corroborated by many studies examining the influence of forest overstories on understory vegetation dynamics. Several variables, including aspect, slope, and stand disturbance and treatment, were not important and contrary to expectation. Predicting understory vegetation attributes to aid assessments of carbon, fuel, and wildlife habitat may be more generalizable across forests of the western U.S. using standardized national inventory data in conjunction with improved measurements.

1. Introduction

Numerous ecological studies have focused on describing how overstory, site, and understory vegetation interact with one another in forested environments throughout the United States (Ffolliott and Clary, 1982; McKenzie et al., 2000; Burton et al., 2014), in Canada (Hart and Chen, 2006), in China (Ahmad et al., 2019), in Africa (Ensslin et al., 2015) in Australia (Bauhus et al., 2001) and in Europe (Tonteri et al., 2016; Hedwall et al., 2013; Sigurdsson et al., 2005). It has been long understood that these interactions are both complex and dynamic, and have profound effects on the composition, structure, and productivity of understory vegetation. Considerable silvicultural research in the past century has revealed the effects of site and overstory tree density upon understory vegetation composition and production through various management interventions such as thinning, prescribed burning, and wildland fire. Enhancement of understory vegetation

* Corresponding author.
E-mail address: michael.krebs@mso.umt.edu (M.A. Krebs).

https://doi.org/10.1016/j.foreco.2019.06.024
Received 22 March 2019; Received in revised form 12 June 2019; Accepted 14 June 2019
0378-1127/

production has been achieved in many of these studies by strategic removal of overstory trees and site preparation through these management practices (Sabo et al., 2008; Bailey et al., 1998; Uresk and Severson, 1989; Hedwall et al., 2013; Tonteri et al., 2016; Bauhus et al., 2001).

Understanding the effects of site and overstory structure and density upon understory vegetation structure improves our understanding of understory fuel loading (Lydersen et al., 2015; Olson and Martin, 1981), wildlife habitat (Hagar, 2007), and carbon stocks (Gray and Whittier, 2014; Johnson et al., 2017; Sigurdsson et al., 2005) and sequestration (Campbell et al., 2018; Suchar and Crookston, 2010), all of which are functions of understory biomass production. Taken together, predicting how forest understory vegetation will change in response to proposed overstory management, disturbance, or natural succession is a paramount endeavor in modeling forest change through time and aiding landscape assessments of ecosystem properties.

Past efforts in modeling understory vegetation response to overstory and site characteristics have been limited to specific scales, forest types, and plant functional groups, and have often only focused on trees or shrubs (Kerns and Ohmann, 2004; Moeur, 1985; Johnson et al., 2017). Moreover, most studies describing understory vegetation have been conducted at local scales limiting the spatial scale of application. Only a few attempts have been made to use data collected at a national scale to derive equations useful for estimating understory vegetation performance across more extensive spatial domains. Gebert et al. (2008) used national inventory data to estimate understory forest fuels but cited the lack of available understory biomass equations for their accurate calculation. Another study predicted understory herbaceous cover using national inventory data for four forest types but only within a three county-wide region in Alabama (Joyce and Baker, 1987). Forest understory species richness has also been modeled and predicted using national inventory data but was restricted to coastal plain pine-dominated forests of the southeastern U.S. (Timilsina et al., 2013). One similarly broad-based study attempted to predict understory cover and biomass as a function of relative stand percent canopy cover using inventory data (Suchar and Crookston, 2010) but focused primarily on the role of climatic factors, disturbance, slope and elevation, and without additional overstory attributes. Another broad-based study (Johnson et al., 2017) utilized national inventory data of understory vegetation for estimation of U.S. carbon stocks in a novel approach incorporating both allometric and spatial models of stand and remotely sensed variables, but not as a direct function of field-measured site and tree-level overstory attributes. Given the paucity of regional to national-scale studies aimed at describing relationships between overstory and understory vegetation attributes and an increasing need to understand ecosystems, it is critical to evaluate the ability of such national datasets to reveal these relationships. The present study is unique in that it signifies both an attempt to utilize national forest inventory data to address such broad-scale questions and to fill an increasing need to develop ecosystem-, regional-, and even national-level assessment and analyses.

The difficulty in not utilizing national inventory data earlier for these ends has been, in large part, due to the lack of a national standard protocol by which to combine understory vegetation data over many forest types for broad-scale analyses (Schulz et al., 2008). Increased emphasis on the importance of understory vegetation inventory data in recent years has provided impetus for establishing more standardized measurements in order to permit consistent national and regional analyses of understory vegetation (Bechtold and Patterson, 2005). Improving measurement consistency in forest inventories is also taking place internationally, as the need for such information increases to address global-scale concerns, such as biodiversity and other effects of climate change (Winter et al., 2008). However, it is unknown whether these standardized datasets are useful for developing predictive equations that estimate understory attributes and what, if any, important relationships exist between such measurements of overstory and understory attributes collected in the field.

We attempt to address some of these needs by using national-scale forest inventory data to examine the relationship between overstory and site attributes with understory vegetation, and by identifying the most important variables from inventory data that contribute to predicting understory vegetation structure (and biomass) using measured attributes of height and cover. Both understory vegetation cover and height have been used in many studies to predict biomass and consequently fuel loading, net primary productivity, and quantification of wildlife habitat (Olson and Martin, 1981; Alaback, 1986; Verschuyl et al., 2018; Ohmann et al., 1981). Based on the extensive historical literature on this topic, we expect that site differences and changes in overstory conditions will similarly affect understory vegetation structure. Specifically, we expect that increases in overstory volume, density, and canopy cover will be associated with decreased height and cover of understory vegetation across forest types and that disturbance, treatment, and site conditions will be influential in determining understory vegetation structure.

2. Methods

2.1. Study area

The geographic scope of this study includes four major forest types common to the western United States spanning nearly 43 million hectares of the Pacific Northwest and Intermountain West regions (Ruefenacht et al., 2008). They include lodgepole pine (*Pinus contorta* var. *latifolia*), Douglas-fir (*Pseudotsuga menziesii* (including interior and coastal varieties)), ponderosa pine (*Pinus ponderosa*), and grand fir (*Abies grandis*) forest types (Eyre, 1980), all of which encompass a broad range of environmental and vegetative conditions (Fig. 1).

2.2. Databases

Forest Inventory and Analysis (FIA) National Program Phase 2 inventory data (O'Connell et al., 2013) were used for this study as they contain a large and robust assemblage of measured abiotic, overstory, and understory vegetation attributes appropriate to the task of understanding how understory vegetation may relate to site and overstory components. FIA data were obtained from two sources covering the western contiguous United States. First, plot data from the Pacific Northwest (hereafter PNW) FIA Region (California, Oregon, and Washington) were obtained directly through FIA as a ready-made MS Access database called the 2011 PNW FIADB Annual Inventory Database (http://www.fs.fed.us/pnw/rma/fia-topics/inventory-data/index.php) (USDA Forest Service, 2013). Second, plot data from the Interior West (hereafter IW) FIA Region (Arizona, Colorado, Idaho, Montana, New Mexico, Nevada, Utah, and Wyoming) were obtained directly through their public website (USDA Forest Service, 2018) (available at: https://apps.fs.usda.gov/fia/datamart/datamart.html) (Fig. 2).

2.3. Plot description

Data collected from the current national standard FIA plot design were chosen for their measurement consistency and ease of scalability. The current design consists of four circular subplots with three subplots concentrically arranged around a single center subplot at a predetermined distance and azimuth (Appendix Fig. 1). Each subplot has a 7.3 m radius (168 m^2 or 1/59th ha) where trees greater than or equal to 12.7 cm diameter at breast height (DBH = 1.37 m) are individually tagged and sampled. Within each subplot is a nested 2.1 m radius microplot (13.5 m^2 or 1/740th ha) for measurement of seedlings less than 2.54 cm DBH and saplings between 2.54 and 12.5 cm DBH. Understory vegetation cover, composition, and canopy layer height, including all seedling, sapling, and mature trees (i.e. tally trees), are measured separately on each entire subplot in an ocular fashion. For very large

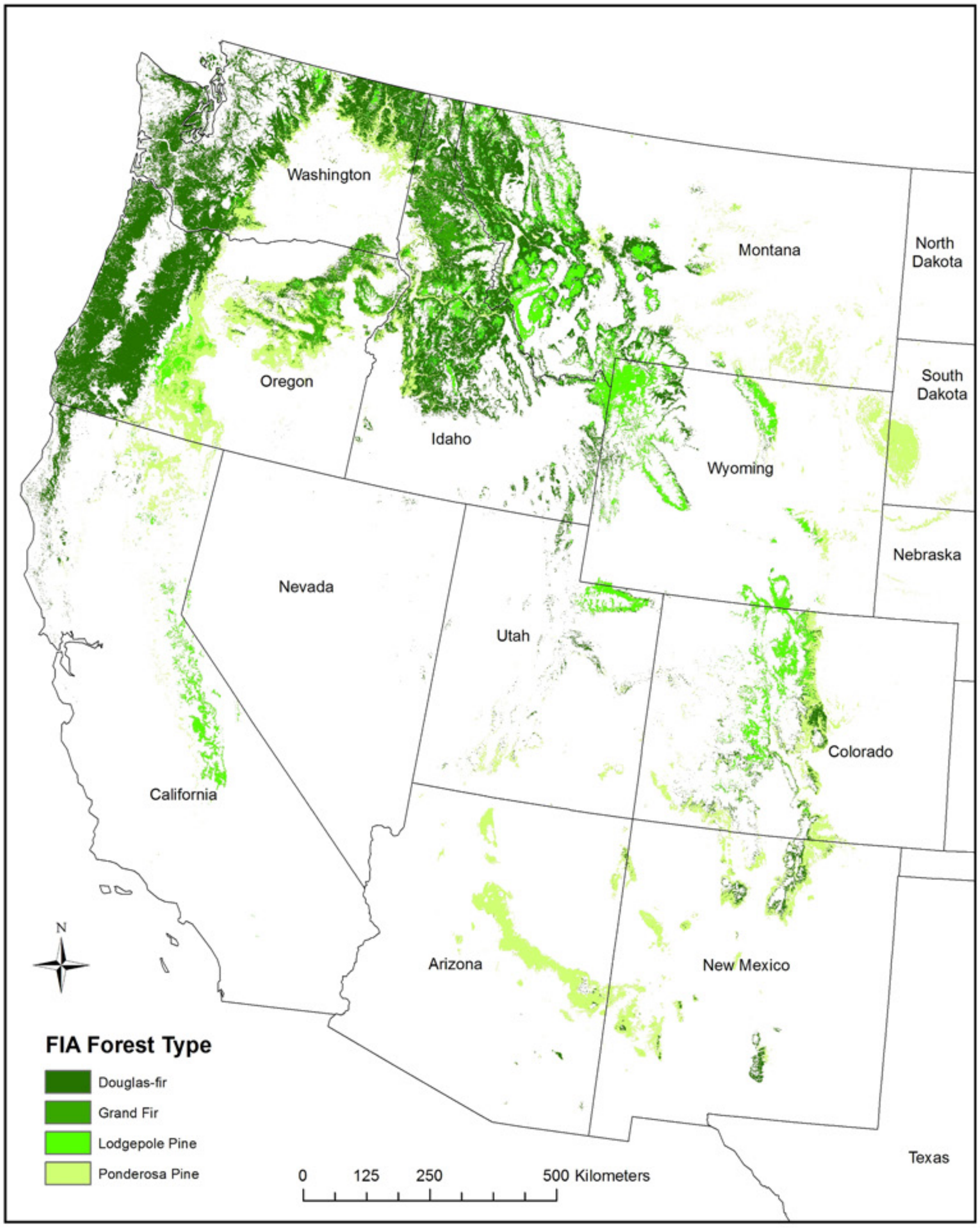


Fig. 1. Spatial extent of the four Forest Inventory and Analysis (FIA) forest types selected for analysis in the western U.S. Total area represented is approximately 42.9 million hectares. Douglas-fir extent is approximately 21.6 million hectares; ponderosa pine - 12.0 million hectares; lodgepole pine - 7.4 million hectares; and grand fir - 1.8 million hectares (Ruefenacht et al., 2008). Selected FIA plots are located within these areas except for South Dakota and Nebraska, which fall within the Northern FIA Region.

diameter trees exceeding a predetermined diameter threshold, an optional 17.9 m radius (1012 m^2 or 1/10th ha) macroplot is also used (USDA Forest Service, 2014). The total area available for sampling, including the macroplot, is approximately 4050 m^2 (or 0.4 ha).

2.4. Data selection

The historical range for all available data used in this study was from 2000 to 2012, with the number of data collection years for most states varying between nine and thirteen years. The notable exception here is with the PNW data, in which only one year could be used (2011) due to inconsistent understory vegetation measurement protocols (described below).

Inventory plots were further delineated comprising only one forested condition, that is, plots containing single uniform stand conditions such as tree density, forest type, stand size class, and ownership, etc. This reduced variability and minimized any edge effects that might be present with more than one forested condition on a given plot. Plots with recently cleared overstories and with complete and consistent vegetation measurements on all four subplots were also included. A total of 6,716 plots in eleven western states were retained for further analysis (Fig. 2 and Table 1).

