Utilizing random forests imputation of forest plot data for landscape-level wildfire analyses

Karin L. Riley[a], Isaac C. Grenfell[b], Mark A. Finney[c], and Nicholas L. Crookston[c]

[a] *College of Forestry and Conservation, University of Montana, 32 Campus Drive, Missoula, Montana, USA, 59812, karin.riley@umontana.edu*
[b] *RTL Networks, US Highway 10 W, Missoula, Montana, USA, 59808, igrenfell@gmail.com*
[c] *Missoula Fire Sciences Laboratory, Rocky Mountain Research Station, U.S. Forest Service,5775 US Highway 10 W, Missoula, Montana, USA, 59808, mfinney@fs.fed.us* [d]*Moscow Forestry Sciences Laboratory, Rocky Mountain Research Station, U.S. Forest Service, 1221 South Main Street, Moscow, Idaho, USA, 83843, nrcrookston@fs.fed.us*

Abstract
Maps of the number, size, and species of trees in forests across the United States are desirable for a number of applications. For landscape-level fire and forest simulations that use the Forest Vegetation Simulator (FVS), a spatial tree-level dataset, or "tree list", is a necessity. FVS is widely used at the stand level for simulating fire effects on tree mortality, carbon, and biomass, but uses at the landscape level are limited by lack of availability of forest inventory data for large contiguous areas. Detailed mapping of trees across large areas is not feasible with current technologies, but statistical methods for matching 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 Landfire vegetation and biophysical predictors at 30m grid resolution. In essence, the random forests method creates a "forest" of decision trees in order to choose the forest plot with the best statistical match for each grid cell in the landscape. Landfire data was used in this project because is publicly available, offers seamless coverage of variables needed for fire models, and is consistent with other datasets, including burn probabilities and flame length probabilities generated for the continental US by Fire Program Analysis (FPA). We used the imputed forest plot data to generate a map of forest cover and height as well as existing vegetation group for a study area in eastern Oregon, and examined correlations with Landfire data. The results showed good correspondence between the two data sets (84-97% within-class agreement, depending on the variable). In future research, the new imputed grid of inventory data will be used for landscape simulation studies to determine risk to terrestrial carbon resources from wildfire as well as to investigate the effect of fuel treatments on burn probability and fire sizes.

Keywords*: imputation; forest plot data; Landfire; random forests*

1. Introduction

For many research applications ranging from estimation of terrestrial carbon resources to the impact of fuel treatment projects on wildfire propagation, to name a few, it is desirable to know the location, size, and species of each tree on the landscape. However, such a mapping effort is not feasible with current technologies. LiDAR and similar technologies may make such a tree-level map a reality in coming years, but in the interim, various statistical efforts can produce spatial models, known as "tree lists", suitable for a wide range of research applications.
To this end, a number of statistical methods have been evaluated in the literature, ranging from gradient nearest neighbour imputation (GNN), linear models (LM), classification and regression trees (CART), kriging, universal kriging (UK) and most similar neighbour (MSN) (Moeur and Stage 1995; Pierce *et al.* 2009). Among these, Pierce *et al.* (2009) found GNN performed best for forest structure variables in Oregon, while LMs and UK demonstrated stronger performance for both forest structure and canopy variables in Washington and California. Drury and Herynk (2011) produced a national tree list by stratifying plot data by existing vegetation type, biophysical setting, succession class, and canopy bulk

density. Because this method generally left several potential matches for each grid cell on the landscape, and because the purpose of the tree list was to model tree mortality from fire, they then identified the median bark thickness for each plot, and chose the plot with the median of the median bark thickness to assign.

