esa

ECOSPHERE

Mapping forest vegetation for the western United States using modified random forests imputation of FIA forest plots

Karin L. Riley,[1,]† Isaac C. Grenfell,[2] and Mark A. Finney[2]

[1]*Forestry Sciences Laboratory, Rocky Mountain Research Station, U.S. Forest Service, 800 East Beckwith, Missoula, Montana 59801 USA*
[2]*Missoula Fire Sciences Laboratory, Rocky Mountain Research Station, U.S. Forest Service, 5775 Highway 10 West, Missoula, Montana 59812 USA*

Citation: Riley, K. L., I. C. Grenfell, and M. A. Finney. 2016. Mapping forest vegetation for the western United States using modified random forests imputation of FIA forest plots. Ecosphere 7(10):e01472. 10.1002/ecs2.1472

Abstract. Maps of the number, size, and species of trees in forests across the western United States are desirable for many applications such as estimating terrestrial carbon resources, predicting tree mortality following wildfires, and for forest inventory. However, detailed mapping of trees for large areas is not feasible with current technologies, but statistical methods for matching the forest plot data with biophysical characteristics of the landscape offer a practical means to populate landscapes with a limited set of forest plot inventory data. We used a modified random forests approach with Landscape Fire and Resource Management Planning Tools (LANDFIRE) vegetation and biophysical predictors to impute plot data collected by the US Forest Service's Forest Inventory Analysis (FIA) to the landscape at 30-m grid resolution. This method imputes the plot with the best statistical match, according to a "forest" of decision trees, to each pixel of gridded landscape data. In this work, we used the LANDFIRE data set for gridded input because it is publicly available, offers seamless coverage of variables needed for fire models, and is consistent with other data sets, including burn probabilities and flame length probabilities generated for the continental United States. The main output of this project is a map of imputed plot identifiers at 30 × 30 m spatial resolution for the western United States that can be linked to the FIA databases to produce tree-level maps or to map other plot attributes. In addition, we used the imputed inventory data to generate maps of forest cover, forest height, and vegetation group at 30 × 30 m resolution for all forested pixels in the western United States, as a means of assessing the accuracy of our methodology. The results showed good correspondence between the target LANDFIRE data and the imputed plot data, with an overall within-class agreement of 79% for forest cover, 96% for forest height, and 92% for vegetation group.

Key words: Forest Inventory Analysis; imputation; LANDFIRE; random forests; tree list.

Received 1 June 2016; accepted 9 June 2016. Corresponding Editor: D. P. C. Peters.

† **E-mail:** kriley@fs.fed.us

Introduction

Geospatial data describing tree species or forest structure are required for many analyses and models of forest landscape dynamics, including estimating stocks of terrestrial carbon (Jenkins et al. 2001), forest biomass (Blackard et al. 2008), forest growth and mortality (Brown and Schroeder 1999, Falkowski et al. 2010), national-level fire planning and risk assessment (Schmidt et al. 2002), simulating continental-scale burn probabilities (Finney et al. 2011), simulating wildfire intensity patterns and fuel treatment strategies (Finney et al. 2007), tree species abundance and distribution (Wilson et al. 2012), basal area (Wilson et al. 2012), estimating timber volume (Franco-Lopez et al. 2001, Muinonen et al. 2001), mapping wildland fuels for simulating fire growth (Keane et al. 2001), and

simulating wildfire risk transmission from federal lands to the wildland–urban interface (Haas et al. 2014). Forest data must have resolution and continuity sufficient to reflect site gradients in mountainous terrain and stand boundaries imposed by historical events, such as wildland fire and timber harvest. Such detailed forest structure data are not available for large areas of public and private lands in the United States, which rely on forest inventory at fixed plot locations at sparse densities of one plot per 6000 acres (Burkman 2005). While direct sampling technologies such as light detection and ranging (LiDAR) may eventually make broad coverage of detailed forest inventory feasible (e.g., Hudak et al. 2008, Latifi et al. 2012), no such data sets at the scale of the western United States are currently available.

Models of geospatial forest structure have utilized various statistical methods to assign the measured plots from a sparse sample to unmeasured locations using a set of predictor variables. These methods have a common goal: to take more detailed observations at relatively few locations (e.g., field plots or stand inventories) and assign their characteristics to the unmeasured locations on the landscape in order to provide seamless information about all locations. The detailed observations are often referred to as the "reference data," while the landscape data (often derived from aerial photographs, stand records, or satellite imagery) are referred to as the "target data." These methods include linear models, image classification (Van Wagtendonk and Root 2003), classification and regression trees (CART), kriging (Krige 1951), universal kriging (UK; Cressie 1990), and an assortment of nearest neighbor methods, including normalized and unnormalized Euclidean distance, Mahalanobis distance, independent component analysis (ICA), most similar neighbor (MSN, also called canonical correlation analysis), gradient nearest neighbor (GNN, also called canonical correspondence analysis), and random forests (RF; Moeur and Stage 1995, Pierce et al. 2009, Hudak et al. 2008, Wilson et al. 2012, Breiman 2001). A distinguishing factor among all of these methods is whether they allow for the use of categorical predictor variables as well as continuous variables.

Linear models use the values of one or more predictor variables and a set of coefficients to predict the value of a response variable. Most classification and regression tree (CART) analyses use a look-up table or classification rules to match a response variable with input characteristics (Breiman 2001, Pierce et al. 2009, Rollins 2009). Random forests uses a set of decision trees (a "forest") to predict which among the reference data (e.g., forest plots) are most similar to the characteristics at a target location, including both continuous and categorical variables (Cutler et al. 2007). Kriging is a form of interpolation, using "a Gaussian process governed by prior covariances" (Krige 1951). Universal kriging is similar to kriging, but with a local trend; it can be viewed as a point interpolation, using a point map as input and returning a raster map with estimations (Cressie 1990). Nearest neighbor methods represent a more recent approach to mapping forest attributes, and use a set of numerical predictors (often spectral and environmental continuous variables) to assess which of the candidate reference data (e.g., field plots) are most similar to each target (map) location (Pierce et al. 2009). Using continuous variables only, most similar neighbor (MSN) uses a similarity measure that employs canonical correspondence analysis to summarize the multivariate relationships between the set of target data and the set of reference data derived from field samples (Moeur and Stage 1995). Similarly, gradient nearest neighbor (GNN) imputation also utilizes canonical correspondence analysis, but incorporates the direct gradient analysis in assigning weights to predictor variables (Ohmann and Gregory 2002, Pierce et al. 2009). GNN and MSN techniques assign values at each target location that are the original values measured at a field plot, while regression-based and interpolation methods assign the modeled values (Pierce et al. 2009). The GNN and MSN techniques give users the option to choose multiple nearest neighbors in order to predict continuous variables, as in k-nearest neighbor techniques (e.g., McRoberts 2009, Wilson et al. 2012).

Recent studies have used a variety of these methods to impute forest structure variables or forest plots to target data, and have in some cases compared the robustness of various methods. In a project focused on estimating the tree mortality, Drury and Herynk (2011) used a CART procedure to create a national-scale 30-m grid of tree-list plots having the median bark thickness

within stratifications of existing vegetation type, biophysical setting, succession class, and canopy bulk density. Among several methods, Pierce et al. (2009) found that GNN performed best for forest structure variables in Oregon, while linear models and universal kriging demonstrated stronger performance for both forest structure and canopy variables in Washington and California. Among the nearest neighbor methods, Hudak et al. (2008) found that random forests produced the best results for predicting the plot-level basal area and tree density and was the most robust and flexible method for a study area in north-central Idaho.

In this study, we describe the methods for developing a tree-list data set for the purpose of using forest inventory data with national-scale wildfire simulations, among others. Potential uses of this tree list include estimating the effect of fuel treatments and wildfires on mortality and carbon stores. For use in these research applications, the tree-list data set needed to be compatible with several other existing data sets: (1) Landscape Fire and Resource Management Planning Tools (LANDFIRE) vegetation and fuels data, which provide landscape inputs to the wildfire simulations (Rollins 2009; data available at www.landfire.gov), and (2) outputs of the wildfire simulations, including continental-scale burn probability grids and flame length grids from Fire Program Analysis project (FPA; Finney et al. 2011). The LANDFIRE program is shared between the wildland fire management programs of the U.S. Department of the Interior and U.S. Department of Agriculture and was precipitated by the National Fire Plan of 2000, which tasked LANDFIRE with producing landscape-scale geospatial products to support planning, operations, and management across land ownership boundaries. Major advantages of LANDFIRE data are that they provide seamless coverage of variables necessary for fire models across the United States, and the data are publicly available. LANDFIRE vegetation and fuels data serve as inputs for a number of fire modeling efforts, including the above-referenced continental runs of the Large Fire Simulator (FSim) conducted by Fire Program Analysis (FPA). Thus, LANDFIRE data were chosen as target data for the tree list, so that the tree list would be consistent with the outputs of previous fire modeling efforts, and could be used in future research, including the calculation of risk to terrestrial carbon resources from wildfire. However, the level of detail in the LANDFIRE data is not sufficient for applications such as basal area determination and treatment effectiveness. Hence, the current effort to produce a tree-list data set links each LANDFIRE pixel to an FIA plot, thus enabling users to link to the extensive database for each FIA plot, which includes the number, size, and species of each tree, among many other attributes.

With the requirement for having compatibility between the tree-list data and the vegetation and fuels data provided by LANDFIRE, most of the methods listed above were precluded. For example, kriging would model the values based on the point plot data, but would not have good correspondence with the LANDFIRE data. In addition, the vegetation group data are categorical, while other predictor variables are numerical and continuous. This limits the set of possible methodologies to classification trees (i.e., Pierce et al. 2009). We chose random forests as our methodology because it leverages a "forest" of classification trees in order to produce high accuracies and model complex interactions among predictor variables, two notable strengths of this methodology over other statistical classifiers (Breiman 2001, Cutler et al. 2007). The modified random forests method used here evaluates a set of forest plots and identifies the best-matching plot for each grid cell on the landscape. Several important differences exist between the random forests methodology used here and that of Drury and Herynk (2011): (1) We limited our scope to a single set of nationally consistent plot data, whereas they obtained a variety of fixed- and variable-radius plot designs from multiple agencies; (2) because tree mortality was not the primary variable of interest, we did not use it as a predictor; and (3) we wanted to identify a single best-matching plot for each point on the landscape rather than utilizing the median plot in a class, thus retaining more variability on the landscape.

Given our objective to find the single best-matching plot for each pixel on the landscape, we optimized our model for a set of response variables linked to the prediction of terrestrial carbon: forest cover, forest height, and existing vegetation group (EVG). Here, we demonstrate

high model accuracies and high levels of agreement between the target (LANDFIRE) and imputed data for a random forests imputation run on all 997,153,322 forested pixels in the western United States at 30-m grid resolution. The primary output of this project is a raster grid of plot identifiers, which can in turn be used to generate a list of the number, size, and species of trees assigned to each pixel.

Methods

In this section, we first describe our data sources, then present a brief description of the modified random forests methodology and how we applied it specifically to our problem. We finish by offering a description of the methods we used for verifying our outputs.