2.5. Independent/dependent variables

We used all measured plot variables that directly described both overstory and site conditions demonstrated in many previous studies to influence understory vegetation patterns. These included plot-, condition-, subplot-, and tree-level measurements that FIA routinely collects at these measurement scales on each plot. Data tables from both FIA regional databases were then further screened for attribute and measurement consistency. Twenty-eight measured attributes were

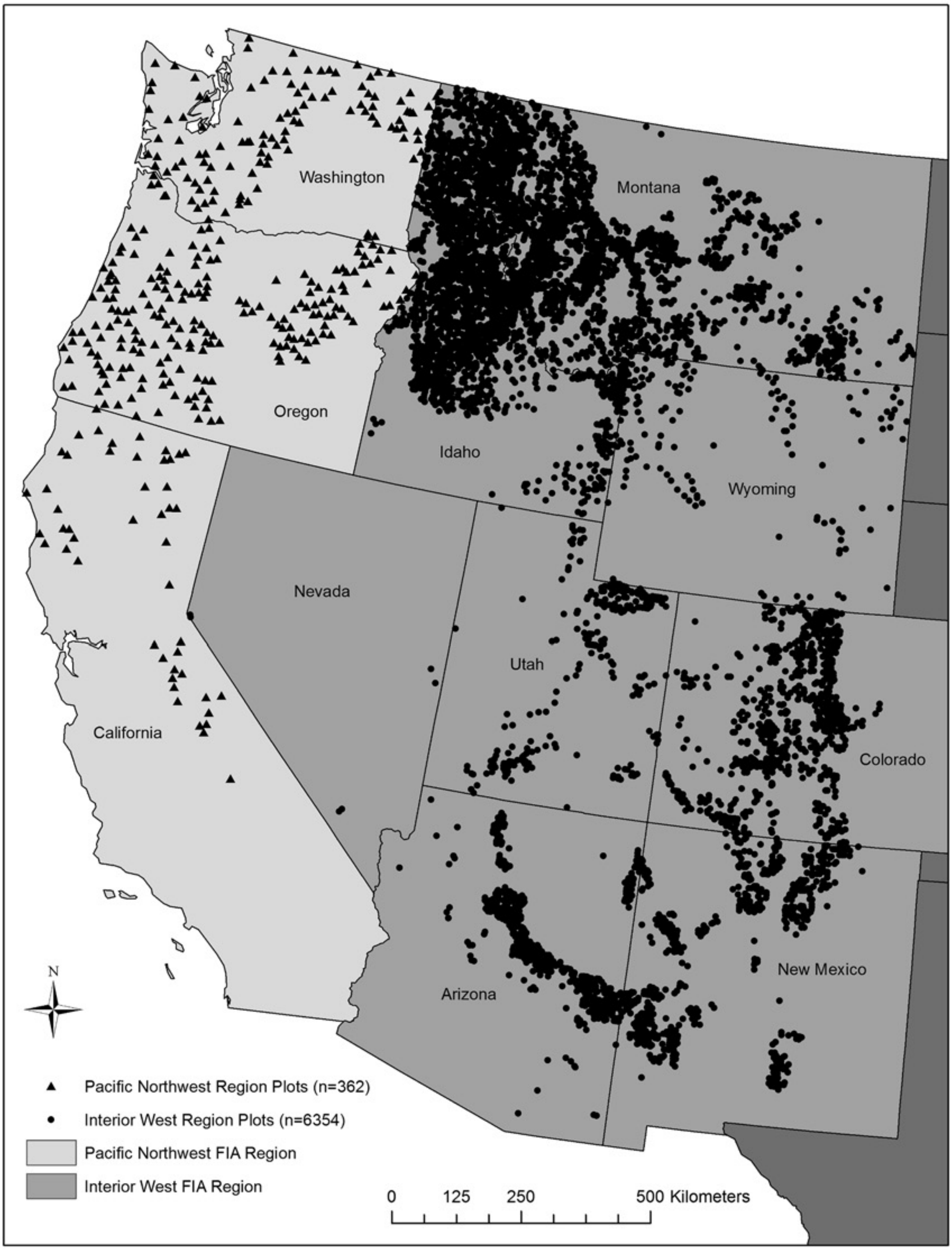


Fig. 2. Distribution of Phase 2 Forest Inventory and Analysis (FIA) plots for both Interior West (IW) and Pacific Northwest (PNW) datasets selected for the present study.

identified as important independent variables (Table 2) for predicting understory vegetation growth and abundance. These included several regional variables collected separately within each of the two FIA regions. In total, 28 separate independent variables, 27 for the IW dataset and 25 for the PNW dataset, were considered for analysis (Table 2).

For the dependent variables, understory vegetation height and percent canopy cover of shrub, grass, and forb lifeform components were used (Table 3). FIA collects vegetation height and cover for these three commonly recognized lifeforms over each entire subplot, both in an ocular fashion. However, due to inconsistent measurement protocols of understory vegetation for most years of the PNW data (2001–2011), only the last year of data (2011) that was provided could be used, since national measurement protocols for understory vegetation were not adopted until then for all FIA units (Patterson and O'Brien, 2011). This resulted in only 362 single condition PNW plots available for analysis (Table 1). This was not the case with the IW-FIA data, where all understory vegetation was either consistently measured or revised appropriately according to national protocols to allow for between-year comparisons.

Table 1
FIA single condition plot summary for Interior West (IW) and Pacific Northwest (PNW) datasets (2011 data for PNW only). All single condition plots contain four accessible subplots where vegetation was equally measured on each. Forest types are taken from those defined by FIA and the Society of American Foresters (SAF).

	FIA/SAF Forest Type	# Plots (single condition)	# Subplots (single condition)	# Plots with live overstory and measured understory vegetation[1]	Total # plots[2]
Interior West (2000–2012)					
Douglas-fir	201	2529	10,116	2438	2529
Ponderosa pine	221	2001	8004	1910	2001
Grand fir	267	332	1328	327	332
Lodgepole pine	281	1492	5968	1359	1492
Total	–	6354	25,416	6034	6354
Pacific Northwest (2011 only)					
Douglas-fir	201	203	812	191	203
Ponderosa pine	221	95	380	85	95
Grand fir	267	27	108	27	27
Lodgepole pine	281	37	148	35	37
Total	–	362	1448	338	362
Combined Total					6716

[1] denotes plots where vegetation was fully measured in all four subplots.
[2] total number of plots for each dataset include those with both live overstories and no overstories (e.g. recently harvested) and with complete vegetation measurements taken over all four subplots.

2.5.1. Scaling variables

1) **Independent variables** – After selection of all predictor variables, it was determined that the plot level was the most appropriate sampling unit by which to conduct analyses. This meant that all variables measured at scales below this required upward scaling and thereby included those that were collected at the individual tree- and subplot-levels. In most cases, scaling was achieved by simply taking the mode, average, or sum over all tree and subplot measurements within a given plot. However, there were several special cases encountered where summarizing measured attributes proved more cumbersome, such as aspect and habitat type (Appendix Table 1, Table 2).
2) **Response variables** – Understory vegetation measured on each subplot also needed to be scaled up to the plot level. FIA collects the percent cover for each lifeform at each of five individual vertical layers. The first four layers are separate height classes (0–61, 62–183, 184–488, > 488, in cm) while the fifth layer represents an aerial percent cover estimate. In determining estimates of average plot lifeform cover, the aerial cover estimate associated with each lifeform for each subplot was used; which is a top-down, overlapping view of the vegetation cover throughout all of the layers. By definition, aerial cover cannot exceed the sum of all cover measured for individual height layers, but can be less. Average plot lifeform cover is then obtained by averaging the aerial percent covers for each lifeform over all four subplots.

For calculation of average plot lifeform height over all four subplots, the height of each lifeform was estimated using cover-weighted heights. This is obtained by first calculating a lifeform's relative cover for each height class layer by summing the total lifeform covers over all four height layers on each subplot and then determining the relative proportion of cover for each individual height layer. The resulting relative cover proportions, which always sum to one, are then multiplied by their respective height class midpoints (i.e. 30, 124, 337, and 488 cm) to obtain its cover-weighted height. These weighted heights are then summed together for each lifeform and averaged over each subplot to obtain an overall plot mean.

In the end, six variables representing height and percent cover for each lifeform (shrub, forb, and grass) were calculated separately in this way for each dataset (IW and PNW), consisting of: a) plot average cover-weighted height and b) plot average percent cover for each of the three non-tree lifeforms measured (Table 3). This resulted in twelve total dependent variables, six for each dataset.

2.6. Analytical methods

We adopted a regression tree classification approach using Random Forests (Breiman, 2001; Liaw and Wiener, 2002) since we wanted to identify important variables and assess their ability and accuracy in predicting understory cover and height from a large number of categorical and numerical predictors. In addition, Random Forests (hereafter RF) does not make any distributional assumptions about the response and predictors nor of the nature of their relationship to one another (Cutler et al., 2007). It is also robust against outliers and prevents the overfitting of data (Breiman, 2001). Lastly, it is capable of imputing missing data values, which is a common occurrence in many ecological datasets.

The understory response variables representing percent cover were expressed as proportions prior to modeling. Height response data were not transformed and were modeled in their original scale. It was not necessary to screen data obtained from both FIA sources prior to analysis, since FIA implements a stringent quality control process for each field measurement (Pollard et al., 2006). Thus, all data were used in their original form, including imperial units. All results were then converted to common metric units.

All statistical analyses were done using R version 3.1.2 (R Development Core Team, 2014) including the following libraries: (a) *caret*, (b) *doParallel*, (c) *ggplot2*, (d) *ggRandomForests*, (e) *gplots*, (f) *gridExtra*, (g) *parallel*, (h) *randomForest*, (i) *randomForestSRC*.

2.6.1. Random Forest parameters/missing data

We implemented the RF algorithm for regression in R using the *randomForestSRC* package (Ishwaran and Kogalur, 2017). The default parameter for the number of trees (*ntree*) was kept at 1000, since results varied little after making several runs and increasing and decreasing the number of trees. Another commonly modified parameter in RF is the number of variables to be used to split a node (*mtry*), where we implemented a training step in R for each model to optimize the number of variables to be tried at each split in an individual tree. This alleviated the need to specify a predefined number of variables beforehand and was tantamount to a 10-fold cross validation.

Missing data were imputed using a modification of the missing data algorithm from Ishwaran et al. (2008), in order to utilize all observations. Imputation for continuous variables are determined by a mean rule and categorical variables via a modal rule. This method is referred to as "on the fly imputation", where data is imputed simultaneously while growing a regression classification tree (Tang and Ishwaran, 2017).

Table 2
Independent variables and their description (O'Connell et al., 2013) for IW and PNW datasets measured from FIA Phase 2 plot measurements. Scaling method refers to the calculation used to upwardly scale a variable to the plot level. NA denotes that no scaling to the plot level was needed and/or the actual measured values were used.

#	Variable name	Description	Variable type	Scaling method
1	Mean tree height	Average of actual tree height (ground to existing tip plus estimated lengths of broken tops) (in meters). Averaged for measured trees greater than or equal to 2.5 cm DBH.	Continuous	Average
2	Mean tree age	Average of tree breast height age, in years, collected for a subset of measured live trees greater than or equal to 2.5 cm DBH for each species, diameter class, and crown class (PNW).	Continuous	Average
3	Mean crown ratio	Average of tree compacted crown ratio, in percent, for measured live trees greater than or equal to 2.5 cm DBH. Crown ratio is the percent of the tree bole supporting live, healthy foliage (the crown is ocularly compacted to fill in gaps) when compared to actual length (ACTUALHT).	Continuous	Average
4	Mean tree diameter	Average diameter (in cm) of all measured live trees greater than or equal to 2.5 cm DBH.	Continuous	Average
5	Mean slope	Mean slope for plot after averaging over each subplot (n = 4) (in percent).	Continuous	Average
6	Mean uncompacted live crown ratio (IW only)	Average of uncompacted live crown ratio, in percent. Measured for sampled live trees greater or equal to 12.5 cm DBH and determined by dividing the live crown length by the actual height. This is a measure of tree crown vigor.	Continuous	Average
7	Crown class	Mode of tree crown class for measured trees greater than or equal to 2.5 cm DBH. Crown class code indicates the position of the tree within the canopy (e.g. open grown, dominant, codominant, intermediate, and overtopped) and consequently the amount of sunlight received.	Categorical	Mode
8	Number of trees	Number of sampled trees per plot greater than or equal to 2.5 cm DBH and used for calculation of tree level attributes. Only live trees are considered.	Continuous	Sum
9	Primary stand disturbance	Indicator variable for disturbance code # 1 for the condition and recoded as presence (1) and absence (0) of disturbance. Disturbance code (DSTRBCD1) is populated for all forested plots and, as collected in the field, indicates the kind of disturbance occurring since the last measurement or within the last 5 years for new plots. The area affected by the disturbance must be at least 0.4 ha in size.	Categorical	NA
10	Secondary stand disturbance	Indicator variable for disturbance code # 2 for the condition and recoded as presence (1) and absence (0) of disturbance. Disturbance code (DSTRBCD2) is populated for all forested plots and, as collected in the field, indicates the kind of disturbance occurring since the last measurement or within the last 5 years for new plots. The area affected by the disturbance must be at least 0.4 ha in size.	Categorical	NA
11	Primary stand treatment	Indicator variable for stand treatment code # 1 recoded as presence (1) and absence (0) of treatment. Treatment code 1 (TRTCD1) indicates the type of stand treatment for the condition that has occurred since the last measurement or within 5 years for new plots. The area affected by the treatment must be at least 0.4 ha in size.	Categorical	NA
12	Secondary stand treatment	Indicator variable for stand treatment code # 2 recoded as presence (1) and absence (0) of treatment. Treatment code 2 (TRTCD2) indicates the type of stand treatment that has occurred for the condition since the last measurement or within 5 years for new plots. The area affected by the treatment must be at least 0.4 ha in size.	Categorical	NA
13	Elevation	Plot elevation above sea level (in meters) (NAD 83 datum).	Continuous	NA
14	Predominant diameter class (measured)	Field stand-size class code (6 codes) for the condition. A coded field measure of predominant diameter class of live trees.	Categorical	NA
15	Forest type	FIA forest type code for the condition that is assigned in the field and based on the tree species or species groups forming a plurality of all live stocking (201 = Douglas-fir, 221 = ponderosa pine, 267 = grand fir, 281 = lodgepole pine).	Categorical	NA
16	Habitat series	Forest habitat series for the condition. This represents the truncation of habitat type (series/type/phase) to the series level only and measured for the condition. This was done to increase the number of observations per habitat type level.	Categorical	NA
17	Percent live canopy cover (IW only)	The percentage of live tree canopy cover for the condition including live tally trees, saplings, and seedlings.	Continuous	NA
18	Percent live and missing canopy cover (IW only)	Live (above) plus estimated missing tree canopy cover for the condition of missing live and dead tally trees, saplings and seedling due to disturbance, treatment, etc., based on observation, stand history, and historical aerial imagery. This percentage cannot exceed 100%.	Continuous	NA
19	Ownership	Ownership code for the condition. See FIA National Core Field Guide for definitions (USDA 2014).	Categorical	NA
20	Physiographic class	Physiographic class code for the condition within xeric, mesic, and hydric categories. It is a coded measure of available moisture to stands as affected by topographic landform.	Categorical	NA
21	Site index	Site index for the condition using 50, 80 or 100 years as the base age. Site index is the estimated /expected average height of dominant/codominant trees at the specified base age.	Continuous	NA
22	Site productivity	Site productivity class code (7 classes). This measure identifies the potential growth (in $m^3\ ha^{-1}\ yr^{-1}$) and is based on the culmination of mean annual increment of fully stocked natural stands. The classes 1–7 are ordered in terms of decreasing potential growth.	Categorical	NA
23	Predominant diameter class (computed)	Stand-size class code (4 codes) for the condition. A coded classification of the predominant diameter class of live trees in the condition using an algorithm.	Categorical	NA
24	Total basal area	Sum of basal area per tree over all subplots of measured live trees (in square meters) greater than or equal to 2.5 cm DBH.	Continuous	Sum
25	Seedling trees per hectare	Sum of number of seedling trees per hectare (unadjusted) from microplot (13.5 m^2 or 1/740 ha) measurements where one tree = 30 trees per hectare).	Continuous	Sum
26	Trees per hectare	Trees per hectare (unadjusted) for all live trees 2.5 cm or greater DBH that are measured on microplots, subplots, and macroplots (PNW only), and summed over all trees for plot-level estimate.	Continuous	Sum
27	Gross tree volume	Sum of gross tree cubic volume, in cubic meters, for each sample tree greater than or equal to 12.5 cm DBH, with rotten, cull or defected wood.	Continuous	Sum
28	Topographic position (PNW only)	Plot topographic position on landscape (9 classes).	Categorical	NA