This project had several specific requirements that precluded use of most of these methods. For this project, we required our tree list to be consistent with two already existing datasets. The Landfire project provides over 20 national geo-spatial layers, including topographic, fuel, and vegetation layers, on 30m grids (www.landfire.gov). These layers, in turn, are used as inputs to Fire Program Analysis (FPA), which runs national-level wildfire simulations to output burn probability and flame length probabilities on a 270m grid, accompanied by a set of modelled fire perimeters. Therefore, we leveraged the Landfire dataset as inputs to our tree list, to ensure consistency with that dataset as well as the FPA outputs. In order to model forest type, introduction of a categorical variable, existing vegetation group, into our model along with numeric variables was necessary. That limited the set of possible methodologies to classification trees. We chose random forests as our methodology, since 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 (Cutler *et al.* 2007). The random forests method as used here in essence evaluates a set of forest plots, and identifies the best-matching plot for each grid cell on the landscape. Our methodology is distinct from that of Pierce *et al.* (2009) in that it has the capacity to use categorical variables and is consistent with Landfire and FPA data. Several important differences exist between our methodology and that of Drury and Herynk (2011) as well: 1) we limited our data to a single set of nationally consistent plot data, whereas they obtained a variety of fixed- and variable-plot designs from multiple agencies, 2) since 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, retaining more variability on the landscape.

Here, we demonstrate high model accuracies for a random forests imputation run on an approximately 40,000 km^2 area of forest in the western U.S., indicating the output would be suitable for a wide range of research applications.

2. **Methods**

In order to test and refine our methodology, we chose a Landfire zone, Zone 9, which lies mainly in eastern Oregon, as our study area (Figure 1). The zone contains large forested areas, amounting to about 25% of the total area of the zone, or approximately 40,000 km^2.

In this random forests imputation, a set of reference observations was imputed to a set of target points (Crookston and Finley 2008). The reference observations consisted of a set of forest plot data acquired from the US Forest Service's Forest Inventory Analysis (FIA) program. Beginning in 1999, FIA has been using a standardized plot design to conduct forest surveys across the US (O'Connell *et al.* 2014). Among the variables collected at these plots that we utilized in the imputation are: elevation, slope, aspect, latitude, and longitude. The Landfire program calculates additional variables needed for the imputation and stores them in their Landfire Reference Database (LFRDB), including forest height, forest cover, and existing vegetation group (EVG) computed from a classification method devised by NatureServe (2009) for Landfire. We derived additional biophysical variables via overlays of plot locations with gridded data from the Landfire project for photosynthetically active radiation, precipitation, relative humidity, maximum temperature, minimum temperature, and vapour pressure deficit. The target points in this study consist of a grid of 30m pixels that comprise the Landfire dataset. So, in essence, we use random forests to find the best matching FIA plot for each 30m pixel, imputing an FIA plot number to each pixel.