Data sources

Forest Inventory Analysis forest plot data.—We obtained the measurements of tree size, height, species, and status (dead or alive) from the US Forest Service's Forest Inventory Analysis (FIA). FIA measures the forest attributes on a network of plots in all 50 states using a standardized plot design that was implemented beginning in 1999 (Fig. 1) (O'Connell et al. 2014). Version 5.1 data were downloaded from the FIA Data Mart (http://apps.fs.fed.us/fiadb-downloads/datamart.html, on 11 April 2012). We restricted the plot data to single-condition forested plots only; conditions are defined by changes in vegetation or land use, and some plots have more than one condition because of harvesting or fire, for example (O'Connell et al. 2014). Because we wanted the plots used in imputation to be more or less homogenous, we used single-condition plots only.

The LANDFIRE Reference Database.—The subset of FIA plots used in this study were further restricted to those in the LANDFIRE Reference Database (LFRDB), a database of plots leveraged by LANDFIRE in the production of their spatial data sets. The LFRDB was the sole source of three stand-level descriptions not available from the native FIA data: existing vegetation cover (EVC), existing vegetation height (EVH), and existing vegetation group (EVG), which were assigned to FIA plots by the LANDFIRE project based on geographic location and the characteristics of the trees recorded. Existing vegetation group describes the ecological system (NatureServe 2009). Existing vegetation cover represents the vertically projected percent cover of the live canopy layer. Existing vegetation height is the average height of the dominant vegetation. Once plots were restricted to single-condition forested plots that appear in the LFRDB, 15,333 plots in the western United States were available for imputation.

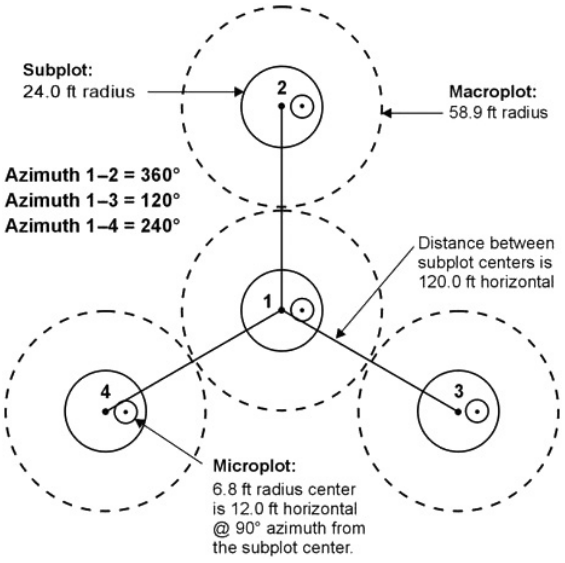

Fig. 1. The nationally standard FIA Phase 2 plot design (from O'Connell et al. 2014). Trees larger than 12.7 cm in diameter at breast height are measured on the four 7.3 m radius subplots, and trees less than 12.7 cm in diameter are measured on the nested 2.1 m radius microplots. Conifer seedlings at least 15.2 cm tall and hardwood seedlings at least 30.5 cm tall are measured on microplots. In some areas, larger trees are measured on the 18.0 m radius macroplots, in order to avoid under- or overcounting of large rare trees.

LANDFIRE Target Landscape Data.—LANDFIRE provides a suite of topographic, biophysical, and vegetation data at 30-m grid resolution for the western United States that served as the target data for this project. Existing vegetation group is assigned to each pixel using a set of hierarchical and iterative CART models, Landsat imagery, biophysical gradients, and training databases developed from the LFRDB (Rollins 2009). Existing vegetation cover and height are mapped using regression tree-based models, empirical models, and spectral mixture models, leveraging spectral information from Landsat imagery and

land cover information from the National Land Cover Database (Homer et al. 2004).

Random forests imputation

In this modified random forests imputation, a set of reference observations comprising the FIA plot data was imputed to a set of target points corresponding to the center of each 30-m pixel of the LANDFIRE landscape grid (Crookston and Finley 2008). The output consists of a raster grid attributed with the best-matching plot ID for each pixel (Fig. 2). The modified random forests model was created by inputting the forest plot data to the yaImpute package in the statistical program R (Crookston and Finley 2008, R Foundation 2016). We refer to our method as a modified random forests approach because yaImpute adapts the randomForest package in several ways, most notably: (1) by using the "nodes" matrix directly to compute proximity without the necessity of holding the often-large proximity matrix in memory and (2) it is possible to have more than one response variable. We chose to use the modified random forests approach because our study design ideally involved the use of more than one response variable. However, although we did utilize the modified random forests approach as coded in yaImpute, we often refer to our method in the remainder of the manuscript as simply "random forests" for the sake of brevity.

Random forests requires all predictor variables to be available for both the reference and target data, which greatly constrains the list of possible variables. After examining the variables used in other imputation efforts and testing a more extensive list of predictor variables, we chose to include the variables listed in Table 1 based on their variable importance scores, the expected relevance to forest characteristics, and the lack of redundancy with other predictor variables: three topographic variables (slope, aspect, and

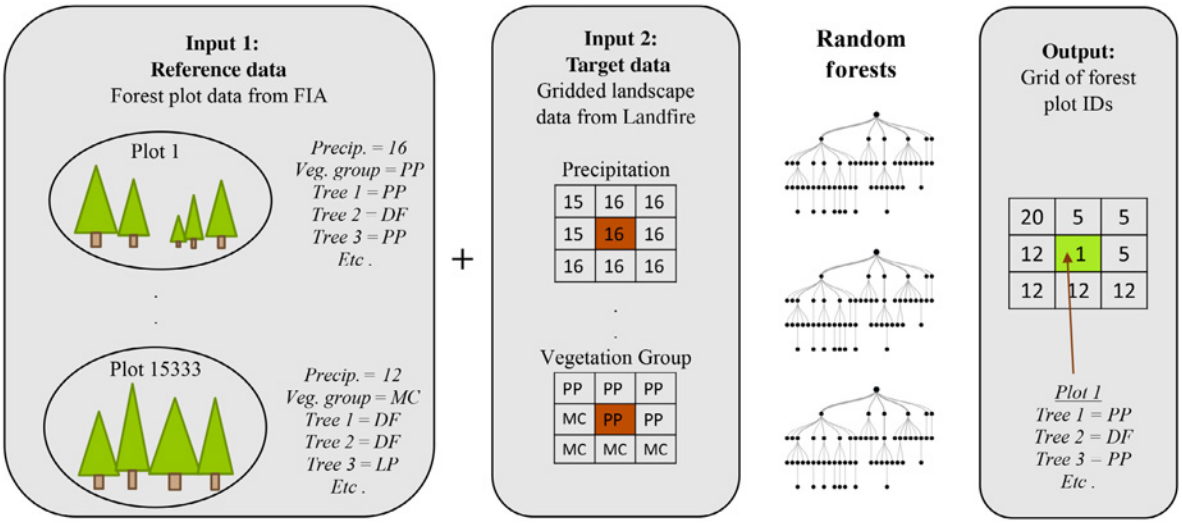

Fig. 2. Random forests requires two sets of input data, which are referred to as the reference data and the target data. The reference data refer to the more detailed observations at sparse points on the landscape, in this case, the FIA plot data. The target data represent the whole landscape and are often derived from satellite data; in this case, we used the raster LANDFIRE data. The same set of predictor variables must be available for both the reference and target data. In this simplified schematic example, we used only two predictors: precipitation and vegetation group. The goal of the random forests imputation is, for each pixel of raster data, to find the forest plot that is the best match, based on a suite of predictor variables. In this example, the characteristics of plot 1 are obviously much closer to the characteristics of the center pixel than those of plot 15,333, so plot 1 would be matched (or "imputed") to the center pixel. The output from our random forests method is a raster grid of forest plot ID numbers, which can be linked to characteristics of the plots, including the number, size, and species of trees.

Table 1. Predictor variables for reference (FIA plots) and target (LANDFIRE raster) data.

Category	Variable	Source for reference (FIA plot) data	Source for target (LANDFIRE raster) data
Topographic	Slope	FIADB	LANDFIRE National topographic layers
	Aspect (sine and cosine)	"	"
	Elevation	"	"
Location	Latitude	MOC with FIA	Center of 30-m pixels for LANDFIRE Refresh 2008 layers
	Longitude	"	"
Vegetation	Existing Vegetation Cover (forest cover)	LFRDB (derived from FIA plot data)	LANDFIRE Refresh 2008 layers
	Existing Vegetation Height (forest height)	"	"
	Existing Vegetation Group (vegetation group)	"	"
Biophysical	Maximum temperature	Overlay of plot location with LANDFIRE biophysical grid	LANDFIRE biophysical grid
	Minimum temperature	"	"
	Relative humidity	"	"
	Precipitation	"	"
	Photosynthetically active radiation	"	"
	Vapor pressure deficit	"	"

elevation), two location variables (latitude and longitude), three vegetation variables (forest cover, forest height, and vegetation group), and six biophysical variables (maximum temperature, minimum temperature, relative humidity, precipitation, photosynthetically active radiation, and vapor pressure deficit).

The predictor variables for the reference (plot) data were derived as follows. FIA collects numerous variables at plots, but for the imputation we directly utilized only elevation, slope, aspect, latitude, and longitude (the latter two attributes are not publicly available, but were acquired via a Memorandum of Cooperation signed between the authors of this manuscript and FIA). The three vegetation variables are calculated based on FIA plot characteristics by the LANDFIRE program, as noted above, which calculates EVH (hereafter referred to as "forest height"), EVC ("forest cover"), and EVG ("vegetation group") for each plot in their LANDFIRE Reference Database. These data are also not publicly available, but were acquired via the Memorandum of Cooperation. A suite of biophysical predictors was derived via an overlay of the plot locations with gridded biophysical data from the LANDFIRE project (maximum temperature, minimum temperature, relative humidity, precipitation, photosynthetically active radiation, and vapor pressure deficit).

The same suite of predictor variables was obtained from various LANDFIRE raster data sets for the target (gridded) data. The three topographic variables (elevation, slope, and aspect) and the three vegetation variables (forest cover, forest height, and vegetation group) are publicly available as 30 × 30 m rasters (www.landfire.gov, version 1.2.0). The biophysical predictors are available at the same resolution from the LANDFIRE project.

Because we wanted to optimize the tree list for estimating carbon storage, we chose the response variables of forest cover, forest height, and vegetation group. Note that these also appear nominally in the list of predictor variables. Although the terms "predictor" and "response variable" are used in the descriptions of the random forests methodology, they do not have the same meaning as in other statistical approaches where the predictor variables are used to predict the value of the response variable. In random forests, the predictor and response variables are used to find the associations among the reference data and to find which observations are most like one another. In that sense, a variable can serve as both predictor and response without conflict. It is important that

the response variable should not be used also as a predictor variable for the target data, and our methodology meets this criterion, as the three response variables were derived using different data sources and methodologies than were used to derive the predictor variables for the target data: For the reference data, the response variables we call forest cover, forest height, and vegetation group were based on the characteristics of the trees in the FIA plots, and for the gridded target data, the predictor variables we call forest cover, forest height, and vegetation group were based on satellite imagery and land cover data, among other inputs. We expect that using several of the predictor variables from the reference data as response variables will have the effect of increasing accuracy for these three variables over more traditional approaches, an innovation of our approach. It is also important to note that although the response variables are forest cover, forest height, and vegetation group, that is not what we are predicting: We are predicting the plot that is the best match for each pixel and outputting a map of plot IDs, which users can link to data in the FIA databases.