Table 3
Dependent variables for IW and PNW datasets calculated from subplot-level FIA understory vegetation measurements. See methods for description of cover-weighted height calculations. Model names for each dataset are also based on the response variables listed below.

#	Variable name	Description	Variable type	Scaling method
1	Forb cover	Mean percent cover of all forb species found on plot based on the average aerial cover layer for each subplot (n = 4).	Continuous	Average
2	Forb height	Mean cover-weighted height of forbs, in meters, based on the cover-weighted height for each subplot (n = 4) and averaged for each plot.	Continuous	Average
3	Grass cover	Mean percent cover of all grass species found on plot based on the average aerial cover layer for each subplot (n = 4).	Continuous	Average
4	Grass height	Mean cover-weighted height of grass, in meters, based on the cover-weighted height for each subplot (n = 4) and averaged for each plot.	Continuous	Average
5	Shrub cover	Mean percent cover of all shrub species found on plot based on the average aerial cover layer for each subplot (n = 4).	Continuous	Average
6	Shrub height	Mean cover-weighted height of shrubs, in meters, based on the cover-weighted height for each subplot (n = 4) and averaged for each plot.	Continuous	Average

Table 4
Summary Random Forest model statistics for IW and PNW datasets. Model performance statistics are based on the out-of-bag (OOB) prediction dataset for each model. Model names are based on the response variables described in Table 3.

		IW (n = 6354)	PNW (n = 362)
Model		Measured Variables	Measured Variables
	Number of predictors	27	25
Forb Cover	% variance explained	27.17	26.15
	OOB prediction error	0.0063	0.0262
Forb Height	% variance explained	8.08	13.82
	OOB prediction error	0.2386	0.1448
Grass Cover	% variance explained	27.27	19.53
	OOB prediction error	0.0115	0.0194
Grass Height	% variance explained	10.42	19.22
	OOB prediction error	0.2409	0.1496
Shrub Cover	% variance explained	30.92	15.53
	OOB prediction error	0.0143	0.052
Shrub Height	% variance explained	39.24	27.4
	OOB prediction error	1.052	2.967

2.6.2. Model assessment and variable importance

The performance of the RF regression model can be examined as the percent variance explained for the out-of-bag (OOB) portion of the dataset. Models containing higher percent variance explained values account for more variation in their data and hence represent better model performance and fit. This is analogous to a traditional R^2 but since the out-of-bag predictions are averaged over a subset of the data (roughly one-third) from which the percent variance explained is calculated, it is a less forgiving and more pessimistically biased measure than R^2; which considers all of the data in other more traditional statistical approaches (James et al., 2013). Moreover, this aspect of model performance is itself a means of cross-validation, since the OOB portion of the RF model is never used for the development of the actual model, but instead is used to test its predictive accuracy (Cutler et al., 2007; Hastie et al., 2009).

Variable importance was calculated using the Breiman-Cutler permutation method as described in Breiman (2001). This is first computed by recording the OOB prediction error for each tree (described above). Then, for each OOB case, a given variable is randomly permuted (i.e. removed) before it is dropped through the tree grown for the in-bag data. This is repeated for each tree and a new OOB estimate of prediction error is calculated. Variable importance is the difference between this value and the estimate of OOB prediction error without the permutation. Thus, large positive values indicate that the variable is predictive, while zero or negative values indicate no predictive value (Ishwaran et al., 2008).

2.6.3. Model prediction

Marginal dependence plots (also called variable dependence plots) were generated for each variable in each model. Such plots may be used to graphically characterize relationships between individual predictor variables and predicted responses of cover and height obtained from the RF regression algorithm. Marginal dependence relationships of variables give an idea of the overall trend between a given response-predictor by showing the predicted response as a function of a covariate of interest. They depict the nature of the relationship of predictor variables as a function of the predicted response not averaging out the effects of the other predictors, as with partial dependence plots (Cook and Weisberg, 1997). In other words, it is the estimated effect of a predictor upon the predicted response not considering the effects of the other predictors in the model. Variable dependence plots can only be generally interpreted, since each predicted response is dependent on the full complement of covariates in the model. However, isolating the most important variables within a model can illustrate how the RF predictors respond in regard to specific variables of interest (Ehrlinger, 2015).

3. Results

3.1. Random Forest models

All twelve final RF regression models exhibited considerable stability in performance, mean prediction error, and rankings of variable importance (Table 4, Fig. 3). In addition to the measured variables (Table 2), we also considered several derived measures of aspect (see Appendix Table 1; Zar, 1999), which we thought would prove influential in predicting understory vegetation structure. However, including these in several combinations resulted in little to no improvement in either model performance or reduction in prediction error (not shown). Moreover, all aspect measures within all models were among the very least important variables in terms of their predictive power. Therefore, we excluded these variables from our final models.

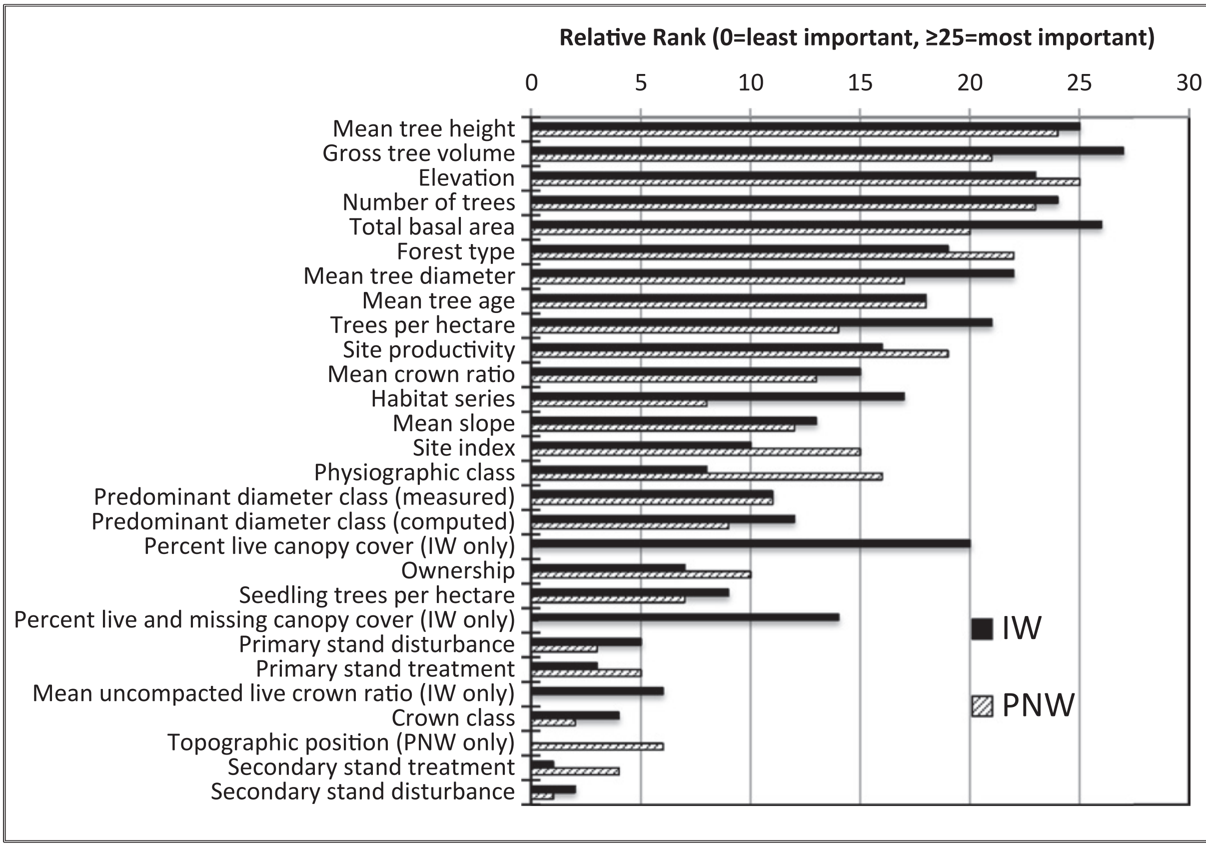


Fig. 3. Combined mean rank variable importance values for the IW and PNW datasets. Relative rank represents the ranked average importance of each variable over all models for each dataset. Variables with higher ranks reflect the most important variables and variables with similar ranks for each dataset reflect similar importance of variables. Note the difference in maximum ranks between the two datasets, which reflects the number of model variables evaluated (i.e. 27 for the IW and 25 for the PNW). See Table 2 for variable descriptions.

3.2. Model performance

Overall, the IW models performed better (i.e. better model fit and predictive ability) for each corresponding lifeform cover and height, except for forb and grass height (Table 4). Average model performance between datasets in terms of percent variance explained was higher overall across IW models (23.85%) than with PNW models (20.28%). For the IW models, percent variance explained (a measure of model fit) ranged from 8.08% for forb height to 39.24% for shrub height. For the PNW models, percent variance explained ranged from 13.82% for forb height and 27.4% for shrub height.

Between corresponding models of each dataset, the differences in percent variance explained were smallest for forb cover (1.02%) and greatest for shrub cover (15.39%), both greater with the IW data. Both forb and grass height for the PNW data represented better fits by 5.74 and 8.8 percent than their IW counterparts, respectively. Overall, both IW and PNW shrub height models represented the best model fits with the highest percent variance explained, where the IW model explained nearly 12% more of the variance in shrub height than the PNW. The forb height models had the lowest percent variance explained, where the PNW model explained nearly 6 percent more of the variance. Overall, cover models from both datasets performed better on average than their corresponding height models, except in the case of shrub height. Greater model performance (and hence model fit) of all models was also demonstrated through lower OOB prediction errors, where the smallest mean prediction errors were observed, in tandem, with the highest values of percent variance explained.

3.3. Patterns of permutation-based variable importance

Clear patterns of variables and their relative importance were found within all models for both IW and PNW datasets (Fig. 3, Table 5), where often the same predictors appeared as the most important and least important in terms of their contribution to predictive accuracy. In nearly all models, the most important variables were most easily discerned within the top 5–10 by their relatively greater importance values. Below this approximate number, importance values tended to converge and more slowly diminish in their predictive contribution, making the top ten most important variables a convenient cutoff for closer comparison of models both within and between datasets (Table 5). Similarly, many of the same variables appeared as least important across all models and a similar cutoff would also allow for a convenient comparison (Fig. 3). However, these are only briefly discussed in the aggregate over all models and we instead focus on the ten most important variables, both in terms of their actual permutation importance and, later, their ranked importance.

For all twelve models combined across each dataset, the top ten variables accounted for between 69 and 86 percent of the permutation importance of all of the variables in each model (Table 5). The top five variables in each model across both datasets alone represented between 40 and 60 percent of the total permutation importance.