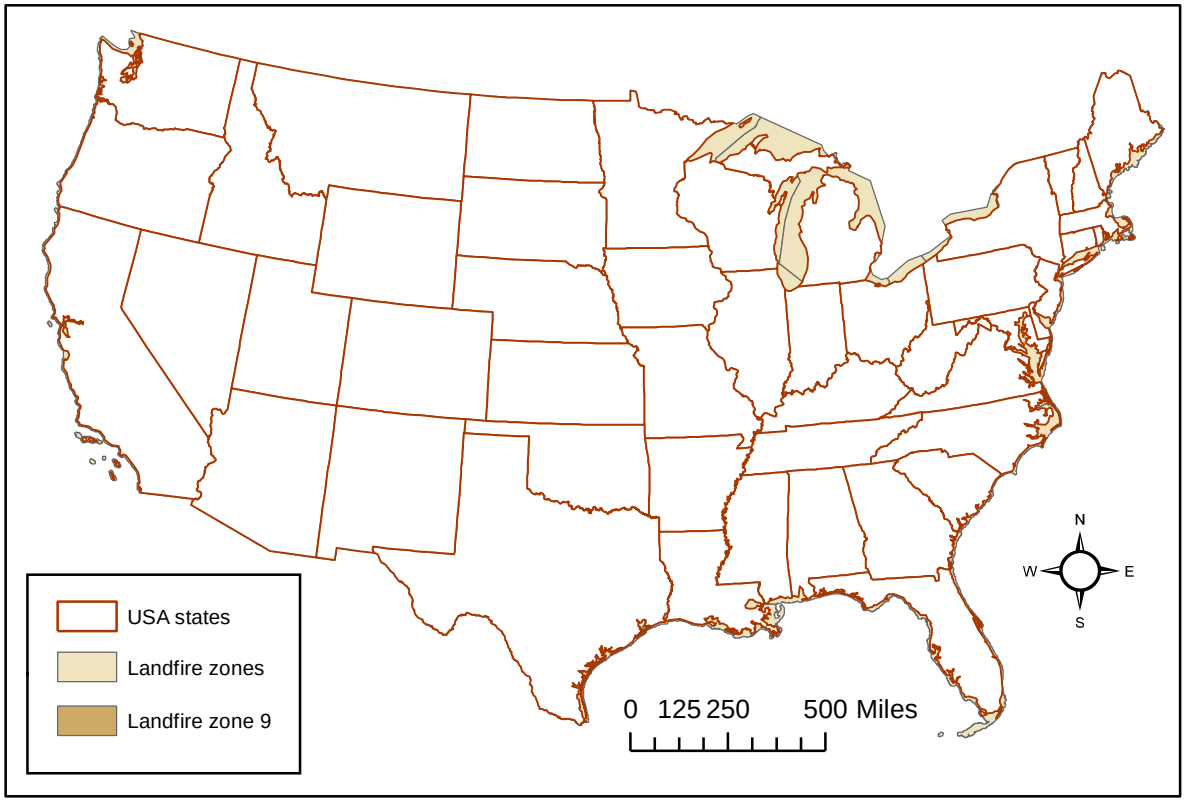

Figure 1. Location of the study area, Landfire Zone 9, in the western United States.

For this project, we retained only FIA plots that utilized the national design, were single condition (meaning they did not cross ownership boundaries or major vegetation types), and appeared in the LFRDB. Thus, we began with 15,333 plots in the western half of the continental United States, and then created a subset consisting only of the plots with EVGs appearing in Zone 9. Then, we formed the random forest model using the yaImpute package in the statistical program R. To do this, we used 250 total decision trees to predict tree height, tree cover, and EVG for each plot. To predict these, we used the following predictors: latitude, longitude, tree cover, tree height, elevation, slope, EVG, PAR (photosynthetically active radiation), precipitation, relative humidity, maximum temperature, minimum temperature, VPD (vapour pressure deficit), cosine of aspect (northing), and sine of aspect (easting). Notice that the variables we wish to predict also appear as predictors. This may appear odd, but the reason we do this is that the objective with random forest imputation is to build a model that assigns a set of predictor values to a plot associated with the response variables. Accuracy in the response variables can be heightened by also including them as predictors. The random forest method involves building many classification trees (in this case, 83 for each response variable, adding to the total 249). Each tree is formed using a subset of the plots, and the remainder (referred to as the out-of-bag observations) are set aside to assess accuracy. As Cutler *et al.* (2007) eloquently describes it, "Observations in the original data set that do not occur in a bootstrap sample are called out-of-bag observations. A classification tree is fit to each bootstrap sample, but at each node, only a small number of randomly selected variables (e.g., the square root of the number of variables) are available for binary partitioning." 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. "The predicted class of an observation is calculated by the majority vote of the out-of-bag-predictions for that observation, with ties split randomly" (Cutler *et al.* 2007).