At the time of this study, the randomForest package in R had a maximum of 32 classes of data it could use, and there were more than 32 vegetation groups present in the plot data; therefore, we performed the imputation on one zone of LANDFIRE data at a time, which always limited the number of classes to less than 32 (Fig. 3). From the list of 15,333 plots in the western United States, we created a subset consisting only of the plots with vegetation groups appearing in each zone. For example, when performing the imputation for zone 9, we limited the population of plots available for imputation to only those plots with vegetation groups that LANDFIRE had mapped in zone 9. Then, we formed the random forest model using the suite of predictor variables listed in Table 1. For each zone, we employed 249 total decision trees, with these trees divided equally among the three response variables: In other words, there were 83 decision trees to predict forest height, 83 for forest cover, and 83 for vegetation group. We decided on 249 trees after finding that the error rates barely differed from those of parameterizing the model using 500 trees, but saved a significant amount of computational time. A short description of how each decision tree is constructed follows, based on Cutler et al. (2007). Each decision tree is formed using a random sample of 66% of the plots, and the remainder (referred to as the out-of-bag observations) are set aside to assess the accuracy. Then, the bootstrap sample is divided into two groups in a process called binary partitioning. In order to determine how to partition the data, a small number of randomly selected variables (in this case, the square root of the number of predictor variables) are examined to see which best minimizes the variance in the response variable. That variable is chosen, at a point called a "node," and the bootstrap sample is divided into two groups, or "buckets." Binary partitioning continues until the variance in each bucket cannot be reduced significantly, or until further divisions cannot be made without reducing the number of observations in a bucket to less than 5. Each "fully grown" decision tree is used to predict the out-of-bag observations, in order to assess the model's accuracy. To better illustrate this process, we show a simplified schematic of two trees in Fig. 4. In random forests, the decision trees cannot be viewed, so we cannot show one here, but Fig. 4 is provided as an illustration of how the method works.

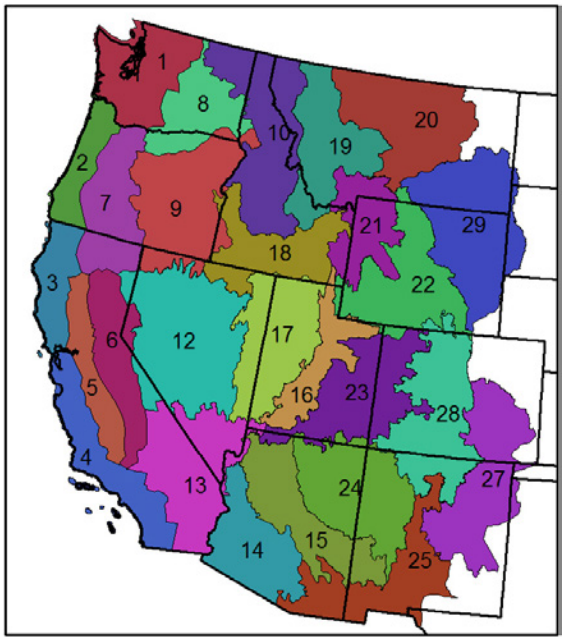

Fig. 3. The 27 LANDFIRE zones in the western United States.

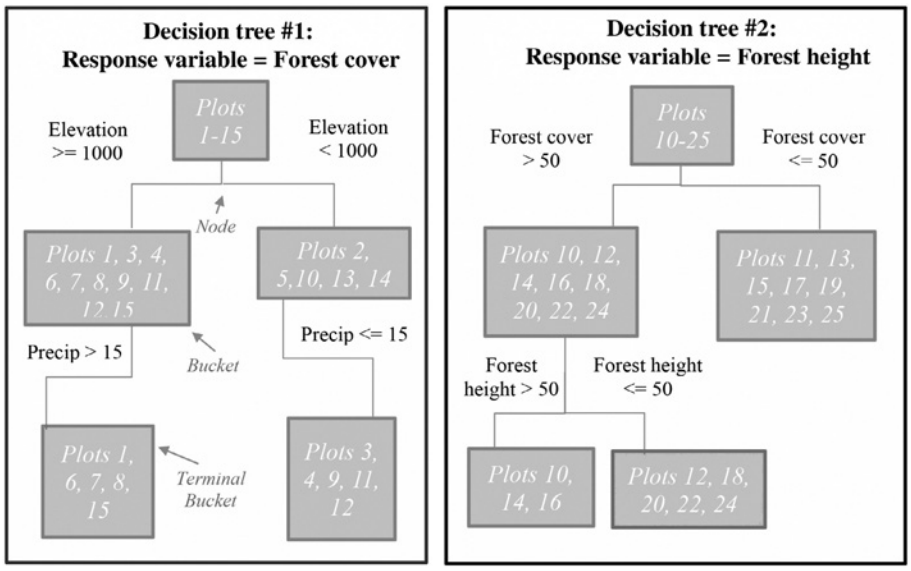

Fig. 4. A "forest" of decision trees was constructed from the forest plot observations. This figure shows a simplified schematic example of how two of the trees might look. Once the forest of trees is constructed, observations from each cell of the target data are passed down the trees, and the plots that end up in the terminal bucket (lowest) with the observation are recorded. The plot assigned to that pixel is the one that appears most frequently in the terminal bucket with it.

Once the "forest" of decision trees has been grown using the plot data, they are used to predict the best-matching plot for each pixel of the gridded target data. For a chosen pixel, its suite of predictor variables are used to determine its terminal bucket for each decision tree. The plots that appear in the same terminal bucket with the pixel observation are recorded for each of the 249 trees (recall that there are 83 decision trees to predict forest height, 83 for forest cover, and 83 for vegetation group). The plot that most frequently co-occurs with the pixel observation across all 249 trees (and thus all three response variables) is chosen as the best match for that pixel, with ties split randomly. In this project, we used random forests to find the best-matching FIA plot for each forested 30-m pixel of LANDFIRE data, imputing an FIA plot number to each pixel, and generating a raster grid of the best-matching plot numbers. The plot numbers can be linked back to the database of plot characteristics, so for any pixel on the landscape, maps can then be made of any number of plot characteristics, ranging from the response variables (cover, height, and vegetation group) to other plot characteristics that were neither predictor nor response variables (such as terrestrial carbon, basal area, or the number of trees).

We obtained an overall accuracy of the model by taking the out-of-bag misclassification for each tree and considering them in aggregate to assess the overall quality of the random forests model. We report the error rates for four randomly chosen zones in Table 2. Error rates were similar across the four zones for the three response variables. The low error rates indicate high model accuracy.

FIA data security restrictions did not allow the distribution of the tree list if any plot with its center located in a pixel was imputed to that pixel, which occurred at 3679 of the 15,333 plot locations. For these plots, we normalized each of the predictor variables to a scale of 0–1, then found

Table 2. Out-of-bag error rates for each response variable in four randomly chosen LANDFIRE zones.

Zones	Forest cover (%)	Forest height (%)	Existing vegetation group (%)
7	7.79	2.77	1.13
9	6.99	1.79	0.897
12	7.98	2.29	1.08
21	9.85	3.02	1.87

the second best-matching plot, and imputed that plot ID instead. In the research (nondistributed) version of the tree list, we retained the original plot ID values.

Fidelity Compared to LANDFIRE Attributes and random chance

Another measure of the performance of this modified random forests method was to compare the characteristics of the imputed plots to the gridded target data. Specifically, we compared the forest cover, forest height, and vegetation group of the imputed plot data with those of the gridded target data and report the percentage agreement for each LANDFIRE zone and the US West as a whole. For the US West, we also calculated and reported Cohen's kappa statistic to account for agreement by random chance in forest cover, forest height, and vegetation group, with complete agreement being $\kappa = 1$ and agreement only by random chance being $\kappa = 0$ (Cohen 1960). Bar plots were used to compare the proportion of the data in each cover, height, and vegetation class across the three data sources (FIA plots, target LANDFIRE data, and imputed data). In addition, we calculated the physical distance from each pixel center to the center of the plot that was imputed, to assess whether random forests preferentially imputed plots from nearby.

Results

Plot identification numbers were imputed at 30 × 30 m resolution to 997,153,322 forested pixels in the US West. We found that plots tend to impute to a cluster of pixels, due to similarities in the topographic and biophysical predictor variables, as clustering is not imposed by the random forests method (Fig. 5).

During fidelity assessment, we compared the values of the response variables (forest cover, forest height, and vegetation group) in the imputed plot data to the LANDFIRE target raster grids in order to obtain the estimated levels of agreement for the tree list. For all forested pixels in the US West, within-class agreement was 79% for forest cover, 96% for forest height, and 92% for vegetation type. In addition, agreement for these variables is high in most zones (Table 3).

Forest height

LANDFIRE maps forest height in five classes: 0–5 m, 5–10 m, 10–25 m, 25–50 m, and greater than 50 m. The LANDFIRE organization also computes the height of FIA forest plots in its LFRDB to tenths of a meter. We compared the height of each imputed plot to the height class mapped by LANDFIRE for the corresponding pixel. Overall within-class agreement for height varied between 82% and 100% across the 27 completed zones (Table 3). For the western United States as a whole, the overall percentage agreement was 96% and Cohen's kappa was 0.93. The imputation reproduced the patterns in the gridded LANDFIRE data (Fig. 6).

The proportion of the landscape in each height class was similar across the LANDFIRE data, the imputed data, and the FIA plots (Fig. 7). However, the distribution of height classes was more similar between the LANDFIRE and imputed data than in the FIA data, with the LANDFIRE and imputed data somewhat underrepresenting the lowest height class (0–5 m) and overrepresenting the 5- to 10-m height class compared with the FIA plot data. As the FIA data constitute a random sample of forested points on the landscape, they likely represent the proportion of height classes present on the landscape quite well. It is not surprising that the imputed data would better match the LANDFIRE data, however, because the LANDFIRE gridded data were used as target data in this project. The close correspondence between the three distributions indicates that LANDFIRE data accurately capture the distribution of height classes present in a random sample of the landscape (as conveyed by the FIA data) and that the imputed data correspond closely with the LANDFIRE target data, an indication of high model accuracy. Within-class agreement was related to the number of plots in a height class, with rarer classes having lower agreement rates (Fig. 8).