For the PNW data, the top ten variables in each model represented between 72 and 86 percent of the total permutation importance (Table 5). The top five most important variables for each model alone represented roughly between 46 and 60 percent of the total permutation importance and included, among others: elevation, mean tree height, forest type, total basal area, and gross tree volume. These latter variables appeared most often as the top five most important variables

Table 5
Top ten most important individual predictors (in descending order of importance) within cover and height models for the IW and PNW datasets. Percent relative importance refers to the permutation importance of each variable relative to all other variables within each model. Regional variables are indicated with an *. See Table 2 for variable descriptions. Model names are based on the response variables described in Table 3.

Model	Rank	IW Variable	% Relative Importance	PNW Variable	% Relative Importance
Forb Cover					
	1	Gross tree volume	8.70%	Site productivity	21.10%
	2	Mean tree height	8.70%	Site index	9.80%
	3	Total basal area	8.00%	Elevation	9.70%
	4	Number of trees	7.90%	Mean tree height	9.50%
	5	Habitat series	7.00%	Mean tree diameter	7.50%
	6	Mean tree diameter	6.30%	Number of trees	5.60%
	7	Trees per hectare	5.90%	Mean tree age	5.40%
	8	*Percent live canopy cover	5.70%	Gross tree volume	5.30%
	9	Mean tree age	5.60%	Physiographic class	5.20%
	10	Elevation	5.40%	Forest type	3.90%
Forb Height					
	1	Gross tree volume	11.20%	Mean tree diameter	11.00%
	2	Total basal area	11.20%	Total basal area	10.90%
	3	Number of trees	10.50%	Gross tree volume	10.90%
	4	Mean tree height	8.80%	Mean tree age	10.80%
	5	Mean tree diameter	7.20%	Elevation	10.30%
	6	Trees per hectare	7.10%	Mean tree height	10.10%
	7	*Percent live canopy cover	6.90%	Site productivity	7.10%
	8	*Percent live and missing canopy cover	4.30%	Number of trees	6.40%
	9	Mean tree age	4.20%	Trees per hectare	5.40%
	10	Site productivity	4.20%	Forest type	3.40%
Grass Cover					
	1	Total basal area	10.80%	Number of trees	21.30%
	2	Number of trees	10.30%	Physiographic class	15.90%
	3	Gross tree volume	9.60%	Site productivity	7.70%
	4	Trees per hectare	7.10%	Forest type	6.50%
	5	*Percent live canopy cover	6.00%	Total basal area	6.40%
	6	Elevation	5.90%	Gross tree volume	6.30%
	7	Mean tree diameter	5.60%	Mean tree height	5.90%
	8	Mean tree height	5.40%	Elevation	5.80%
	9	Habitat series	4.70%	Mean tree diameter	4.80%
	10	Mean slope	3.90%	Trees per hectare	4.10%
Grass Height					
	1	Gross tree volume	12.00%	Number of trees	18.50%
	2	Total basal area	11.20%	Elevation	15.50%
	3	Number of trees	11.20%	Forest type	9.90%
	4	Mean tree diameter	9.00%	Mean tree height	9.30%
	5	Mean tree height	8.30%	Total basal area	6.50%
	6	Trees per hectare	7.30%	Trees per hectare	6.20%
	7	*Percent live canopy cover	6.30%	Gross tree volume	5.90%
	8	Forest type	5.50%	Physiographic class	4.30%
	9	Mean crown ratio	4.10%	Ownership	3.90%
	10	Mean tree age	3.60%	Mean tree diameter	3.60%
Shrub Cover					
	1	Habitat series	11.0%	Elevation	16.6%
	2	Elevation	10.2%	Forest type	8.4%
	3	Forest type	9.8%	Mean tree height	7.4%
	4	Mean tree height	9.2%	Site index	6.8%
	5	Total basal area	7.3%	Gross tree volume	6.6%
	6	Gross tree volume	6.8%	Site productivity	6.1%
	7	Number of trees	5.3%	Mean tree age	5.4%
	8	Site productivity	5.2%	Number of trees	5.2%
	9	*Percent live canopy cover	5.0%	Total basal area	4.8%
	10	Trees per hectare	4.7%	Predominant diameter class	4.3%
Shrub Height					
	1	Elevation	15.1%	Elevation	15.9%
	2	Forest type	10.9%	Mean tree age	10.7%
	3	Site productivity	9.8%	Forest type	8.9%
	4	Mean slope	9.4%	Mean tree height	7.6%
	5	Mean crown ratio	6.0%	Gross tree volume	7.5%
	6	Mean tree height	5.7%	Total basal area	7.5%
	7	Mean tree age	5.0%	Number of trees	5.9%
	8	Total basal area	5.0%	Mean tree diameter	5.5%
	9	Gross tree volume	5.0%	Site productivity	4.8%
	10	Number of trees	4.9%	Physiographic class	4.3%

Table 6
Frequency of most and least important variables in cover and height models for the IW and PNW datasets. Frequency of variables refers to the number of times a predictor appears as a top ten ranked variable in each dataset model. Least important variables (in asterisks) are those that appeared only in the bottom ten variables for all models. Frequency values range from zero to six, reflecting the maximum number of models for each dataset. Most important variables appearing most frequently in models as top ten variables are indicated in bold. Marginal plots of these variables are shown in Fig. 4. NA refers to regionally-specific variables that were not measured. See Table 2 for variable descriptions.

Variable	IW	PNW
Mean tree height	**6**	**6**
Mean tree age	4	4
Mean crown ratio	2	0
Mean tree diameter	4	5
Mean slope	1	0
Mean uncompacted live crown ratio (IW only)*	0*	NA
Crown class*	0*	0*
Number of trees	**6**	**6**
Primary stand disturbance*	0*	0*
Secondary stand disturbance*	0*	0*
Primary stand treatment*	0*	0*
Secondary stand treatment*	0*	0*
Elevation	**4**	**6**
Predominant diameter class (measured)	0	0
Forest type	**3**	**6**
Habitat series	3	0
Percent live canopy cover (IW only)	5	NA
Percent live and missing canopy cover (IW only)	1	NA
Ownership	0	1
Physiographic class	0	4
Site index	0	2
Site productivity	2	5
Predominant diameter class (computed)	0	0
Total basal area	**6**	**5**
Seedling trees per hectare	0	0
Trees per hectare	5	3
Gross tree volume	**6**	**6**
Topographic position (PNW only)	NA	0

in PNW models.

For the IW data, the top ten variables represented between 69 and nearly 78 percent of the total permutation importance (Table 5). The top five most important variables for each model alone represented between 40 and 52 percent of the total permutation importance and included, among others: total basal area, gross tree volume, mean tree height, number of trees, and elevation. These latter variables also appeared most often as the top five most important variables in IW models.

3.4. Patterns of ranked variable importance (all models)

Comparing the mean ranks of all model variables for each model in both datasets also revealed distinct similarities with both most and least influential variables (Fig. 3). Overall, the top ten most influential variables across both datasets for all lifeform cover and height models, in descending order of importance from highest to lowest, were: mean tree height, gross tree volume, elevation, number of trees, total basal area, forest type, mean tree diameter, mean tree age, trees per hectare, and site productivity. The top ten least influential variables overall (excluding regional variables specific to either IW or PNW regions), in ascending order of importance from lowest to highest, were: secondary stand disturbance, secondary stand treatment, crown class, primary stand treatment, primary stand disturbance, seedling trees per hectare, ownership, predominant diameter class of live trees (computed), predominant diameter class (measured) of live trees, and physiographic class.

Between the IW and PNW datasets, similar patterns of variable importance existed but with two notable exceptions. For the IW data, both trees per hectare and percent live canopy cover were also considered among the ten most important variables. Thus, for the IW models overall, the top ten most influential variables, in descending order of importance from most to least influential, were: gross tree volume, total basal area, mean tree height, number of trees, elevation, mean tree diameter, trees per hectare, percent live canopy cover, forest type, and mean tree age (Fig. 3).

The notable exceptions for the PNW data were physiographic class and site productivity, which were among the top ten most important variables, and habitat series, which was among the least influential variables (Fig. 3). Thus, the top ten most influential variables across all PNW models, in descending order of importance, were: elevation, mean tree height, number of trees, forest type, gross tree volume, total basal area, site productivity, mean tree age, mean tree diameter, and physiographic class. A complete table containing all variables and their ranking (based on permutation importance) over all cover and height models by lifeform for each dataset can be found in Appendix Table 2.

3.5. Patterns of permutation-based variable importance for individual grass, forb, and shrub models

Clear patterns of variable importance were also found within individual lifeform models for each dataset. Within grass cover and height models for the IW and PNW data, number of trees, total basal area, gross tree volume, trees per hectare, mean tree height, and mean tree diameter appeared in all four models as important variables (Table 5). Forest type appeared as a top ten variable in all grass models except for IW grass cover, and percent live canopy cover, which was only measured for the IW data, appeared in both IW models. Elevation appeared in the top ten most important variables in all models except for grass height. Physiographic class appeared as important in both PNW grass models.

In both forb cover and height models for the IW and PNW data, mean tree height, gross tree volume, number of trees, mean tree age, and mean tree diameter were the most common important predictors in all four models (Table 5). Elevation appeared in all models except for IW forb height while site productivity appeared in all models except for IW forb cover. Total basal area and trees per hectare appeared in all models except for PNW forb cover, while forest type appeared in both PNW forb models. Percent live canopy cover appeared in both IW forb models and percent missing live canopy cover appeared as important predictors in the IW forb height model. Both of these variables are regional.

In shrub cover and height models for the IW and PNW data, elevation, forest type, mean tree height, total basal area, gross tree volume, number of trees, and site productivity were the most common important predictors in all four models (Table 5). This represents the greatest number of commonly appearing important variables (seven) across lifeform models. Mean tree age appeared in all models except for IW shrub cover. The IW regional variable, percent live canopy cover, appeared only in the shrub cover model.

For all models combined, several variables for each model within and between datasets appeared as most commonly important, that is, occurring as top variables in all models. Across the six models for the IW dataset, the most frequently occurring top ten important variables that were found include mean tree height, number of trees, and gross tree volume. These three variables appeared as top ten variables in all six IW models (Table 6). Similarly, for the PNW, the most commonly occurring most important variables that were found in all six PNW models also include mean tree height, number of trees, and gross tree volume. In addition, both elevation and forest type also appeared within the top ten most important variables in all six PNW models.

For both datasets overall, the most important variables common to all twelve models included mean tree height, number of trees, and gross tree volume. The most commonly occurring least important variables that were found across all models include primary and secondary disturbance, primary and secondary treatment, and crown class (Fig. 3, Table 6).

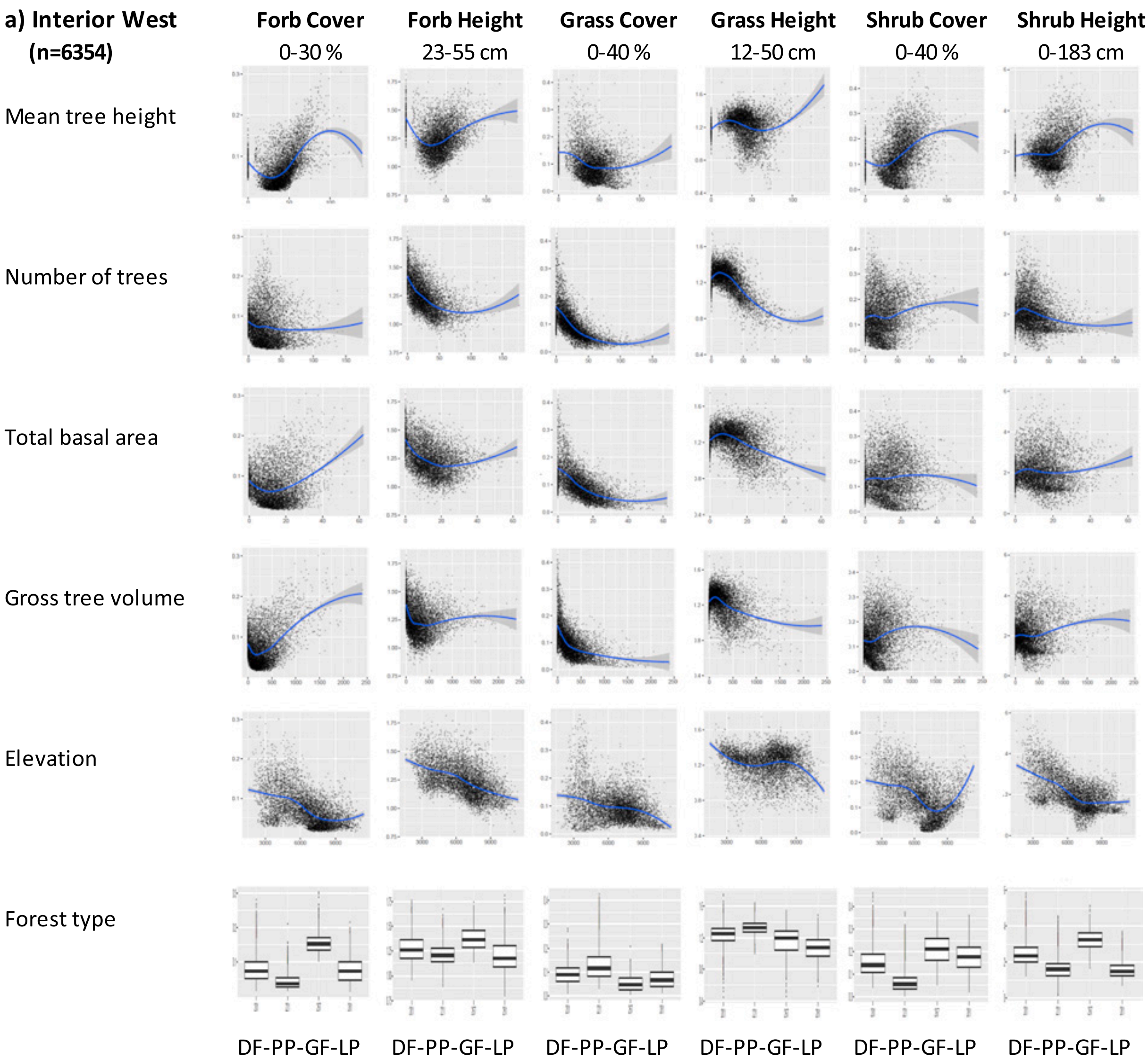


Fig. 4. (a and b). General predicted marginal relationships of most frequently occurring top six important variables across all models for the IW and PNW datasets. Mean predicted responses are shown along with 95% confidence intervals. Ranges for response axes for cover models are given directly below each model name in percent and in centimeters for height models. For forest type plots, the x-axis is given at bottom (DF = Douglas-fir, PP = ponderosa pine, GF = grand fir, LP = lodgepole pine). See Table 2 for variable descriptions. Model names are based on the response variables described in Table 3.