We can obtain 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 forest model. In this case,

the out-of-bag error rates for Zone 9 were 6.99%, 1.79%, and 0.897% for forest cover, forest height, and EVG respectively, an indication of high model accuracy.
Once the forest of decision trees is in hand, we can impute new target observations to determine which reference plots are most closely associated with the targets (pixels, in this case). Our dataset consisted of 44,138,635 forested pixels. The imputation is done by evaluating the target predictor variables for each pixel through each of the 3 sets of trees associated with each response variable. Then, the plot most frequently imputed amongst all 500 trees is considered the winner. Once we obtain this list of reference plots, we can build imagery of the variables of interest associated with each imputed plot. In this case, we output raster images of forest cover, forest height, existing vegetation group, and plot number. The plot number can be used to reference the number, size, and species of trees in each plot via a lookup table.
Validation consisted of assessing within-class accuracy for the three response variables (forest cover, forest height, and existing vegetation group) using confusion matrices and the kappa statistic. Barplots were used to assess the proportion of pixels in each class in the target data versus the imputed data.

3. Results and Discussion

Accuracies for imputed forest height were high. Landfire maps forest height in four classes: 0-5 m, 5-10 m, 10-25 m, and greater than 25 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 (Table 1). The resulting confusion matrix represents a type of accuracy assessment of the outputs. We do not assess the accuracy of the Landfire data itself, which has its own error rates, but we compute the accuracy of our imputation compared to the gridded target data. Within-class accuracy for forest height was 97% in Zone 9. The number of pixels in each height class compared quite favourably across the imputed plot data and the Landfire gridded target data (Figure 2).

Table 1. Confusion matrix of forest height in meters in gridded Landfire data and imputed forest plot data.

		Imputed plot				
		0-5m	**5-10m**	**10-25m**	**>25m**	**Accuracy**
Gridded Landfire	**0-5m**	1,059,777	98,042	820	13	**0.91**
	5-10m	47,313	7,842,084	6,227	63	**0.99**
	10-25m	852	66,957	23,609,236	107,949	**0.99**
	>25m	316	2,734	1,100,876	10,195,377	**0.90**
	Accuracy	**0.96**	**0.98**	**0.96**	**0.99**	**0.97**

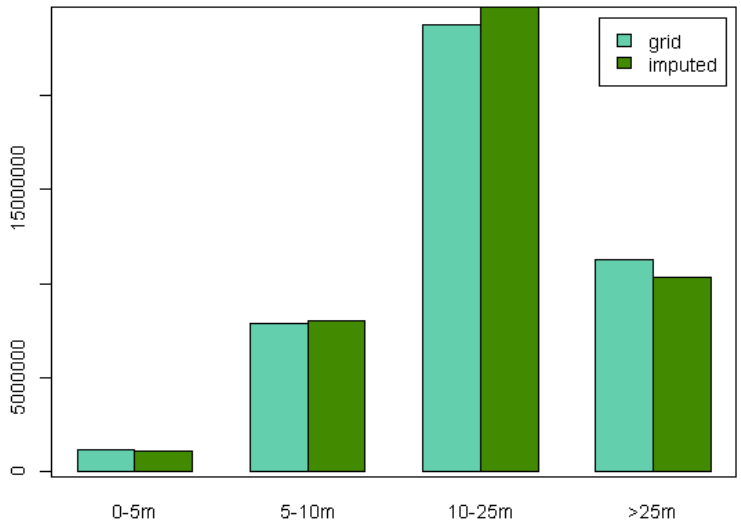

Figure 2. Barplots comparing height class of imputed forest plot data and gridded Landfire reference data. Units of the y axis are numbers of pixels.

The accuracy of the imputation was also high for forest cover. Forest cover is mapped in nine classes in Landfire (Table 2), with areas of tree cover less than 10% not considered forested. For FIA plots, forest cover is estimated to the nearest percent in the LFRDB. Overall accuracy for the zone was 86%, with within-class accuracy ranging from less than 1% in the two densest forest cover classes (80-89% and 90-100%) to 98% in a moderate cover class (30-39%). The proportion of cover classes compared favourably across the imputed forest plots and gridded target Landfire data (Figure 3). The largest discrepancies were in the sparsest cover class (10-19%), which was underestimated by the imputed plot data, and the 20-29% cover class, which was conversely overestimated by the imputed plot data. Forest cover greater than 60% was rare in the Landfire reference data.