Forest cover

Forest cover is mapped in nine classes by LANDFIRE: 10–19%, 20–29%, 30–39%, 40–49%, 50–59%, 60–69%, 70–79%, 80–89%, and 90–100%, with areas of tree cover less than 10% not considered forested. For FIA plots, forest cover is estimated to the nearest percentage in the LFRDB. In the 27 LANDFIRE zones in the western United

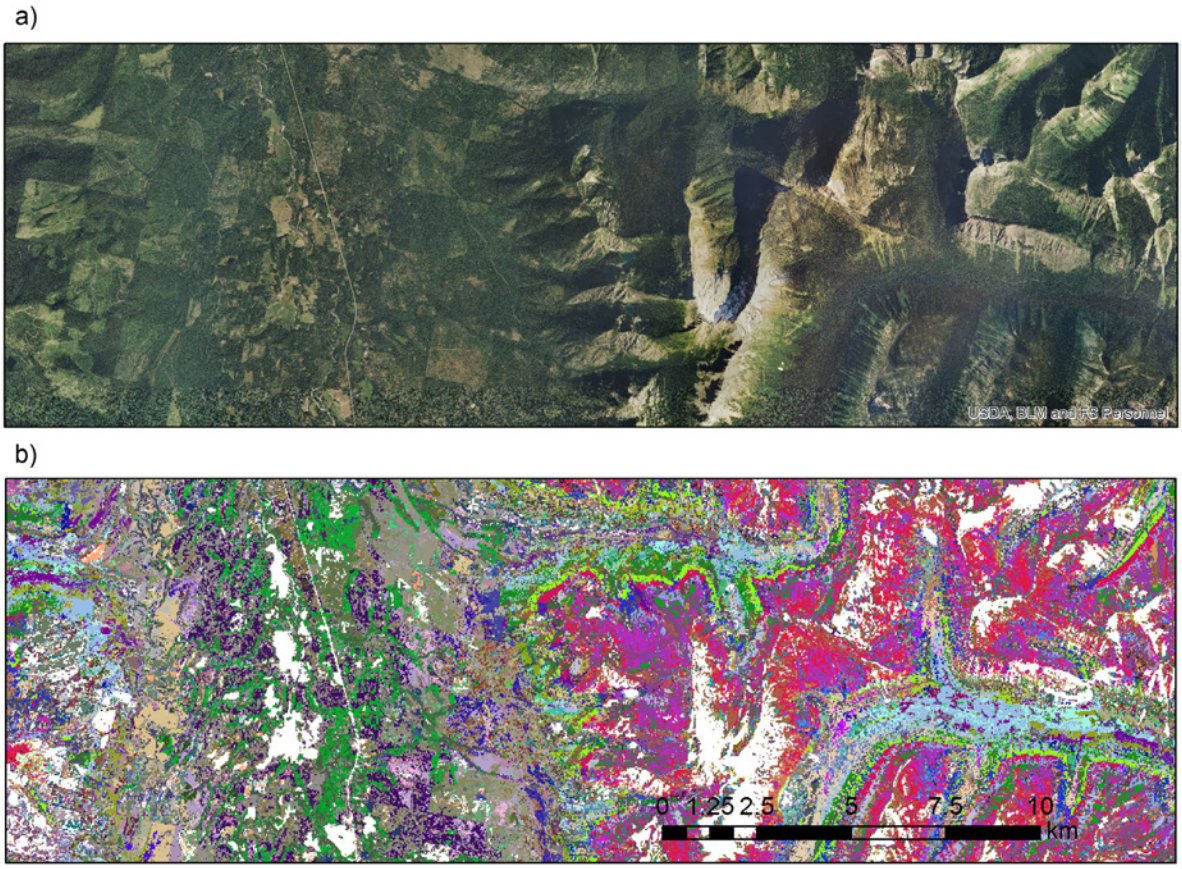

Fig. 5. A subset of the landscape in zone 19, the Swan Valley of Montana. The top panel is NAIP 1-m imagery provided by the US Forest Service image server. The lower panel shows the plot IDs for the plots most frequently assigned to each pixel, with each color representing a unique plot imputed to the same subset of the landscape as in the top panel (white signifies nonforested pixels). In the left half of the imagery, it is apparent that the landscape is dominated by a checkerboard pattern, the legacy of extensive timber harvest on private lands, and less extensive harvest on public lands. On the right side of the imagery, vegetation is dominated by topographic gradients in a mountainous landscape. The imputation was able to pick up these patterns, with the outline of the checkerboard visible in the left half of the lower panel and the topographic gradients visible in the clustering of the plots on the right half of the lower panel.

States, overall within-class agreement for cover varied between 40% and 95% (Table 3). For the western United States as a whole, percentage agreement was 79% and Cohen's kappa was 0.75. In general, landscape patterns of forest cover in the target data were well reproduced by the imputed data (Fig. 9). The proportion of the landscape in each cover class was similar for the LANDFIRE target data, the imputed data, and the FIA plots (Fig. 10). The imputed data, however, underrepresented the lowest cover class (10–19%) compared with the FIA and LANDFIRE data. The number of plots in a cover class affected the within-class agreement rates, with rarer cover classes having lower rates of agreement (Fig. 11).

Vegetation group

The third response variable, vegetation group, is mapped to the gridded target data by LANDFIRE, and assigned to FIA forest plots in the LFRDB. Note that all vegetation groups appear in the LANDFIRE and imputed data appear in the FIA plots (n = 36), but not all vegetation groups represented by FIA plots appear in

Table 3. Within-class agreement in percentage between LANDFIRE target data and imputed plot data, summarized by LANDFIRE zones in the US West.

Zones	Forest cover	Forest height	Vegetation group	No. of pixels
z01	68	90	92	78,355,830
z02	73	82	91	39,468,642
z03	60	91	94	35,678,822
z04	57	94	82	19,058,444
z05	65	95	88	4,431,257
z06	70	95	92	59,959,878
z07	73	96	92	73,522,690
z08	40	93	87	3,897,792
z09	86	97	84	44,138,635
z10	80	96	97	113,675,398
z12	92	99	94	32,167,916
z13	90	85	96	2,311,150
z14	92	100	97	695,799
z15	85	98	93	58,799,790
z16	84	100	93	42,812,230
z17	95	99	95	23,553,327
z18	86	92	63	7,734,012
z19	87	99	96	57,959,737
z20	59	96	83	11,108,613
z21	81	96	92	45,645,598
z22	88	98	48	5,777,871
z23	92	100	96	38,898,791
z24	91	98	97	36,193,270
z25	70	98	91	18,027,523
z27	85	97	95	12,839,794
z28	78	98	93	103,288,833
z29	77	97	91	27,151,679
Overall	79	96	92	997,153,322

Note: Results are reported for the three response variables: forest cover, forest height, and vegetation group.

the LANDFIRE (n = 31) or imputed (n = 30) data. In other words, FIA plots could be keyed to a vegetation group that did not appear in the gridded LANDFIRE data (details are given in Table 4). Vegetation group 701 (introduced riparian vegetation) appeared in the gridded LANDFIRE data, but only one FIA plot keyed to this vegetation group, and it was never used in the imputation, as it was presumably not a good match for the pixels where it appeared in the LANDFIRE data in terms of the other predictors (cover, height, x, y, and biophysical variables).

Overall agreement for vegetation group varied between 48% and 97% across the 27 zones (Table 3). For the western United States, within-class agreement was 92% and Cohen's kappa was 0.92. In general, the random forests imputation accurately reproduced the patterns in vegetation group in the LANDFIRE data (Fig. 12), but the rates of agreement for the Western Riparian Woodland and Shrubland category (group 635) were low in many zones. This category had few plots (n = 135), and most of these plots were located in mesic coastal areas of Washington and Oregon. Hence, random forests rarely imputed them in the drier colder continental sections of the US West and instead tended to impute the plots with other vegetation types common to the area. Even with this limitation, the proportion of the landscape in each vegetation group was similar across the FIA plots, LANDFIRE, and imputed data (Fig. 13). Vegetation group 615 (Douglas-fir–Western Hemlock Forest and Woodland) was somewhat overrepresented in both the LANDFIRE and imputed data, while 622 (Lodgepole Pine Forest and Woodland) was somewhat underrepresented in both the LANDFIRE and imputed data compared with the FIA plots. Interestingly, the LANDFIRE data overrepresented 635 (Western Riparian Woodland and Shrubland) compared with the FIA data, while the proportion imputed by random forests was similar to that of the FIA plots.

Similar to the results for height and cover, agreement was lower in rarer classes of vegetation group, although there were exceptions (Fig. 14). This result makes sense, because it is unlikely in rare types that random forests can match all three of the response variables (forest cover, height, and vegetation group) when choosing from a limited pool of candidate forest plots, and must in essence choose which of these response variables is most important to match.

Distance

In most cases, random forests chose nearby plots for imputation to a pixel. In some cases, likely when a rare plot was required, the plots were imputed from over 1500 km away (Fig. 15). Distance from the pixel center to the imputed plot center is shown for a subset of the landscape in Fig. 16. Nearby plots are preferred for imputation not only because of similarity in the x and y coordinates, but because of the similarity in biophysical variables, which indicate the similarity in site descriptors including species composition, site productivity, disturbance history and regime, and climatic characteristics that were not directly included in the imputation.

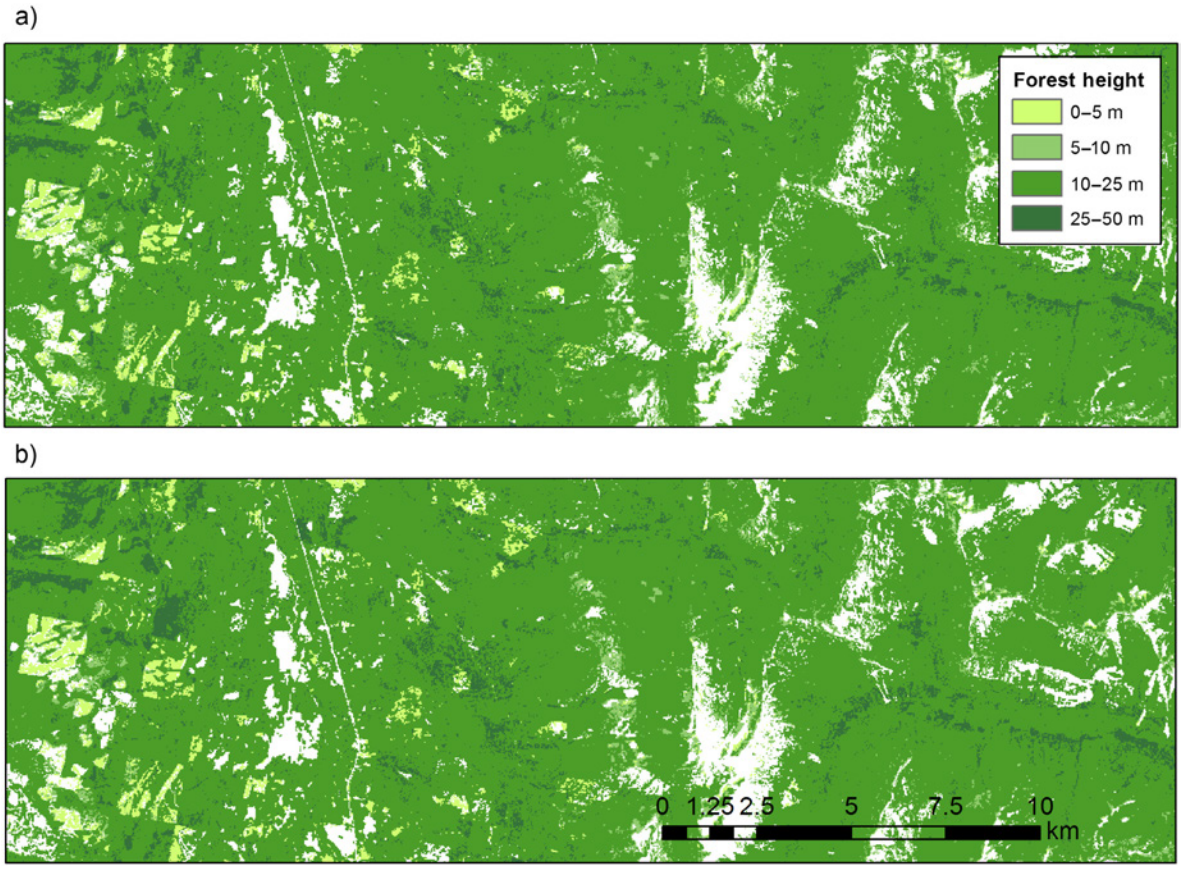

Fig. 6. Forest height class is shown below for (a) the imputed tree-list data and (b) the LANDFIRE data. High within-class agreement is evident, as is the checkerboard pattern that also appeared in the NAIP imagery at left and the topographic gradients at right.