3.6. Predictive patterns of most important variables

In considering the six most important variables (mean tree height, number of trees, total basal area, gross tree volume, elevation, and forest type) based both on their mean rank and frequency within cover and height models from both datasets (Table 6), we see the predicted model responses as a function of each of these predictor variables (Fig. 4). The predicted mean cover and height response for all shrub and forb models generally increased with increasing mean tree height but decreased with grass models in both datasets. According to the variable prediction models, understory cover and height decreased with increasing tree number for most models except for PNW forb cover and shrub height, where it was generally increasing. For both shrub cover models, there was little to no relationship between predicted shrub cover and number of trees.

Predicted cover and height for most models in both datasets generally decreased with increasing total basal area and gross tree volume. The exceptions here were with PNW forb cover and shrub height, where forb cover was predicted by our models to increase with increasing basal area and tree volume, and with IW shrub cover and height, where little to no relationship was evident. Predicted response for both cover and height generally decreased in all models with increasing elevation except for PNW grass cover and grass height, where cover and height both increased with elevation.

Mean predicted response of understory cover and height also varied with forest type (representing the four major forest types considered here). Forest type was the most important categorical predictor and was present as a top ten variable in all six of the PNW models and in three of the IW models (Table 6). In all of the models except for grass cover and height, the highest predicted cover and height responses were found in either Douglas-fir or grand fir forest types, and lowest in both ponderosa and lodgepole pine forest types. In the four grass height and cover models, mean predicted grass cover and height were highest in the ponderosa pine forest type, while lowest in the Douglas-fir type of the PNW and in the grand fir and lodgepole pine forest types of the IW (Fig. 4).

4. Discussion

In this study, we considered many field-measured variables, both biotic and abiotic, for inclusion in our models, relying on those readily measured in the field by FIA as part of their annual plot inventories. We demonstrated that clear and similar patterns of measured variables exist

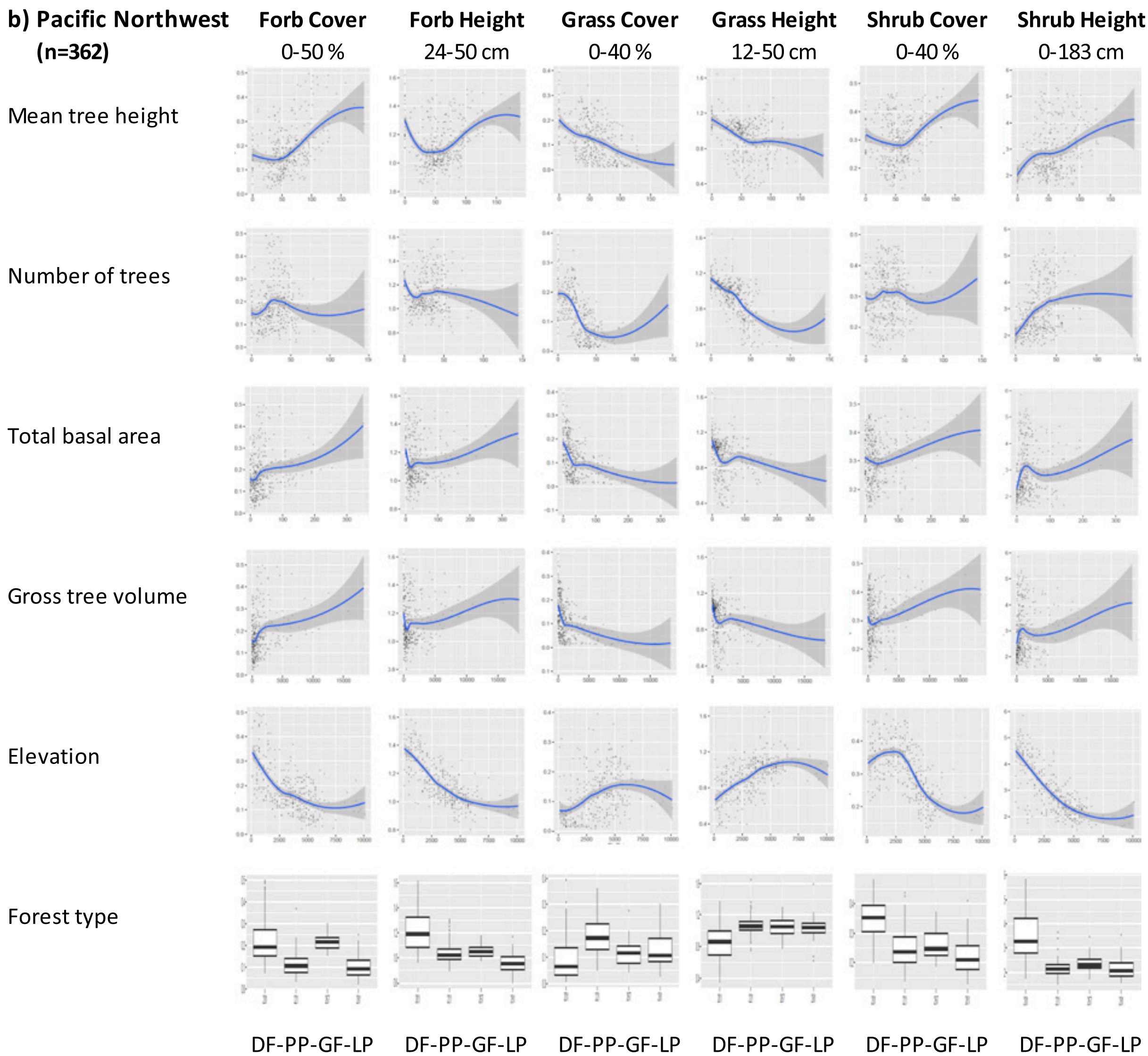


Fig. 4. (*continued*)

that are influential in predicting understory vegetation cover and height over all lifeform cover and height models in the forest types we considered. We also found similarly consistent patterns with regard to the least influential variables.

We were not overly surprised that several of these variables representing overstory conditions proved influential, since they depict longstanding relationships supported by considerable previous work of the effects of overstory on understory vegetation (discussed in Section 4.2 below). These include tree height, density, canopy cover, and volume. Elevation was also very influential and this has also been well established as an important topographic factor in many such studies.

We were surprised, however, that several other variables were not important in contributing to predictions of understory vegetation structure. These include additional topographic variables such as aspect and slope, which have been well established to play a significant role in determining understory vegetation patterns, distribution, and abundance. We were also surprised that habitat series (Pfister et al., 1977), and stand disturbance and treatment histories were not more prominent in our models. We discuss possible reasons for these results below in Section 4.3.

Nonetheless, these results show promise in using national and regional forest inventory data to predict understory vegetation structure and biomass and would greatly facilitate efforts to estimate terrestrial carbon stocks, fuel loading, and wildlife habitat. However, more work is clearly needed as we discuss below in Section 4.4.

4.1. Model implications

Individual model results of understory cover and height were highly variable and were generally weak to moderate in their predictive accuracies when applied to the test OOB data. Overall, they did not account particularly well in explaining the variation of understory lifeform height and cover within the plots we considered. This was especially true of forb and grass height models, which explained only between 8% and 19% of the test data variance. Shrub height models represented the best model fits with between 27% and 39% of the variance explained. These results underscore both the longstanding difficulty and complexity in capturing understory vegetation structure and dynamics, especially over the large geographic regions we considered.

Our results are comparable to Suchar and Crookston (2010), whose work is most closely aligned with the present study. They assessed understory vegetation cover and biomass using regional inventory data from national forests of the Pacific Northwest and focused on climatic, disturbance history, topographic, soil, and overstory canopy cover attributes as predictors. The percent variance explained by their RF models for understory cover and biomass indices ranged from 37 to 44 percent, a result similar to ours. They found that climate, elevation, and slope were highly significant predictors of understory structure and, in the case of elevation, corroborate our findings that shrub and herb cover all decrease with increasing elevation. Slope played an influential predictive role in their study but was a variable of only intermediate importance in our models. In both studies, disturbance history had been expected to play a greater role but these data suffered from inconsistent and anecdotal records. A key difference between the present study and theirs is that they focused more on site attributes while we emphasized stand and overstory characteristics, but with comparable results. Thus, in comparison, we conclude that our focus on plot-level overstory attributes did not improve predictability of understory vegetation.

In another related study, FIA inventory data was used to estimate understory species richness in coastal plain forests of southeastern U.S. (Timilsina et al., 2013). Among the most important variables in their models included forest type, stand age, and disturbance and treatment histories. Their regression tree models explained approximately 57% of the variation in herbaceous species richness. With the exception of forest type and stand age, the importance of these variables is in contrast to our results where stand disturbance and treatment history were among the least important variables. This is likely due to the much greater availability of these data in these highly managed forests but nonetheless highlights the potential role that such variables could have played in our models if such stand history information were present in our data.

Nonetheless, and despite the limited ability of these predictors to better account for understory structure, a substantial finding is the identification of the most and least important field-collected variables that commonly contributed to these models. The top ten variables for each model alone accounted for more than two-thirds of the total variable importance, while the least important variables accounted for typically less than five percent. Despite these limitations however, the presence of these variables and their predicted relationships with understory vegetation structure appear to coincide, at least in part, with our ecological expectation of them based on previous studies. This is discussed in the following section.

4.2. Most important variables

The majority of top variables for each of the cover and height models are ones that relate exclusively to overstory structure conditions and canopy closure. In these cases, the variables predicted to be important by the RF algorithm appear to, in part, substantiate ones typically found in many silvicultural studies of the effects of overstory conditions upon understory vegetation structure. In the aggregate, these variables represent individual measures of overstory tree size and density; attributes both of which have been shown to influence understory vegetation structure through their effects upon canopy closure and competition for light and nutrients.

For example, increases in tree size (basal area and height), density (number of trees and trees per hectare), and volume (gross tree volume) as part of stand development have typically resulted in decreases in understory vegetation abundance in many forests (Ffolliott and Clary, 1982; Alaback, 1982; Hedwall et al., 2013; Tonteri et al., 2016; Bakker and Moore, 2007; Bataineh et al., 2006). However, the predicted responses of vegetation cover and height in our models were very mixed here and clear predictive trends were sometimes difficult to discern. Predicted forb cover generally increased with increasing average tree height, tree density, total basal area, and gross tree volume, whereas predicted forb height generally decreased against these predictors. This finding was complemented by Jules et al. (2008), who examined a 420-year chronosequence of interior Douglas-fir stands, where forb cover also increased with stand age and overstory canopy closure. They explained this increase in very old stands as likely due to both the greater structural complexity of the overstory canopy allowing gaps in light penetration and the presence of dispersal-limited forb species that very gradually expand and recolonize over centuries. This may well be the case within our data but we do not have enough detailed information to substantiate this.

Conversely, predicted shrub cover tended to decrease with the PNW data and increase with IW data against these four overstory variables. This was also observed by Jules et al. (2008) and seems consistent with the notion that the IW region is generally more arid than the PNW, where tree cover (which is on average greater in the PNW region) increases light attenuation, and becomes more of a limiting factor constraining shrub growth. Increasing tree cover in the IW region may indicate increased water availability and hence improved growing conditions for both trees and shrubs, but tree canopy cover closure may not be so pronounced as to create such a light-limited environment restricting shrub growth.

In contrast, both predicted grass cover and height uniformly decreased in both datasets as each of these overstory predictors increased. Considering also the mostly concomitant predicted increases in both forb and shrub cover and height, this suggests that grass abundance may be constrained through canopy closure by increased forb and shrub abundance, as well as by increased overstory size and density. This complements other studies, including those in longleaf pine forests of the southeastern U.S., where (C_4) grasses tended to decline with increased overstory basal area and density (Veldman et al., 2013), and in graminoid cover of older unmanaged interior Douglas-fir stands, which favored more shade tolerant species (Jules et al., 2008).

Admittedly, many of these tree-based variables are biometrically correlated with one another and both the IW and PNW data exemplified this to similar degrees. For example, tree diameter is correlated with total tree basal area and tree height, and consequently tree volume. Tree diameter is also correlated with tree age, especially when site conditions are considered. The number of trees is also directly correlated with total trees per hectare. For all of these correlated predictors, there is evidence that suggests that the RF algorithm tends to select those variables with high collinearity (Nicodemus et al., 2010). Our results illustrate the unique selection procedures of the RF algorithm when compared with other classification methods by spreading the variable importance across several variables, instead of removing them. This potentially guards against the elimination of variables that are good predictors of the response and that may be ecologically important despite being correlated with other predictors (Cutler et al., 2007). In addition to these reasons, we decided to keep these variables since they also represented ones that are easily measured and were an important objective of this study. Moreover, studies like this are not numerous and thus the information about these variables is valuable.