Table 2. Confusion matrix of forest cover in percent in gridded Landfire data and imputed forest plot data.

Imputed plot

Gridded Landfire	10-19%	20-29%	30-39%	40-49%	50-59%	60-69%	70-79%	80-89%	90-100%	Accuracy
10-19%	9,052,029	3,321,677	345,886	42,645	26,675	1,066	3,105	3	0	**0.71**
20-29%	171,230	7,960,416	62,318	4,945	10,694	42	45	2	0	**0.97**
30-39%	76,663	79,301	7,984,875	34,055	6,778	3	734	0	0	**0.98**
40-49%	13,698	38,271	289,665	9,001,039	109,372	25,566	1,255	0	0	**0.95**
50-59%	23,696	22,807	151,700	464,473	3,859,002	362,658	64,946	1,741	2,154	**0.78**
60-69%	123	1,071	12,445	39,498	64,997	306,547	60,585	9,197	2,100	**0.62**
70-79%	32	34	1,087	1,472	2,456	4,044	5,367	1,860	83	**0.33**
80-89%	2	35	2,066	2,182	308	3,220	154	38	396	**0.00**
90-100%	0	1	4	0	2	0	0	0	0	**0.00**
Accuracy	**0.97**	**0.70**	**0.90**	**0.94**	**0.95**	**0.44**	**0.04**	**0.00**	**0.00**	**0.86**

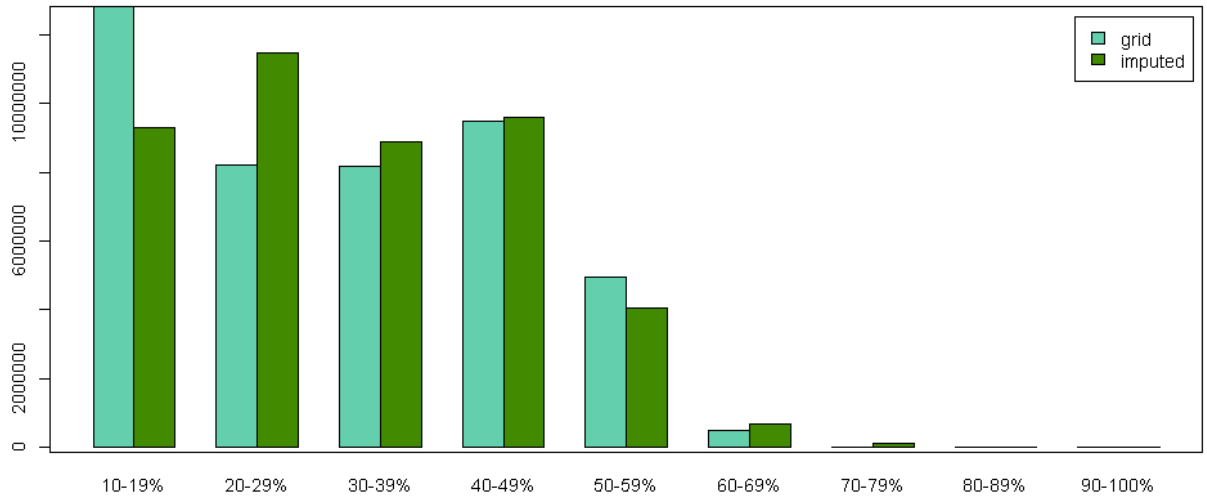

Figure 3. Barplots comparing forest cover class of imputed forest plot data and gridded Landfire target data. Units of the y axis are numbers of pixels.

The third response variable in our study was existing vegetation group (EVG). There were 14 EVGs in Zone 9 (Table 3). EVG is mapped to the gridded target data by Landfire, and also assigned to FIA forest plots in the LFRDB. Within-class accuracy of EVG was 84% for the study area as a whole. Class proportions compared favourably across the gridded Landfire data and the imputed plot data, but the most common EVG classes tended to be somewhat over-represented in the imputed plot data (Figure 4).