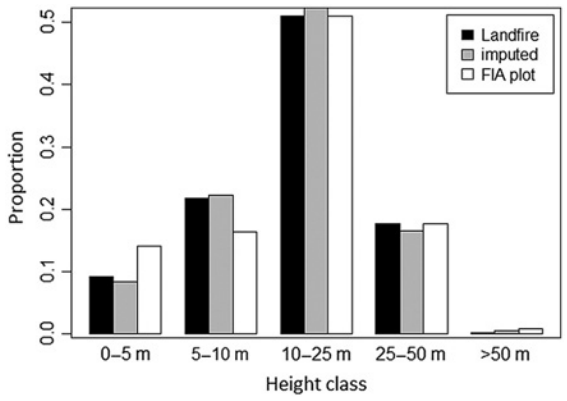

Fig. 7. The proportion of the plots or pixels in each height class for the LANDFIRE target data, the imputed data, and the FIA plots in the western United States.

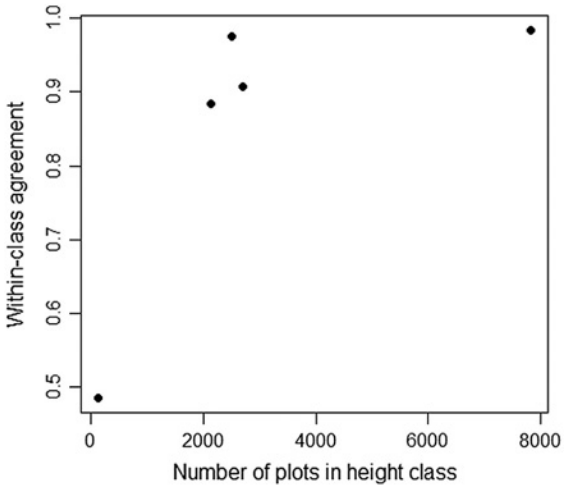

Fig. 8. Within-class agreement vs. the number of plots in each height class.

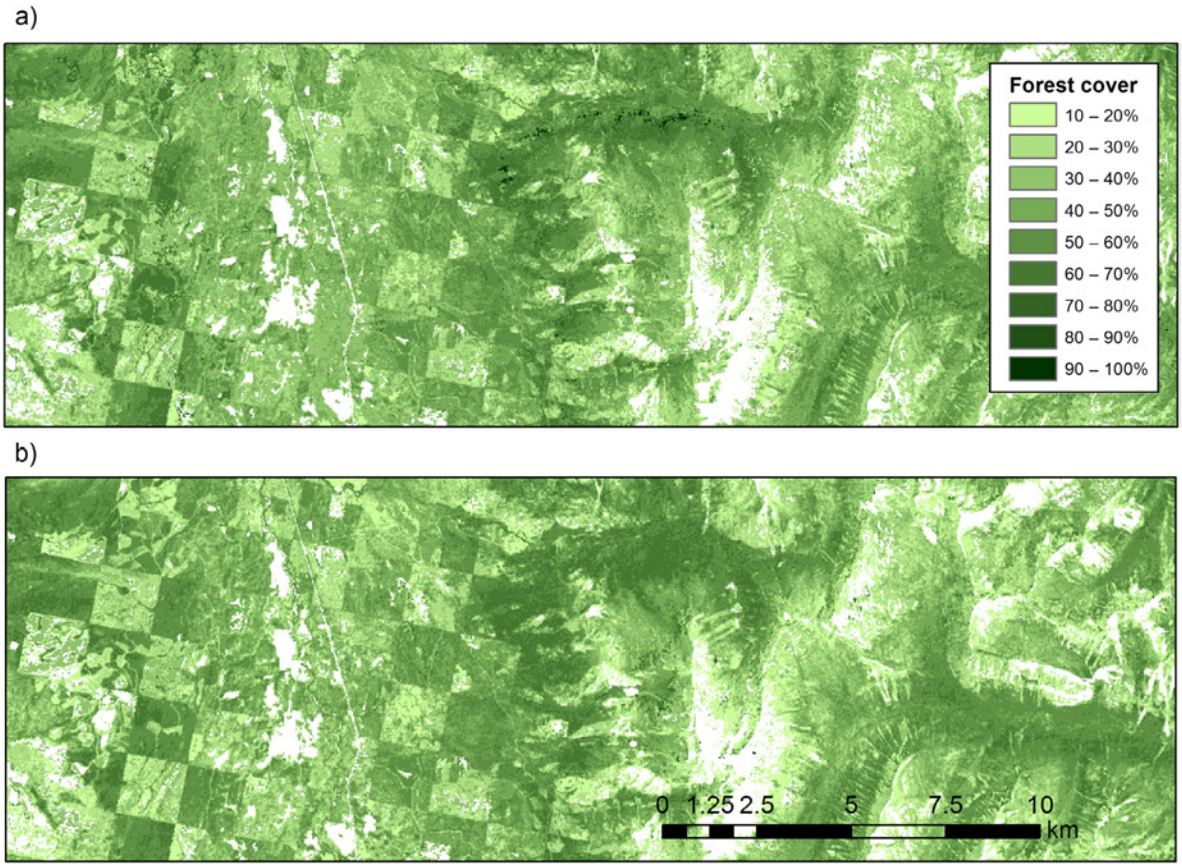

Fig. 9. Forest cover for the same subset of the landscape as in Fig. 7: (a) the imputed data and (b) the target LANDFIRE data. Here also, the checkerboard pattern of timber harvest is evident in the left half of the image, with the topographic gradients visible in the right half of the image.

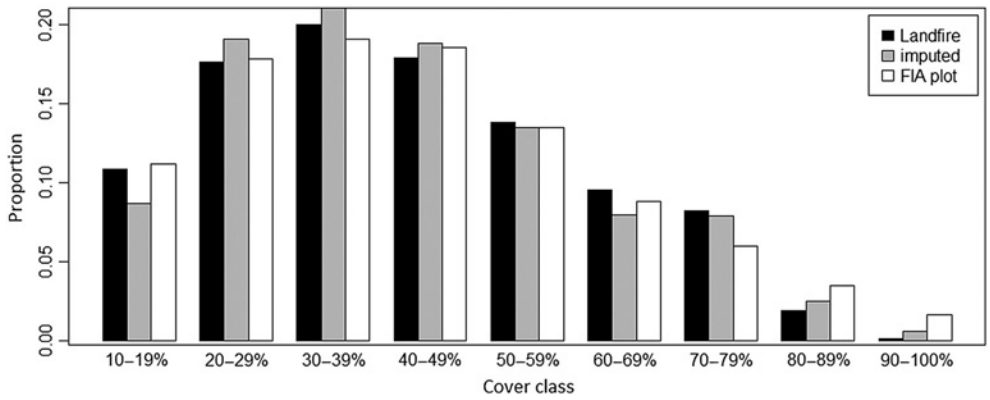

Fig. 10. The proportion of the plots or pixels in each cover class for the LANDFIRE target data, the imputed data, and the FIA plots in the western United States.

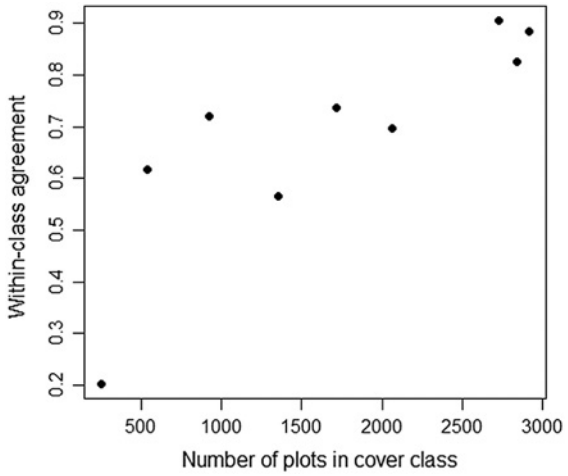

Fig. 11. Within-class agreement vs. the number of plots in each cover class.

Frequency of imputation

Plots tended to impute to a number of pixels, with counts between 10,000 and 100,000 being the most frequently occurring (Fig. 17). It was rare that a plot imputed to fewer than 10 pixels, with the number of times this occurred being around 1500.

As mentioned above, of the 15,333 plots used in this imputation, 3679 (24%) of the plots imputed to the actual pixel where their centroid was located. There are many reasons why a plot might not impute to the same pixel as its centroid location. The footprint of a single plot covers approximately 13 pixels of a 30-m grid, due to the splayed four-subplot design seen in Fig. 1, even though the combined area of a single plot is less than that of a single 30-m grid cell. We found that the characteristics of a pixel of LANDFIRE data often do not match the characteristics recorded by FIA for a plot centered on that pixel. This may result because the characteristics of the plot as a whole (including the cover, height, and vegetation type) will be driven more by the three subplots that are not located at the center pixel than by the one subplot located at the center pixel. Thus, the summary characteristics of a plot would not necessarily be expected to be a good representation of the characteristics of the center pixel, especially where landscape variability is high. In addition, LANDFIRE gridded data contain some level of error, as do the measurements taken at FIA plots. There may also be discrepancies between the plot characteristics and LANDFIRE data due to temporal mismatches between when the plot was measured and the year the LANDFIRE data were mapped. Two examples of this are as follows: (1) The forest at the pixel grew between the time the plot was measured and the year LANDFIRE data were mapped, resulting in higher cover values and a higher number of shade-tolerant species, causing the vegetation group to change, and (2) the forest burned, resulting in changes in cover, height, and vegetation group. Changes in either type could cause the plot centered on that pixel to no longer be the best match for the pixel, and a different plot to be chosen by random forests. Temporal mismatches are not important to our analysis, because we wanted to choose the plot that best represented the characteristics of each pixel circa 2008, regardless of when the plot itself was measured.