Elevation was also an important variable in all of the PNW models and in four of the IW models. It was also the variable with one of the most consistent and distinct predicted relationships within all of the models that were considered. Predicted values of both understory cover and height generally decreased with increasing elevation in ten of the twelve models. The exception to this was with grass height and cover for the PNW, where the predicted height and cover of grass increased with increasing elevation. This relationship is also consistent with the findings of Suchar and Crookston (2010), where elevation was deemed a significant predictor, with both understory biomass and cover decreasing with elevation. Ensslin et al. (2015) also reported decreases in understory biomass, especially shrubs, over a broad elevation gradient, owing largely to differences in precipitation patterns. This decline may also be due to colder temperatures, shorter growing season, and thinner, less productive soils associated with higher elevations. Elevation for PNW plots ranged from 30 to 3050 m and from 470 to 3475 m for IW plots, which represent very broad elevational gradients for both regions that would allow for this possibility. It is possible too that the predicted relationship of increasing grass cover and height with increasing elevation for the PNW data reflects an overall dominance of non-graminoid species associated with these plots in the much more productive forest environments found at lower elevations.

Forest type also exhibited trends that seem somewhat consistent with our ecological understanding of the four forest types we considered. Predicted cover and height for both forb and shrub were highest in the grand fir and Douglas-fir forest types for each dataset. Both of these types are typically more productive and would be expected to support greater amounts understory vegetation (Pfister et al., 1977). Alternatively, highest predicted grass cover and height was observed in ponderosa pine forests, which represent here the most drought tolerant forest type. Since graminoids are known to thrive and are well adapted to more arid conditions, this also supports the expectation of their greater abundance in these forests (Gibson, 2009).

Very similar to elevation, the trends in predicted cover and height for forbs and shrubs declined as site productivity declined (not shown). This was not the case for grass cover and height, where mean predicted cover and height increased for the PNW data as site productivity declined (there was no little to relationship present with the IW data). Again, these trends are consistent with ecological expectation in that sites containing greater growth potential will also generally support greater amounts of understory vegetation, shown by many studies (Hedwall et al., 2013; Hart and Chen, 2006). A possible reason why grass was seen in our data to increase with declining site productivity may again be due to its ability to withstand and better compete in drier conditions and on less productive sites (Gibson, 2009).

One last important predictor deserves mention. It is percent live canopy cover, which is a regional variable that was measured only on the IW plots and appeared as a top ten variable in five of the six IW models (except for shrub height). Percent live canopy cover is a measure of overstory tree crown closure for the plot ranging in value from 0 to 99 percent, and is an important measure of overstory density used in many related studies (e.g. Suchar and Crookston, 2010; McKenzie et al., 2000; Jules et al., 2008). Predicted forb and grass height, and grass cover all tended to decrease with increasing live canopy cover (not shown), a result largely corroborated by these studies (McKenzie et al., 2000; Jules et al., 2008) and many others; due to greater light attenuation of increasing crown closure on more shade intolerant species. Later releases of the PNW data included the measurement of percent live canopy cover on all plots and may have proved to be of similar predictive importance.

4.3. Least important variables

Primary/secondary stand disturbance and treatment history, and crown class were all found to be the most consistently least important variables in all models and contributed the least overall to their overall predictive ability. In addition, the regional variables mean uncompacted live crown ratio (IW) and topographic position (PNW) were also consistently among the least important variables. We expected a greater influence of both stand disturbance and treatment history measures, but there was a general lack of collected information present to discern any meaningful effect, despite recoding and simplifying these variables as presence/absence. Many of our selected plots are located in areas where little to no disturbance or treatment was evident both in the previous 5–10 years and covering an area at least 0.4 ha in size, which are the thresholds FIA crews use for recording these data. Similarly, crown class offered little variation or contrast to discern any effects, which was likely made worse by using the mode as a summary measure for each plot. For mean uncompacted live crown ratio, nearly one quarter of the observations were missing and were subsequently imputed, which would have reduced overall variation in the plots. Lastly, the PNW regional variable, topographic position, was also not very discriminating in revealing any effect, with over three quarters of all plots occurring either on upper, middle, or lower hillsides.

Several of the least important variables had negative importance values in most of the PNW models (not shown), which may be related to the much smaller sample size of these data, or simply an artifact of the permutation process. Since variable importance here uses the method of randomly permuting the variable and determining the difference between the OOB prediction errors both with and without this variable, positive importance values signify that the model predictive accuracy is reduced without the variable. Consequently, an importance value of zero means that the variable contributes nothing to model predictive accuracy and a negative value actually improves model accuracy if that variable was not included; implying that model noise is actually more important when that variable is absent versus present (Ehrlinger, 2015).

It was surprising that measures of both slope and aspect did not play a greater role in our models, since it is well known that vegetation is greatly affected by topography, or the combination of elevation, aspect, and slope (Warren, 2008; White et al., 2005). In the case of aspect, this may have been partly due to the relative difficulty in obtaining accurate aspect measurements in highly variable landscapes, which FIA itself has acknowledged in its own quality control efforts (Pollard et al., 2006). We included slope in our final models since it was a measured field variable that was simply averaged over each subplot; but it represented only intermediate importance throughout. Aspect was measured similarly too but not included in our final models, since we needed to compute several different measures to arrive at a mean plot aspect (see Appendix Table 1) and hence considered them derived predictors. Nonetheless, all the computed measures of aspect that we tested in earlier iterations were consistently among the least important predictors (not shown).

Lastly, habitat series was of intermediate importance for the IW data but not important for the PNW, which was also surprising to us, especially in this latter case. This was likely due in part, to the fact that nearly 23% of the PNW cases for this variable were missing and needed to be imputed (using a mode rule for categorical predictors). Also, the number of plots per factor level was highly variable and unbalanced, with several habitat series only represented by less than ten plots.

4.4. Model improvement

One of the more likely and obvious approaches for model improvement lies with stratification, which would serve to partition the considerable variation encountered in our data. Stratification could occur geographically based on some measure of climate regime or ecoregion; both possibly important predictive attributes that may not have been adequately represented and integrated within our chosen suite of predictors. For example, the Douglas-fir forest type identified here (which includes both interior and coastal varieties) has such broad ecological amplitude that differences in climate driving vegetation growth and structure may not be readily discernible within such a coarse classification.

However, many habitat series occurred within each of the four forest types we considered; a level which could signify a better indicator of differences in climate. If so, then stratification could also occur for these habitat series, since they would better represent an integrated measure of physiographic, climatic, and edaphic factors driving vegetation growth and structure in western interior forests of the U.S. (Kusbach et al., 2014; Roberts, 2015). Moreover, as stated previously, habitat types in this study were truncated to the series level only, in order to increase the number of observations per habitat type level. By stratifying and conducting a more focused analysis by habitat type (including type and phase) may reveal stronger associations between vegetation types and other environmental indicators. Lastly, stratification could also occur along other similarly integrated metrics in our data, such as site index and site productivity.

Another considerable limitation in these models is the obvious and inherent difficulty in obtaining precise cover and height data for understory vegetation from plots, as well as timing such measurements to coincide with annual peak foliar biomass conditions. Moreover, in this study, height class midpoints were used since actual heights were not collected and are not part of the current FIA measurement protocol. As previously described, lifeform heights were determined using a cover-weighted approach, which uses midpoint class heights instead of actual height (again, which FIA does not measure) for each of four FIA height layers. For example, if a lifeform had only cover present in the first height layer of 0–61 cm for all four subplots, then the cover-weighted height for that lifeform is 30.5 cm, the midpoint height. For cover-weighted height to exceed 30.5 cm for the entire plot, then some cover in the next height layer (64–183 cm) would have to be present on at least one subplot. The relative coarseness of height data obtained using this approach had the net effect of constraining variability in heights to more discrete and less continuous values. Nonetheless, shrub height models performed comparatively well, perhaps due to the fact that shrubs are more easily measured in the field, given their uniformly greater stature.

In the absence of using cover-weighted heights to better approximate vegetation height, the remaining option would be to use the height layers themselves as classes; a coarser option which would have likely resulted in poorer prediction estimates. Additionally, height measurements for the IW data were collected for three height classes until 2011, when a fourth height class protocol was adopted. All previously collected height data prior to 2011 were then adjusted by FIA to accommodate the new national standard. This was not the case with the PNW data since they adopted the new standards in 2011, the only year of data that could be used.

Ocular estimates of vegetation cover over an entire subplot are also relatively subjective and vary from observer to observer, despite many historical efforts to correct for this. However, FIA implements a strict quality assurance program for their measurements through "blind checks" of measured variables. For vegetation cover, compliance results indicate that these measurements are very repeatable at the ± 5 and 10 percent levels for all height layers, although the aerial cover estimate (i.e. fifth layer) was found to be typically lower than the height layer covers, regardless of the tolerance. This was likely due to the inherent difficulty in combining vegetation at different heights to obtain an overall aerial cover (Patterson and O'Brien, 2011). Nevertheless, more precise measurements in both of these dimensions would likely improve the predictive ability of these models, especially as it concerns lifeform height.

Also, as additional variables are being considered for inclusion in the FIA measurement protocol nationwide, these may also prove valuable in predicting understory vegetation structure. One noteworthy example already encountered here is the later addition of percent live tree canopy cover for the PNW data, which may have had a similar (or greater) importance in our models as with the IW data, had it been available at the time.

At the same time, with the adoption of more probabilistic standardized measurements nationally of understory vegetation, such as with FIA, improvements in overall prediction of understory structure and composition would be more easily achievable, as well as scalable (Gray et al., 2012). Such vegetation databases created from nationally consistent protocols are being developed and increasingly used to address long term ecological and management-oriented questions. Other datasets could be explored which more strongly emphasize understory vegetation measurement; although admittedly, such data would not likely be part of a national or regional inventory strategy.

Sample size and missing data were also issues that likely compromised both model performance and predictive power, especially with the PNW data. With the relatively small number of useable plots with suitable understory vegetation measurements combined with the higher percentage of missing data for several of the predictors resulted in a substantial loss of information. Sample size was not an issue for the IW data but missing data was also significant here, and for both datasets the imputation of these missing cases is never a substitute for actual data. Improving sample size and reducing missing cases should lead to improved model predictions.

One approach that was adopted by Suchar and Crookston (2010) in their study was to analyze plot data at the subplot level; a scale where understory vegetation patterns are more likely governed by their natural processes than in the aggregate at the plot level. This could address some of the above concerns by increasing heterogeneity of understory measurements by using actual observation-level data and avoiding the loss of information from scaling, while at the same time greatly enlarging sample sizes (e.g. a fourfold increase in the present study). Additionally, they considered biomass (plant cover × height) as response variables, which may or may not have improved our model predictions.

4.5. Model applications

One potential and direct application of this work will be to implement these results in the U.S Forest Service (USFS) Forest Vegetation Simulator (FVS), a model which simulates a broad range of silvicultural treatments (Dixon, 2002). It addresses the effects of thinning, fire, insect and disease, and climate upon forest growth and yield while simulating calibrated treatments and disturbances over distinct geographical regions throughout the United States; but generally lacks any current mechanism for estimating understory structure in forests (Crookston and Dixon, 2005). In addition, nearly all of the variables we considered, and especially those most important and most frequently occurring in our models (Tables 5 and 6), are ones that FIA readily measures or that can be easily computed from them. For forest simulation models such as FVS, which directly incorporates FIA data for its predictions (Shaw, 2009), this work signifies an attempt to integrate FIA site, overstory, and understory data to reveal these relationships in making understory vegetation cover and height (and biomass) predictions.

Our model results may also prove useful in improving estimates of the contributions of understory vegetation for assessing terrestrial carbon stocks. Less attention has been given to understory vegetation because it is so variable and the fraction of carbon it represents is small relative to total ecosystem biomass (Johnson et al., 2017). Nonetheless, models estimating understory vegetation are still critical over broad spatial scales (Russell et al., 2014). Moreover, in the context of climate change, there is an increasing need to base carbon models on empirical data that are regularly monitored (Woodall, 2012).

Likewise, the contribution of understory vegetation to the fuelbed profile is another extension of this work. Understory fuel loading and bulk density are among the most important predictors of wildland fire intensity and rate of spread (Keane, 2015). The recent and increasing accumulation of ladder fuels in many western forests can exacerbate fire severity by exposing tree crowns, leading to catastrophic wildfires with often dramatic post-fire effects on forest ecosystems (Kramer et al., 2016). Estimation of understory vegetation cover and height based on field measurements should help improve estimates of fuel loading, and especially ladder fuels, which remain largely underrepresented (Kramer et al., 2016).

In addition, such estimates of understory structure and biomass by lifeform will help enhance wildlife diversity by providing more refined assessments of browse and cover suitability for habitats of many species (Hagar, 2007). For example, estimates of shrub biomass have been useful for assessing browse and forage availability in ungulate species (Visscher et al., 2006) and smaller herbivores. For avian species, herbaceous vegetation is an important forage component of habitat and has direct effects upon their overall breeding success (Betts et al., 2010). Shrubs, especially those in early-seral stages of forest development are important for all stages of avian breeding cycles (Hagar, 2007). Further understanding of the habitat relationships and understory vegetation requirements of forest faunal species are critical for understanding their demography and for devising sound conservation strategies.