Table 3. Landfire Existing Vegetation Groups (EVGs) in zone 9, by numeric code and text description

EVG Code	EVG Description
602	Aspen Forest, Woodland, and Parkland
603	Aspen-Mixed Conifer Forest and Woodland
614	Douglas-fir Forest and Woodland
620	Juniper Woodland and Savanna
621	Limber Pine Woodland
622	Lodgepole Pine Forest and Woodland
625	Douglas-Fir-Ponderosa Pine-Lodgepole Pine Forest and Woodland
628	Mountain Mahogany Woodland and Shrubland
630	Pinyon-Juniper Woodland
631	Ponderosa Pine Forest and Woodland and Savanna
635	Western Riparian Woodland and Shrubland
639	Spruce-Fir Forest and Woodland
640	Subalpine Woodland and Parkland
643	Douglas-fir-Grand Fir-White Fir Forest and Woodland

Table 4. Confusion matrix of Existing Vegetation Group (EVG) in gridded Landfire data and imputed forest plot data.

Imputed plot

Gridded Landfire

	602	603	614	620	621	622	625	628	630	631	635	639	640	643	Accuracy
602	314,840	249,598	95,978	1,385,619	7,993	59,758	334,829	107,570	64,309	535,962	1,714	19,841	7,251	586,985	**0.08**
603	4,550	154,112	23,785	25,385	609	13,355	213,861	743	135	50,873	222	8,258	823	96,633	**0.26**
614	38	349	2,571,261	16,626	1,162	860	47,797	42	3,563	10,060	294	86	894	2,609	**0.97**
620	730	118	16,363	5,807,709	751	5,691	27,071	15,083	89,344	60,399	2,000	816	21,819	3,970	**0.96**
621	0	0	2	0	2,439	186	0	0	4	52	0	266	1,076	461	**0.54**
622	3	3,139	919	2,883	659	402,396	10,222	7	1	70,926	249	31,867	826	9,664	**0.75**
625	540	359	3,831	6,904	375	171,844	15,500,221	107	253	36,256	1,593	25,714	141	5,903	**0.98**
628	602	9,372	19,380	62,687	901	3,432	70,707	304,824	9,576	112,518	187	7,306	3,469	41,835	**0.47**
630	0	159	100	13,145	0	376	1,120	700	205,741	7,978	0	2,451	102	21,165	**0.81**
631	0	2,605	2,590	89,693	702	6,818	119,542	163	69,293	7,527,132	51	7,345	7,183	225	**0.96**
635	209	62,056	59,712	493,731	39	102,195	141,945	4,645	2,221	383,109	646,482	21,820	2,907	70,895	**0.32**
639	101	2,506	1,108	1,892	11	1,911	2,818	120	11	7,804	395	1,023,592	0	338	**0.98**
640	1	0	31	0	0	843	28	16	0	461	0	298	347,194	10	**1.00**
643	1,126	4,455	139,042	1,802	3	13,250	214,882	161	8	7,863	658	57,759	3,668	2,211,945	**0.83**
Accuracy	**0.98**	**0.32**	**0.88**	**0.73**	**0.16**	**0.51**	**0.93**	**0.70**	**0.46**	**0.85**	**0.99**	**0.85**	**0.87**	**0.72**	**0.84**