Case study in zone 22:When random forests is wrong. Or is it?

In a departure from most other LANDFIRE zones, zone 22 had low agreement (48%) between the imputed data and target data for vegetation group. Examination of the confusion matrix for vegetation group for this zone suggests that misclassification for vegetation group number 621 (Limber Pine Woodland) is driving the low accuracies for the zone (Table 5). Zone 22 comprises the Wyoming Basin, a high-elevation inland cold plateau with few trees (Fig. 3). Many, if not most, of the forested pixels in zone 22 were located in riparian areas, which LANDFIRE had generally classified as Western Riparian Woodland and Shrubland (shown in the table as vegetation group number 635). As noted above, most of the FIA plots available for imputation in this vegetation group were located in the milder mesic climate of coastal Oregon and Washington, and were not suitable matches in this environment, based on the biophysical predictor variables. Because there were no biophysically appropriate Western Riparian Woodland and Shrubland plots available for imputation, random forests tended to choose plots with a vegetation group of Limber Pine Woodland (621) for imputation in zone 22 instead (Fig. 18). We checked the FIA plots in

Table 4. Existing vegetation group names and codes that were assigned to FIA plots in the western United States in the LANDFIRE Reference Database.

Name	Code	Present in LANDFIRE gridded target data	Present in imputed data set
Unclassified Forest and Woodland	201		
Unclassified Savanna	204		
Aspen Forest, Woodland, and Parkland	602	x	x
Aspen–Mixed Conifer Forest and Woodland	603	x	x
Bigtooth Maple Woodland	605	x	x
Chaparral	607	x	x
Conifer–Oak Forest and Woodland	610	x	x
Deciduous Shrubland	612		
Desert Scrub	613		
Douglas-fir Forest and Woodland	614	x	x
Douglas-fir–Western Hemlock Forest and Woodland	615	x	x
Grassland and Steppe	618		
Juniper Woodland and Savanna	620	x	x
Limber Pine Woodland	621	x	x
Lodgepole Pine Forest and Woodland	622	x	x
Douglas-fir–Ponderosa Pine–Lodgepole Pine Forest and Woodland	625	x	x
California Mixed Evergreen Forest and Woodland	626	x	x
Mountain Hemlock Forest and Woodland	627	x	x
Mountain Mahogany Woodland and Shrubland	628	x	x
Western Oak Woodland and Savanna	629	x	x
Pinyon–Juniper Woodland	630	x	x
Ponderosa Pine Forest and Woodland and Savanna	631	x	x
Red Alder Forest and Woodland	632	x	x
Red Fir Forest and Woodland	633	x	x
Redwood Forest and Woodland	634	x	x
Western Riparian Woodland and Shrubland	635	x	x
Sitka Spruce Forest	638	x	x
Spruce-Fir Forest and Woodland	639	x	x
Subalpine Woodland and Parkland	640	x	x
Western Hemlock–Silver Fir Forest	642	x	x
Douglas-fir–Grand Fir–White Fir Forest and Woodland	643	x	x
Western Larch Forest and Woodland	644	x	x
Western Red-cedar–Western Hemlock Forest	645	x	x
Bur Oak Woodland and Savanna	659		
Juniper–Oak	696	x	x
Introduced Riparian Vegetation	701	x	

Note: Note that not all vegetation groups assigned to FIA plots are present in the forested pixels of the LANDFIRE gridded target data for the same region, or in the imputed tree-list data set.

the area and found that the most frequent vegetation group in this area was in fact limber pine rather than riparian, as suggested by the imputation. In this case, where the imputation appeared to be "wrong" based on comparison with the LANDFIRE data, it in fact accurately reflected the attributes of FIA plots in the vicinity. Random forests was able to discern that in this case, the biophysical and location predictor variables were more important than the vegetation group predictor variable, and assign the appropriate plots from a limited population.

Discussion

Where the sparseness of forest inventory data limits direct estimates of forest biomass (Blackard et al. 2008), a method such as imputation can fill in the interstitial data values. Our effort to employ a random forests imputation suggests that it holds advantages over other methods because it allows both categorical and continuous predictor variables and the fidelity of predicted multivariate response variables can be easily quantified. The technique is also repeatable when revisions

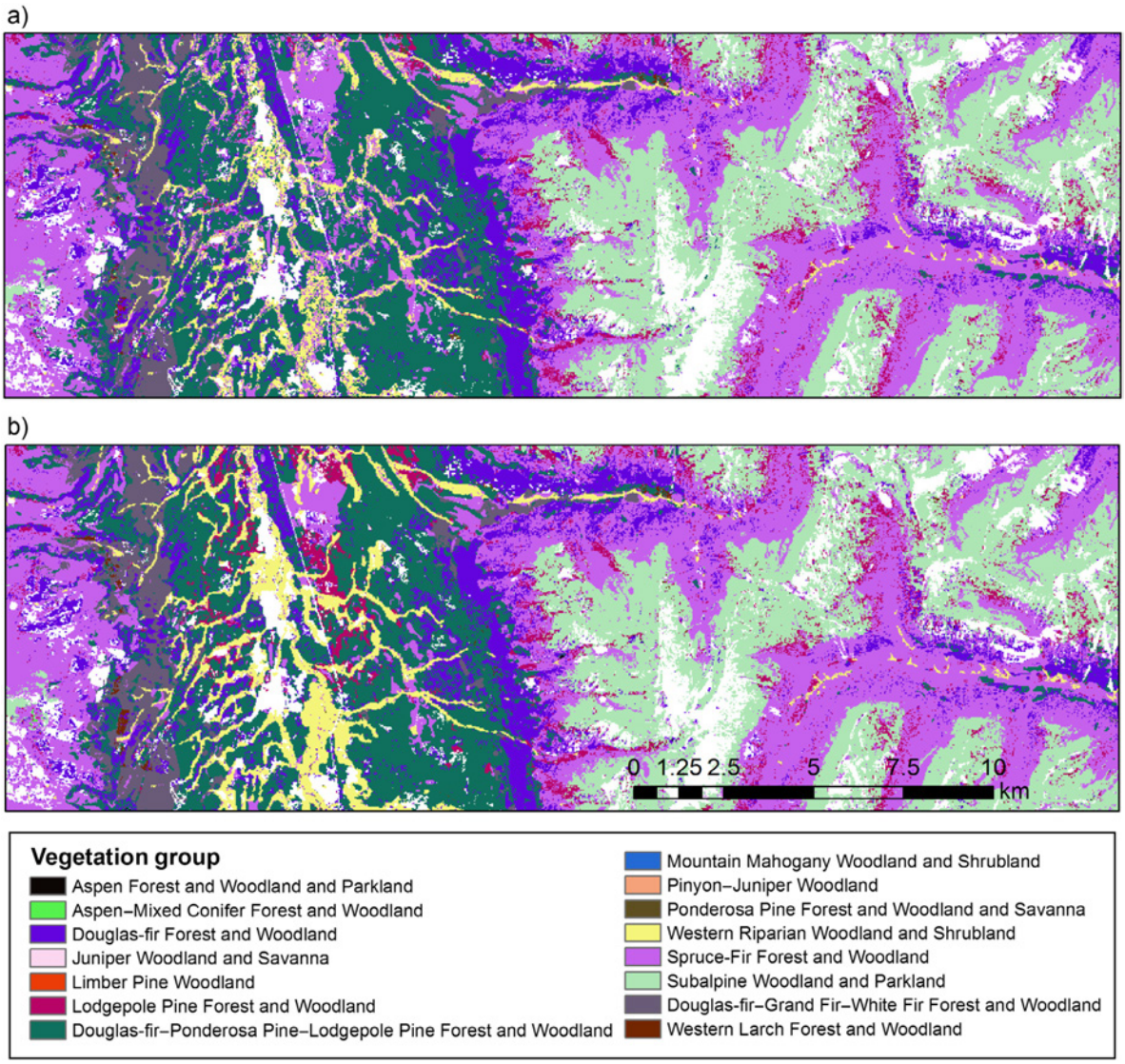

Fig. 12. Vegetation group for the same subset of the landscape shown in the figures above: (a) the imputed data and (b) the LANDFIRE data. While the agreement between the two data sources is generally high, the Western Riparian Woodland and Shrubland category is assigned less often in the imputed data than in the LANDFIRE data.

or updates of target data or reference data become available.

While not new here, the use of multivariate response variables is relatively recent (Crookston and Finley 2008) and constitutes one of the strengths of the modified random forests method used here. In order to optimize the output for predicting risk to carbon from wildfire, we chose three response variables: forest cover, forest height, and vegetation group. We allocated the number of decision trees predicting each equally (83 for forest cover, 83 for forest height, and 83 for vegetation group). However, for other applications, if one of the response variables was considered more critical than the others, it could be weighted more heavily (by allocating more of the decision trees to it).

While we found strong agreement between the gridded target LANDFIRE data and the imputed data, agreement could be improved in the future

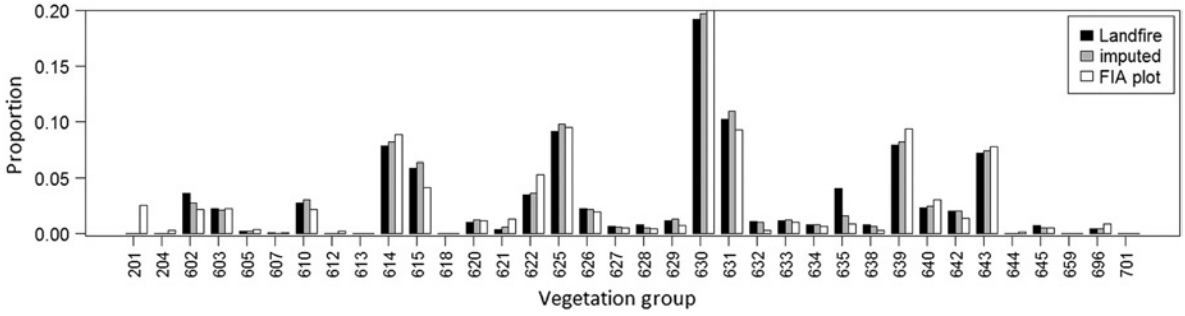

Fig. 13. The proportion of the plots or pixels in each existing vegetation group for the LANDFIRE target data, the imputed data, and the FIA plots in the western United States. The codes can be used to look up the name of the vegetation group in Table 4.