4.6. Conclusions

In this study, we examined the role of various overstory and site variables measured in the field to predict understory vegetation cover and height on permanent plots in selected forests of the western United States. We found the most important variables responsible for predictions of understory structure to be those representing overstory conditions as well as elevation, forest type, and to a slightly lesser degree, site productivity. Several variables proved surprisingly unimportant such as slope, aspect, habitat series, and disturbance and treatment history; likely due to a combination of data limitations and perhaps analytical approach in scaling. The use of standardized measurements also allow for the addition of variables for improved predictions of understory vegetation, as in the case here with percent canopy cover, an important overstory variable for the IW data but only recently introduced in the PNW region.

Despite the relatively weak ability of our models to predict understory structure, many of the most influential variables measured over a large geographic region conform to our ecological understanding and expectation of their role in such predictions based on considerable previous research. This, in itself, is an important finding and signifies an attempt to address the prediction of understory structure using standardized field-collected data obtained over such an extensive area. It is promising and suggests that the prediction of understory vegetation structure may be more generalizable, despite its current elusiveness, through the addition of more pertinent and targeted data collected in the field (e.g. non-ocular estimates of vegetation cover and height, measurements during periods of peak foliar biomass, soil characteristics); and even supplemented with additional outside data, such as those pertaining to climate.

More refinement and work is clearly needed to improve predictions but our present results are encouraging and advantageous for several reasons: 1) considerable capability currently exists within FVS for implementation of these models, 2) the use of standardized data permits broad-scale predictions of understory structure both regionally and nationally, and 3) these data are being routinely collected and allow for regularly updated estimates of understory structure (and biomass) for improved assessments of carbon stocks, fuel loading, and wildlife habitat.

Acknowledgements

For a project of this scope, the help and expertise of many individuals were needed. We thank the following individuals for their valuable input, expertise, oversight, and suggestions throughout the course of this study: John Shaw (IW-FIA), Chris Toney (IW-FIA), Mike Van Dyck (USFS), Benjamin Bird (USFS), Nick Crookston (USFS-retired), Stephanie Rebain (USFS), Karen Waddell (PNW-FIA), and Andrew Gray (PNW-FIA).

This work would not have been possible without funding support provided by the Joint Fire Science Program (Project ID # 12-1-02-15), a federal interagency organization conducting wildland fire research. They played no role during any stage of the study including the design, data collection, interpretation of results, or the writing of this paper. This manuscript was prepared by public employees as a part of their official duties and is therefore in the public domain.

Author contributions

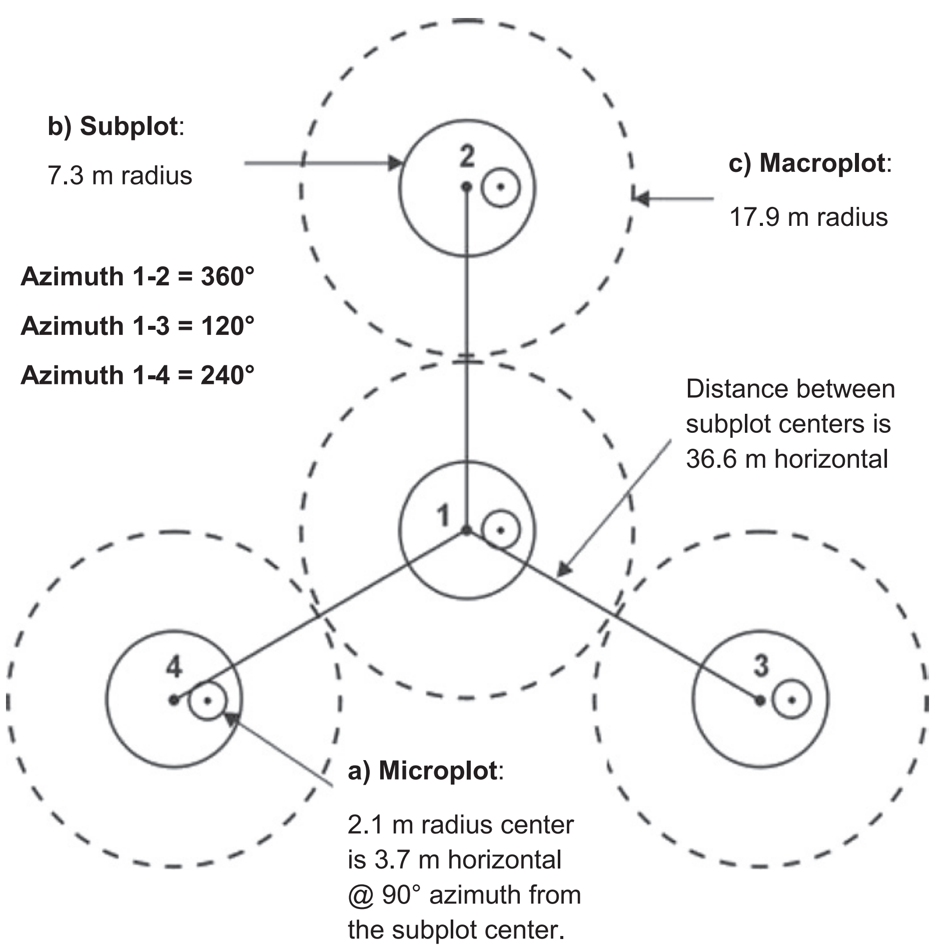
All authors contributed to conceiving and designing the analytical approach. M.K. and M.R. conceived the study design and performed the gathering of study data, preparation and variable selection, and S.B. designed and modeled the data. All authors took part in writing and revising the manuscript.

Declaration of Competing Interest

The authors declare no conflict of interest.

Appendix

See Appendix Fig. 1 and Appendix Tables 1 and 2



Appendix Fig. 1. Forest Inventory and Analysis (FIA) national annual inventory plot design. Approximate sampling areas are as follows: (a) microplot (13.5 m^2), (b) subplot (168 m^2), and (c) macroplot (1012 m^2) (O'Connell et al., 2013; Gray et al., 2012).

Appendix Table 1
Description of the several aspect variables considered in IW and PNW understory cover and height models. Variables were not included in final models since they were all derived and were found to contribute minimally to overall model predictive accuracy.

#	Variable Name	Description	Variable Type	Scaling Method
1	Eastness (derived)	Subplot aspect (n = 4), in degrees, that is averaged for the plot and transformed into eastness (sin[aspect]). See Zar (1999) for methods.	Continuous	Circular Average
2	Northness (derived)	Subplot aspect (n = 4), in degrees, that is averaged for the plot and transformed into northness (cos [aspect]). See Zar (1999) for methods.	Continuous	Circular Average
3	Final_Mean_Aspect (derived)	Transformation of average eastness and northness components into mean aspect (in degrees). See Zar (1999) for methods.	Continuous	Circular Average
4	Coded_Aspect (derived)	Reclassification of Final_Mean_Aspect into eight groups starting with 337.5 degrees and containing 45 degree classes (e.g. 1 = 337.5–22.5 degrees, 2 = 22.5–67.5 degrees,…., 0 = flat areas).	Categorical	NA
5	Cardinal_Direction (derived)	Recoding of Coded_Aspect into eight cardinal directions, including flat areas.	Categorical	NA

Appendix Table 2
Variable importance rankings for all measured variables for cover and height models of the IW and PNW datasets. Ranks are presented in ascending order of importance (i.e. 1 - most important) and range from 1 to 27 for the IW data and 1 to 25 for the PNW data. Blank entries indicate regionally-specific variables that were not measured. See Table 2 for variable descriptions. Model names are based on the response variables described in Table 3.

Variable	Forb Cover		Forb Height		Grass Cover		Grass Height		Shrub Cover		Shrub Height	
	IW	PNW	IW	PNW	IW	PNW	IW	PNW	IW	PNW	IW	PNW
Mean tree height	2	4	4	6	8	7	5	4	4	3	6	4
Mean tree age	9	7	9	4	13	11	10	12	13	7	7	2
Mean crown ratio	14	15	11	13	16	16	9	11	12	11	5	12
Mean tree diameter	6	5	5	1	7	9	4	10	11	12	11	8
Mean slope	20	11	18	17	10	17	16	13	14	14	4	11
Mean uncompacted live crown ratio (IW only)	21	–	21	–	22	–	21	–	22	–	23	–
Crown class	24	23	23	23	25	23	23	22	24	23	25	22
Number of trees	4	6	3	8	2	1	3	1	7	8	10	7
Primary stand disturbance	22	19	24	20	24	22	24	20	23	22	22	25
Secondary stand disturbance	25	21	26	22	26	21	26	25	27	25	27	24
Primary stand treatment	27	24	25	25	23	18	25	19	26	19	24	19
Secondary stand treatment	26	22	27	24	27	20	27	17	25	21	26	23
Elevation	10	3	13	5	6	8	11	2	2	1	1	1
Predominant diameter class (measured)	17	17	12	12	18	24	14	18	18	10	18	16
Forest type	11	10	15	10	14	4	8	3	3	2	2	3
Habitat series	5	16	19	21	9	12	17	21	1	18	12	20
Percent live canopy cover (IW only)	8	–	7	–	5	–	7	–	9	–	14	–
Percent live and missing canopy cover (IW only)	15	–	8	–	11	–	12	–	15	–	15	–
Ownership	23	18	20	19	21	19	18	9	21	16	20	17
Physiographic class	18	9	22	14	15	2	22	8	16	13	21	10
Site index	13	2	17	11	19	13	20	16	17	4	19	18
Site productivity	12	1	10	7	17	3	15	14	8	6	3	9
Predominant diameter class (computed)	16	20	14	15	12	14	13	24	19	17	17	13
Total basal area	3	13	2	2	1	5	2	5	5	9	8	6
Seedling trees per hectare	19	25	16	16	20	15	19	15	20	24	16	15
Trees per hectare	7	14	6	9	4	10	6	6	10	15	13	14
Gross tree volume	1	8	1	3	3	6	1	7	6	5	9	5
Topographic position (PNW only)	–	12	–	18	–	25	–	23	–	20	–	21

References

Ahmad, B., Wang, Y., Hao, J., Liu, Y., Bohnett, E., Zhang, K., 2019. Optimizing stand structure for tradeoffs between overstory and understory vegetation biomass in a larch plantation of Liupan Mountains, Northwest China. For. Ecol. Manage. 443 (2019), 43–50. https://doi.org/10.1016/j.foreco.2019.04.001.

Alaback, P.B., 1982. Dynamics of understory biomass in Sitka spruce-western hemlock forests of southeast Alaska. Ecology 63, 1932–1948. https://doi.org/10.2307/1940131.

Alaback, P.B., 1986. Biomass regression equations for understory plants in coastal Alaska: effects of species and sampling design on estimates. Northwest Sci. 60, 90–103.

Bailey, J.D., Mayrsohn, C., Doescher, P.S., St. Pierre, E., Tappeiner, J.C., 1998. Understory vegetation in old and young Douglas-fir forests of western Oregon. For. Ecol. Manage. 112, 289–302.

Bakker, J.D., Moore, M.M., 2007. Controls on vegetation structure in southwestern ponderosa pine forests, 1941 and 2004. Ecology 88 (9), 2305–2319. https://doi.org/10.1890/06-1775.1.

Bataineh, A.L., Oswald, B.P., Bataineh, M.M., Williams, H.M., Coble, D.W., 2006. Changes in understory vegetation of a ponderosa pine forest in northern Arizona 30 years after a wildfire. For. Ecol. Manage. 235, 283–294. https://doi.org/10.1016/j.foreco.2006.09.003.

Bauhus, J., Aubin, I., Messier, C., Connell, M., 2001. Composition, structure, light attenuation and nutrient content of the understorey vegetation in a *Eucalyptus sieberi* regrowth stand 6 years after thinning and fertilization. For. Ecol. Manage. 144, 275–286. https://doi.org/10.1016/S0378-1127(00)00403-5.

Bechtold, W.A., Patterson, P.L., (Eds.). 2005. The enhanced Forest Inventory and Analysis program—national sampling design and estimation procedures. Gen. Tech. Rep. SRS-80. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. p. 85.

Betts, M.G., Hagar, J.C., Rivers, J.W., Alexander, J.D., McGarigal, K., McComb, B.C., 2010. Thresholds in forest bird occurrence as a function of the amount of early-seral broadleaf forest at landscape scales. Ecol. Appl. 20 (8), 2116–2130. https://doi.org/10.1890/09-1305.1.

Breiman, L., 2001. Random Forests. Machine Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Burton, J.I., Ganio, L.M., Puettmann, K.J., 2014. Multi-scale spatial controls of understory vegetation in Douglas-fir–western hemlock forests of western Oregon, USA. Ecosphere 5 (12), 1–33.

Campbell, M.J., Dennison, P.E., Hudak, A.T., Parham, L.M., Butler, B.W., 2018. Quantifying understory vegetation density using small-footprint airborne lidar. Remote Sens. Environ. 215, 330–342. https://doi.org/10.1016/j.rse.2018.06.023.

Cook, R.D., Weisberg, S., 1997. Graphics for assessing the adequacy of regression models. J. Am. Stat. Assoc. 92, 490–499. https://doi.org/10.1080/01621459.1997.10474002.