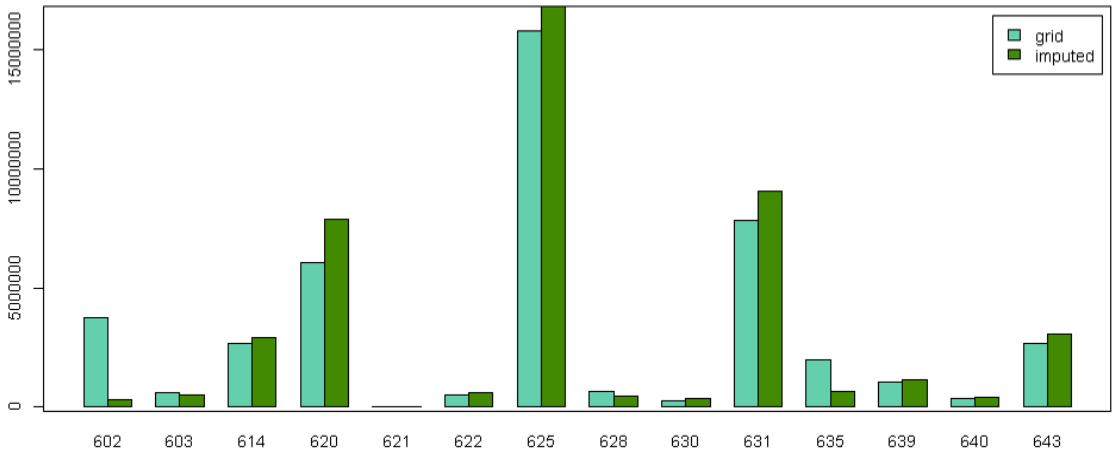

Figure 4. Barplots comparing Existing Vegetation Group (EVG) of imputed forest plot data and gridded Landfire reference data. Units of the y axis are numbers of pixels.

In general, within-class accuracies were lower in rarer classes. This result makes sense, since it is unlikely in rare types that random forests can match all three of the response variables (forest cover, height, and existing 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. If increased accuracy was desired in future implementations of this methodology, increasing the sample size of rare types would likely be the most effective way to boost accuracies.

4. **Conclusions**

Here, we have demonstrated that a modified random forests approach is a feasible method for imputing forest plots to a set of target landscape grids. This method produces a seamless grid of tree data at the landscape level. The modified random forests method produced high correlations between the target gridded data and the imputed plot data for the response variables of forest cover, forest height, and existing vegetation group (86%, 97%, and 84% respectively), an indication of high model accuracy. Very high classification accuracy is one of the strengths of the random forest method, along with its ability to utilize categorical as well as numerical variables (Cutler *et al.* 2007). Due to the high accuracy, the output imputed forest plot data should perform well in a number of applications, including estimation of risk from wildfire to terrestrial carbon resources, and analysis of the effect of fuel treatments on fire sizes and landscape-level burn probability.

5. References

Crookston, NL, Finley, AO (2008) yaImpute: an R package for kNN imputation. Journal of Statistical Software 23, 1-15.

Cutler, DR, Edwards, TC, Beard, KH, Cutler, A, Hess, KT, Gibson, J, Lawler, JJ (2007) Random forests for classification in ecology. Ecology 88, 2783-2792.

Drury, S, Herynk, J, 2011. The national tree-list layer: a seamless spatially-explicit tree-list layer for the continental United States.

Moeur, M, Stage, AR (1995) Most similar neighbor: an improved sampling inference procedure for natural resource planning. Forest Science 41, 337-359.

NatureServe (2009) 'International ecological classification standard: terrestrial ecological classifications.' NatureServe Central Databases. Arlington, VA, U.S.A. Data current as of 06 February 2009.).

O'Connell, BM, LaPoint, EB, Turner, JA, Ridley, T, Pugh, SA, Wilson, AM, Waddell, KL, Conkling, BL (2014). The Forest Inventory and Analysis Database: Database Description and User Guide Version 6.0 for Phase 2. Available at http://www.fia.fs.fed.us/library/database-documentation/current/ver6.0/FIADB_user%20guide_6-0_p2_5-6-2014.pdf [Accessed 16 July 2014].

Pierce, KB, Ohmann, JL, Wimberly, MC, Gregory, MJ, Fried, JS (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.