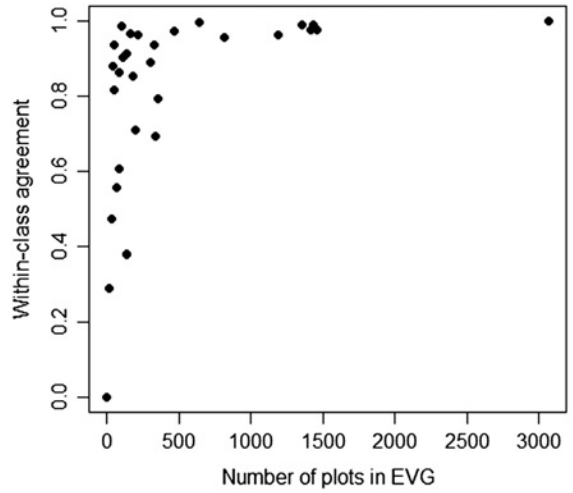

Fig. 14. The number of plots in each vegetation group and the corresponding agreement (as a proportion).

by including more forest plots, especially those in rare types. For the purposes of this project, a "type" can be considered the combination of forest cover, forest height, and vegetation group for a plot. We were restricted in the number of plots by both the FIA and LFRDB databases. However, we considered the reliability of the data in the FIA plots to be paramount and superior to other sources. The population of possible predictor variables was severely restricted because the variables must appear in both the reference (plot) and target (gridded) data sets, and of this population, we deemed that forest cover, forest height, and vegetation type were necessary to our study. It was possible to obtain these variables for the FIA plots only in the LFRDB, so we were thus restricted in the number of plots available for imputation. Nonetheless, the 15,333 plots in our data set represent a large amount of the variability present in the western United States.

The data set described here will enable future research in a number of directions. Because the imputation in essence assigns the best statistically matching FIA plot to each 30 × 30 m pixel, the resulting gridded map of plot IDs can be linked back to the FIA databases, from which many characteristics of the plot can be extracted. Thus, the map of plot IDs can be used to generate maps of basal area, or lists of the trees present at any pixel or group of pixels, among many other applications. Because each pixel is linked to a tree list, this data set provides the necessary information for initializing models such as the Forest Vegetation Simulator, including tree number, species, size, and status (Dixon 2002). Although the tree list was designed for this type of use, we have recently become aware of a number of other potential uses, including initializing wind fields in WindNinja and initializing forest stands in FireBGC.

The original purpose of this research effort was to predict the carbon risk from wildfire. To this end, the tree list can be intersected with modeled burn probability and fire intensity distributions from the western United States (Finney et al. 2011) to estimate risk to terrestrial carbon resources from wildfire. In this process, mortality and carbon storage for each plot are modeled at each fire intensity class; summaries by vegetation type or geographic area are obtained by

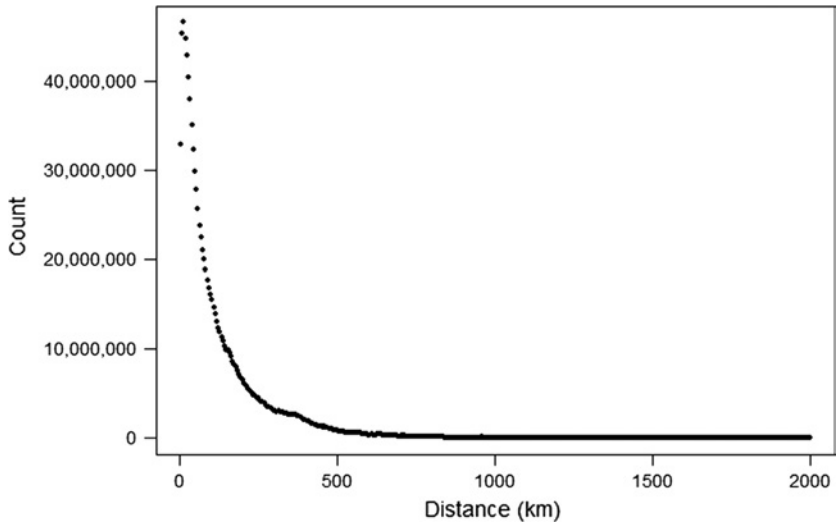

Fig. 15. Histogram of distances from imputed plot center to pixel center for the US West.

weighting the plot results by probabilities of each intensity. Stand and landscape effects of earlier wildfires and intentional fuel treatments on carbon or forest structure can be estimated by introducing changes to the landscape and re-running the fire simulations (which changes the fireline intensity probability distributions). Thus, the implications of fuel treatments on risk to carbon from wildfire can be illuminated.

The imputations are also useful for examining the potential thinning volume associated with fuel treatments and alternative fire risk reduction activities. Thinning is a critical part of restoration treatments prior to the introduction of prescribed burning in many low- to mid-elevation forests of ponderosa pine and mixed conifer in the western United States (Graham et al. 2004, Agee and Skinner 2005, Martinson and Omi 2013). Because this data set can be linked back to the number, size, and species of trees modeled for each pixel, the data set can be used to apply treatment prescriptions and estimate the characteristics of trees that would thus be removed. Estimating the potential volume of merchantable and nonmerchantable forest products available from treatment activities is helpful for treatment planning for national forests, as well as economic cost-benefit analyses.

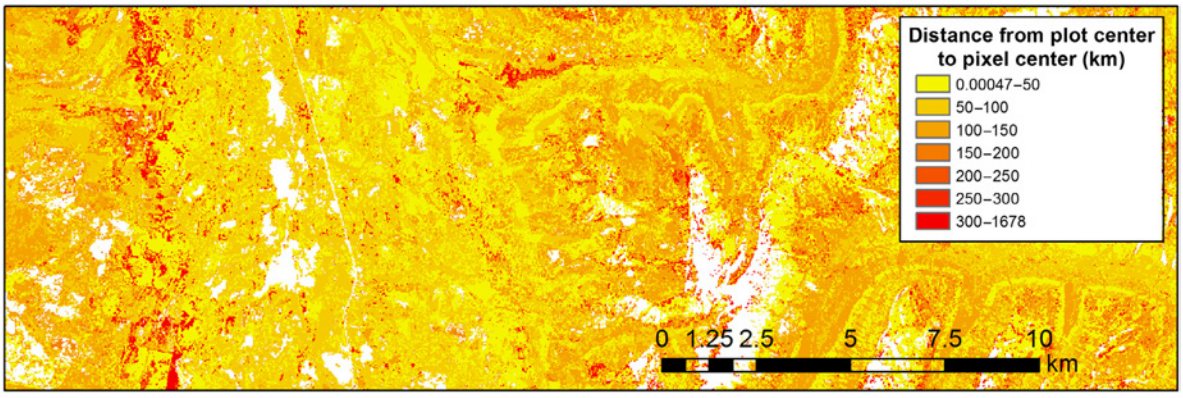

Fig. 16. Distance from the pixel center to the center of the plot that imputed to that pixel, for the same subset of the landscape as in previous figures.

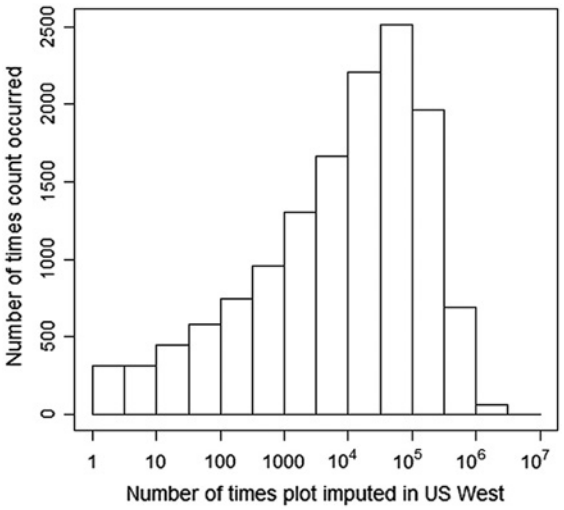

Fig. 17. A histogram of the number of pixels to which an individual plot imputed.

Conclusions

The modified random forest approach presented here produced high within-class agreement with gridded landscape data for the western United States for our three response variables. Within-class agreement for forest cover was 79%, agreement for forest height was 92%, and agreement for vegetation group was 96%. The methodology presented here is novel in that the three response variables served also as predictor variables for the reference data, with the expected result of producing high rates of agreement between the gridded target landscape data and the imputed data for these three variables. We chose this methodology in order to optimize the imputed data for the prediction of risk to carbon resources from wildland fire and for biomass calculations. In that sense, this tree list might not be the optimal data set for all applications (e.g., mapping of tree species envelopes). However, the methodology presented here is flexible and allows users to select predictor and response variables best suited to their research goals.

Because the spatial data set presented here can be linked to the extensive list of attributes recorded by FIA for each plot, a number of forest characteristics can be estimated directly from the data set. In addition, the data set can be used to initialize a number of forest simulation models. Because the data set in essence provides a tree-level model of the forests in the western United States, it greatly augments the information available to researchers and managers who previously relied on data from sparse forest plots or stand inventories.

Table 5. Confusion matrix for vegetation group in zone 22.

	Imputed plot vegetation group													
	602	603	614	620	621	622	625	628	630	631	635	639	640	Accuracy
LANDFIRE (target) vegetation group														
602	*283,552*	1976	16,957	1080	214,371	3699	28,111	109	2149	38,398	22,699	1033	117	0.46
603	173	*10,911*	4425	0	1592	9083	4744	1	0	2678	0	4523	2	0.29
614	1989	965	*26,626*	3	7640	213	1163	715	729	1128	1419	145	0	0.62
620	13	4	239	*2025*	70,455	37	222	14	7570	982	910	20	8	0.02
621	1307	566	5746	628	*1,228,828*	2708	3350	154	1902	18,686	11,871	1963	2363	0.96
622	6	0	358	0	12	*52,342*	2	0	0	0	0	67	16	0.99
625	560	155	10,448	0	1246	1636	*18,505*	75	61	122	115	3654	95	0.50
628	637	131	12,741	340	234,846	14,103	9454	*116,512*	29,603	47,933	24,971	33,397	1414	0.22
630	42	3	176	92	4061	59	65	13	*643,964*	1107	1217	247	87	0.99
631	37	15	27	0	36,888	1211	31	397	1296	*78,507*	69	46	17	0.66
635	11,069	5062	42,199	3809	1,700,003	49,878	47,304	24,667	106,053	31,430	*255,595*	6292	700	0.11
639	57	18	961	0	2305	2018	726	21	111	585	227	*31,434*	0	0.82
640	170	132	526	0	1430	3970	1927	27	108	820	0	122	*3195*	0.26
Accuracy	0.95	0.55	0.22	0.25	0.35	0.37	0.16	0.82	0.81	0.35	0.80	0.38	0.40	0.48

Notes: Overall, agreement for this zone was the lowest for any zone in the western United States at 48%. The most plentiful vegetation group in this zone was 621, Limber Pine Woodland. The low agreement in that vegetation group is one of the major drivers of the low agreement for the zone as a whole, due to the large proportion of pixels mapped as 621. Italics denote the instances where the model predicted a given vegetation group correctly. (Note that the codes can be looked up in Table 4 to yield the name of the vegetation group.)