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.
Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J.C., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88 (11), 2783–2792. https://doi.org/10.1890/07-0539.1.
Dixon, G.E., comp. 2002. Essential FVS: A user's guide to the Forest Vegetation Simulator. Internal Rep. Fort Collins, CO: U. S. Department of Agriculture, Forest Service, Forest Management Service Center. 226 p. (Revised: September 24, 2018).
Ensslin, A., Rutten, G., Pommer, U., Zimmermann, R., Hemp, A., Fischer, M., 2015. Effects of elevation and land use on the biomass of trees, shrubs and herbs at Mount Kilimanjaro. Ecosphere 6 (3), 45. https://doi.org/10.1890/ES14-00492.1.
Ehrlinger, J. 2015. ggRandomForests: Random Forests for regression. Cleveland Clinic. https://arxiv.org/abs/1501.07196v2 (accessed January 2018).
Eyre, F.H. (Ed.) 1980. Forest cover types of the United States and Canada. Washington D. C. Society of American Foresters. p. 148.
Ffolliott, P.F., Clary, W.P. 1982. Understory-overstory vegetation relationships: an annotated bibliography. Gen. Tech. Rep. INT-136. Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Forest and Range Experiment Station. p. 39.
Gebert, K.M., Schuster, E.G., Woudenberg, S., O'Brien, R. 2008. Estimating forest fuels in the Southwest using forest inventory data. In: Proceedings of the 2002 Fire Conference: Managing fire and fuels in the remaining wildlands and open spaces of the Southwestern United States, December 2-5, 2002, San Diego, CA. Gen. Tech. Rep. PSW-GTR-189. Albany, CA: U.S. Department of Agriculture, Forest Service, Pacific Southwest Research Station. pp. 39–48.
Gibson, D.J., 2009. Grasses and Grassland Ecology. Oxford University Press, New York 320 p.
Gray, A.N., Brandeis, T.J., Shaw, J.D., McWilliams, W.H., Miles, P.D. 2012. Forest Inventory and Analysis Database of the United States of America (FIA). In: Dengler, J., Oldeland, J., Jansen, F., Chytry, M., Ewald, J., Finckh, M., Glockler, F., Lopez-Gonzalez, G., Peet, R.K., Schaminee, J.H.J., (Eds). Vegetation databases for the 21st century. Biodiversity and Ecology (4), 225–231. https://doi.org/10.7809/b-e.00079.
Gray, A.N., Whittier, T.R., 2014. Carbon stocks and changes on Pacific Northwest national forests and the role of disturbance, management, and growth. For. Ecol. Manage. 328, 167–178. https://doi.org/10.1016/j.foreco.2014.05.015.
Hagar, J.C., 2007. Wildlife species associated with non-coniferous vegetation in Pacific Northwest conifer forests: a review. For. Ecol. Manage. 246, 108–122. https://doi.org/10.1016/j.foreco.2007.03.054.
Hart, S.A., Chen, H.Y.H., 2006. Understory vegetation dynamics of North American boreal forests. Crit. Rev. Plant Sci. 25, 381–397. https://doi.org/10.1080/07352680600819286.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed. Springer, pp. 745 p..
Hedwall, P.-O., Brunet, J., Nordin, A., Bergh, J., 2013. Changes in the abundance of keystone forest floor species in response to changes in forest structure. J. Veg. Sci. 24, 296–306. https://doi.org/10.1111/j.1654-1103.2012.01457.x.
Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S., 2008. Random survival forests. Ann. Appl. Statist. 2, 841–860.. https://arxiv.org/pdf/0811.1645.pdf.
Ishwaran, H., Kogalur, U. 2017. Random Forests for survival, regression, and classification (RF-SRC), R package version 2.5.1.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning: with Applications. R. Springer, New York 426 p.
Johnson, K., Domke, G., Russell, M., Walters, B., Hom, J., Peduzzi, A., Birdsey, R., Dolan, K., Huang, W., 2017. Estimating aboveground live understory vegetation carbon in the United States. Environ. Res. Lett. 12, 125010. https://doi.org/10.1088/1748-9326/aa8fdb.
Joyce, L.A., Baker, R.L., 1987. Forest overstory-understory relationships in Alabama forests. For. Ecol. Manage. 18, 49–59. https://doi.org/10.1016/0378-1127(87)90124-1.
Jules, M.J., Sawyer, J.O., Jules, E.S., 2008. Assessing the relationships between stand development and understory vegetation using a 420-year chronosequence. For. Ecol. Manage. 255, 2384–2393. https://doi.org/10.1016/j.foreco.2007.12.042.
Keane, R.E., 2015. Wildland Fuel Fundamentals and Applications. Springer International Publishing, Switzerland. https://doi.org/10.1007/978-3-319-09015-3. 191 p.
Kerns, B.K., Ohmann, J.L., 2004. Evaluation and prediction of shrub cover in coastal Oregon forests (USA). Ecol. Ind. 4, 83–98. https://doi.org/10.1016/j.ecolind.2003.12.002.
Kramer, H.A., Collins, B.M., Lake, F.K., Jakubowski, M.K., Stephens, S.L., Kelly, M., 2016. Estimating ladder fuels: a new approach combining field photography with LiDAR. Remote Sens. 8, 766. https://doi.org/10.3390/rs8090766.
Kusbach, A., Shaw, J.D., Long, J.N., 2014. Discriminant analysis reveals limited association between forest habitat types and the environment in western United States land classification. Appl. Veg. Sci. 18 (1), 158–168. https://doi.org/10.1111/avsc.12121.
Liaw, A., Wiener, M., 2002. Classification and regression by RandomForest. R News 2, 18–22.
Lydersen, J.M., Collins, B.M., Knapp, E.E., Roller, G.B., Stephens, S., 2015. Relating fuel loads to overstorey structure and composition in a fire-excluded Sierra Nevada mixed conifer forest. Int. J. Wildland Fire 24, 484–494. https://doi.org/10.1071/WF13066.
McKenzie, D., Halpern, C.B., Nelson, C.R., 2000. Overstory influences on herb and shrub communities in mature forests of western Washington, U.S.A. Can. J. For. Res. 30, 1655–1666. https://doi.org/10.1139/x00-091.
Moeur, M. 1985. COVER: a user's guide to the CANOPY and SHRUBS extension of the Stand Prognosis Model. General Technical Report INT-190. Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Research Station. p. 49.
Nicodemus, K.K., Malley, J.D., Strobl, C., Ziegler, A., 2010. The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC Bioinf. 11, 110. https://doi.org/10.1186/1471-2105-11-110.
O'Connell, B.M., LaPoint, E.B., Turner, J.A., Ridley, T., Boyer, D., Wilson, A.M., Waddell, K.L., Pugh, S.A., Conkling, B.L. 2013. The Forest Inventory and Analysis Database: Database Description and Users Manual Version 5.1.6 for Phase 2. Fort Collins, CO: U. S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 556 p (with appendices). https://www.fia.fs.fed.us/library/database-documentation/index.php (accessed 13 March 2013).
Ohmann, L.F., Grigal, D.F., Rogers, L.L. 1981. Estimating plant biomass for undergrowth species of northeastern Minnesota forest communities. General Technical Report USFS-GTR-NC-61. USDA Forest Service North Central Forest Experiment Station, St. Paul, MN. p. 10.
Olson, C.M., Martin, R.E. 1981. Estimating biomass of shrubs and forbs in Central Washington Douglas-fir stands. USDA Forest Service, Pacific Northwest Forest and Range Experimental Station Research Note PNW-380, Bend, OR. p. 5. https://doi.org/10.2737/PNW-RN-380.
Patterson, P.L., O'Brien, R.A. 2011. Understory vegetation data quality assessment for the Interior West Forest and Inventory Analysis program. Res. Pap. RMRS-RP-87. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. p. 9.
Pfister, R.D., Kovalchik, B.L., Arno, S.F., Presby, R.C. 1977. Forest habitat types of Montana. Gen. Tech. Rep. INT-34. Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Forest and Range Experiment Station. p. 174.
Pollard, J.E., Westfall, J.A., Patterson, P.L., Gartner, D.L., Hansen, M., Kuegler, O. 2006. Forest Inventory and Analysis National Data Quality Assessment Report for 2000 to 2003. Gen. Tech. Rep. RMRS-GTR-181. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. p. 43.
R Development Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/.
Roberts, D.W., 2015. Potential natural vegetation and environment: a critique of Kusbach, Shaw, and Long. Appl. Veg. Sci. 18 (4), 733–738. https://doi.org/10.1111/avsc.12177.
Ruefenacht, B., Finco, M.V., Nelson, M.D., Czaplewski, R., Helmer, E.H., Blackard, J.A., Holden, G.R., Lister, A.J., Salajanu, D., Weyermann, D., Winterberger, K., 2008. Conterminous U.S. and Alaska forest type mapping using Forest Inventory and Analysis Data. Photogramm. Eng. Remote Sens. 74 (11), 1379–1388.
Russell, M.B., D'Amato, A.W., Schulz, B.K., Woodall, C.W., Domke, G.M., Bradford, J.B., 2014. Quantifying understorey vegetation in the US Lake States: a proposed framework to inform regional forest carbon stocks. Forestry 87, 629–638. https://doi.org/10.1093/forestry/cpu023.
Sabo, K.E., Hart, S.C., Sieg, C.H., Bailey, J.D., 2008. Tradeoffs in overstory and understory aboveground net primary productivity in southwestern ponderosa pine stands. For. Sci. 54 (4), 408–416.
Schulz, B., Oswalt, S., Moser, W.K. 2008. Vegetation inventory data: How much is enough? In: McWilliams, W., Moisen, G., Czaplewski, R., comps. 2008. 2008 Forest Inventory and Analysis (FIA) Symposium, October 21-23, 2008, Park City, UT. Proc. RMRS-P-56CD. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. p. 11.
Shaw, J.D. 2009. Using FIA data in the Forest Vegetation Simulator. In: McWilliams, W., Moisen, G., Czaplewski, R., comps. 2009. 2008 Forest Inventory and Analysis (FIA) Symposium, October 21-23, 2008, Park City, UT. Proc. RMRS-P-56CD. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. p. 16.
Sigurdsson, B.D., Magnusson, B., Elmarsdottir, A., Bjarnadottir, B., 2005. Biomass and composition of understory vegetation and the forest floor carbon stock across Siberian larch and mountain birch chronosequences in Iceland. Ann. For. Sci. 62, 881–888. https://doi.org/10.1051/forest:2005079.
Suchar, V.A., Crookston, N.L., 2010. Understory cover and biomass indices predictions for forest ecosystems of the Northwestern United States. Ecol. Ind. 10, 602–609. https://doi.org/10.1016/j.ecolind.2009.10.004.
Tang, F., Ishwaran, H., 2017. Random forest missing data algorithms. Statist. Anal. Data Mining: ASA Data Sci. J. 10, 363–377. https://doi.org/10.1002/sam.11348.
Timilsina, N., Cropper Jr., W.P., Escobedo, F.J., Lima, J.M.T., 2013. Predicting understory species richness from stand and management characteristics using regression trees. Forests 4, 122–136. https://doi.org/10.3390/f4010122.
Tonteri, T., Salemaa, M., Rautio, P., Hallikainen, V., Korpela, L., Merilä, P., 2016. Forest management regulates temporal change in the cover of boreal plant species. For. Ecol. Manage. 381, 115–124. https://doi.org/10.1016/j.foreco.2016.09.015.
Uresk, D.W., Severson, K.E., 1989. Understory-overstory relationships in ponderosa pine forests, Black Hills, South Dakota. J. Range Manage. 42 (3), 203–208. https://doi.org/10.2307/3899473.
USDA Forest Service. 2013. PNW-FIADB Annual Inventory Database 2001–2011. Release February 11, 2013. Portland, OR, USDA Forest Service, Forest Inventory and Monitoring program. https://www.fs.fed.us/pnw/rma/fia-topics/inventory-data/index.php (accessed April 2013).
USDA Forest Service. 2014. Forest Inventory and Analysis national core field guide. Volume 1: Field data collection procedures for Phase 2 plots, version 6.1. Washington, D.C. USDA Forest Service, Forest Inventory and Analysis program. p. 433. https://www.fia.fs.fed.us/library/field-guides-methods-proc/docs/2014/Core%20FIA%20field%20guide_6-1.pdf (accessed 30 March 2015).
USDA Forest Service. 2018. Forest Inventory and Analysis Database, St. Paul, MN: U.S. Department of Agriculture, Forest Service, Northern Research Station. https://apps.fs.usda.gov/fia/datamart/datamart.html (accessed 16 April 2018).
Veldman, J.W., Mattingly, W.B., Brudvig, L.A., 2013. Understory plant communities and the functional distinction between savanna trees, forest trees, and pines. Ecology 94

(2), 424–434. https://doi.org/10.1890/12-1019.1.
Verschuyl, J., Clark, L., Loehle, C., 2018. Predicting shrub biomass and current annual growth from field measurements in the Oregon Coast Range. Northwest Sci. 92 (1), 9–17. https://doi.org/10.3955/046.092.0103.
Visscher, D.R., Merrill, E.H., Fortin, D., Frair, J.L., 2006. Estimating woody browse availability for ungulates at increasing snow depths. For. Ecol. Manage. 222, 348–354. https://doi.org/10.1016/j.foreco.2005.10.035.
Warren, R.J., 2008. Mechanisms driving understory evergreen herb distributions across slope aspects: as derived from landscape position. Plant Ecol. 198 (2), 297–308. https://doi.org/10.1007/s11258-008-9406-1.
White, A.B., Kumar, P., Tcheng, D., 2005. A data mining approach for understanding topographic control on climate-induced inter-annual vegetation variability over the United States. Remote Sens. Environ. 98, 1–20. https://doi.org/10.1016/j.rse.2005.05.017.
Winter, S., Chirici, G., McRoberts, R.E., Hauk, E., Tomppo, E., 2008. Possibilities for harmonizing national forest inventory data for use in forest biodiversity assessments. Forestry 81 (1), 33–44. https://doi.org/10.1093/forestry/cpm042.
Woodall, C.W., 2012. Where did the U.S. forest biomass/carbon go? J. Forest. 110, 113–114.
Zar, J.H. 1999. Biostatistical Analysis, fourth edition. Prentice Hall, Upper Saddle River, NJ. p. 662, plus appendices.