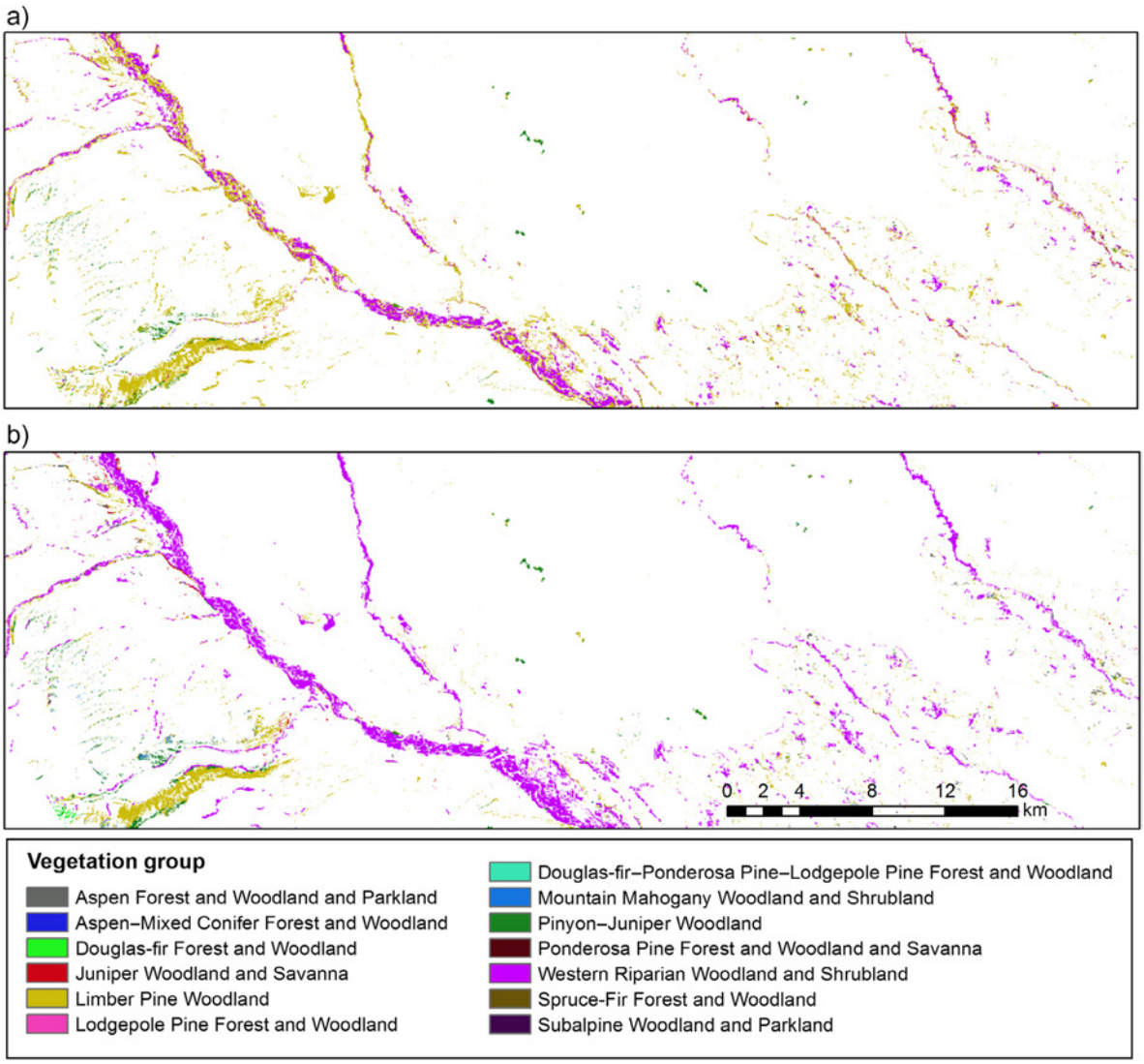

Fig. 18. A subset of the landscape in zone 22, with forested pixels appearing along riparian corridors: (a) the random forests imputation and (b) the LANDFIRE data. Random forests favored the imputation of Limber Pine Woodland plots for many pixels where LANDFIRE had classified the pixel as Western Riparian Woodland and Shrubland.

Acknowledgments

The Rocky Mountain Research Station and the National Fire Decision Support Center supported this effort. We are grateful to Nicholas Crookston for assistance in understanding and using his yaImpute package in R. Elizabeth Burrill at FIA and Chris Toney with the USFS Fire Laboratory in Missoula assisted us with obtaining FIA and LFRDB data via a Memorandum of Cooperation. We would also like to thank two anonymous reviewers for their comments on a previous draft.

Literature Cited

Agee, J. K., and C. N. Skinner. 2005. Basic principles of forest fuel reduction treatments. Forest Ecology and Management 211:83–96.

Blackard, J. A., et al. 2008. Mapping U.S. forest biomass using nationwide forest inventory data and moderate resolution information. Remote Sensing of Environment 112:1658–1677.

Breiman, L. 2001. Random forests. Machine Learning 45:5–32.

Brown, S. L., and P. E. Schroeder. 1999. Spatial patterns of aboveground production and mortality of woody biomass for eastern U.S. forests. Ecological Applications 9:968–980.

Burkman, B. 2005. Forest Inventory and Analysis Sampling and Plot Design. FIA Fact Sheet Series. http://www.fia.fs.fed.us/library/fact-sheets/data-collections/Sampling%20and-20Plot%20Design.pdf

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educational and Psychosocial Measurement 20:37–46.

Cressie, N. A. C. 1990. The origins of kriging. Mathematical Geology 22:239–252.

Crookston, N. L., and A. O. Finley. 2008. yaImpute: an R package for kNN imputation. Journal of Statistical Software 23:1–15.

Cutler, D. R., T. C. Edwards, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. 2007. Random forests for classification in ecology. Ecology 88:2783–2792.

Dixon, G. E., comp. 2002. Essential FVS: a user's guide to the Forest Vegetation Simulator. USDA Forest Service, Forest Management Service Center, Fort Collins, Colorado, USA.

Drury, S., and J. Herynk. 2011. The national tree-list layer: a seamless spatially-explicit tree-list layer for the continental United States. USDA Forest Service Rocky Mountain Research Station General Technical Report RMRS-GTR-254.

Falkowski, M. J., A. T. Hudak, N. L. Crookston, P. E. Gessler, E. H. Uebler, and A. M. S. Smith. 2010. Landscape-scale parameterization of a tree-level forest growth model: a k-nearest neighbor imputation approach incorporating LiDAR data. Canadian Journal of Forest Research 40:184–199.

Finney, M. A., C. W. McHugh, I. C. Grenfell, K. L. Riley, and K. C. Short. 2011. A simulation of probabilistic wildfire risk components for the continental United States. Stochastic Environmental Research and Risk Assessment 25:973–1000.

Finney, M. A., R. C. Seli, C. W. McHugh, A. A. Ager, B. Bahro, and J. K. Agee. 2007. Simulation of long-term landscape-level fuel treatment effects on large wildfires. International Journal of Wildland Fire 16:712–727.

Franco-Lopez, H., A. R. Ek, and M. E. Bauer. 2001. Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. Remote Sensing of Environment 77:251–274.

Graham R. T., S. McCaffrey, and T. B. Jain, technical editors. 2004. Science basis for changing forest structure to modify wildfire behavior and severity. USDA Forest Service General Technical Report RMRS-GTR-120.

Haas, J. R., D. E. Calkin, and M. P. Thompson. 2014. Wildfire risk transmission in the Colorado Front Range, USA. Risk Analysis 35:226–240.

Homer, C., C. Huang, L. Yang, B. Wylie, and M. Coan. 2004. Development of a 2001 National Land-Cover Database for the United States. Photogrammetric Engineering and Remote Sensing 7:829–840.

Hudak, A. T., N. L. Crookston, J. S. Evans, D. E. Hall, and M. J. Falkowski. 2008. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. Remote Sensing of Environment 112:2232–2245.

Jenkins, J. C., R. A. Birdsey, and Y. Pan. 2001. Biomass and NPP estimation for the mid-Atlantic region (USA) using plot-level forest inventory data. Ecological Applications 11:1174–1193.

Keane, R. E., R. B. Burgan, and J. van Wagtendonk. 2001. Mapping wildland fuels for fire management across multiple scales: integrating remote sensing, GIS, and biophysical modeling. International Journal of Wildland Fire 10:301–319.

Krige, D. G. 1951. A statistical approach to some mine valuations and allied problems at the Witwatersrand. Master's thesis. University of Witwatersrand, Johannesburg, South Africa.

Latifi, H., F. Fassnacht, and B. Koch. 2012. Forest structure modeling with combined airborne hyperspectral and LiDAR data. Remote Sensing of Environment 121:10–25.

Martinson, E. J., and P. N. Omi. 2013. Fuel treatments and fire severity: a meta-analysis. Research Paper RMRS-RP-103WWW. USDA Forest Service, Rocky Mountain Research Station, Fort Collins, Colorado, USA.

McRoberts, R. 2009. A two-step nearest neighbors algorithm using satellite imagery for predicting forest structure within species composition classes. Remote Sensing of Environment 11:532–545.

Moeur, M., and A. R. Stage. 1995. Most similar neighbor: an improved sampling inference procedure for natural resource planning. Forest Science 41:337–359.

Muinonen, E., M. Maltamo, H. Hyppanen, and V. Vainikainen. 2001. Forest stand characteristics estimation using a most similar neighbor approach and image spatial information. Remote Sensing of Environment 78:223–228.

NatureServe. 2009. International ecological classification standard: terrestrial ecological classifications. NatureServe Central Databases, Arlington, Virginia, USA.

O'Connell, B. M., E. B. LaPoint, J. A. Turner, T. Ridley, S. A. Pugh, A. M. Wilson, K. L. Waddell, and B. L. Conkling. 2014. The Forest Inventory and Analysis Database: Database Des-

cription and User Guide Version 6.0 for Phase 2. http://www.fia.fs.fed.us/library/databasedocumentation/current/ver6.0/FIADB_user%20guide_6-0_p2_5-6-2014.pdf

Ohmann, J. L., and M. J. Gregory. 2002. Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, U.S.A. Canadian Journal of Forest Research 32:725–741.

Pierce, K. B., J. L. Ohmann, M. C. Wimberly, M. J. Gregory, and J. S. Fried. 2009. Mapping wildland fuels and forest structure for land management: a comparison of nearest neighbor imputation and other methods. Canadian Journal of Forest Research 39:1901–1916.

R Foundation. 2016. The R Project for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org

Rollins, M. G. 2009. LANDFIRE: a nationally consistent vegetation, wildland fire, and fuel assessment. International Journal of Wildland Fire 18:235–249.

Schmidt, K., J. P. Menakis, C. C. Hardy, W. J. Hann, and D. L. Bunnell. 2002. Development of coarse-scale spatial data for wildland fire and fuel management. U.S. Forest Service, Rocky Mountain Research Station GTR-RMRS-87.

Van Wagtendonk, J. W., and R. R. Root. 2003. The use of multi-temporal Landsat Normalized Difference Vegetation Index (NDVI) data for mapping fuel models in Yosemite National Park, USA. International Journal of Remote Sensing 24:1639–1651.

Wilson, B. T., A. J. Lister, and R. I. Riemann. 2012. A nearest-neighbor imputation approach to mapping tree species over large areas using forest inventory plots and moderate resolution raster data. Forest Ecology and Management 271:182–198.