Available online at www.sciencedirect.com
ScienceDirect
Remote Sensing of Environment 112 (2008) 2232–2245
Remote Sensing of Environment
www.elsevier.com/locate/rse

Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data

Andrew T. Hudak [a,*], Nicholas L. Crookston [a], Jeffrey S. Evans [a], David E. Hall [a], Michael J. Falkowski [b]

[a] *Forest Service, U.S. Department of Agriculture, Rocky Mountain Research Station, Forestry Sciences Laboratory, 1221 South Main St., Moscow, ID 83843, United States*
[b] *University of Idaho, Department of Forest Resources, 6th and Line Streets, Moscow, ID 83844-1133, United States*

Received 21 February 2007; received in revised form 25 September 2007; accepted 6 October 2007

Abstract

Meaningful relationships between forest structure attributes measured in representative field plots on the ground and remotely sensed data measured comprehensively across the same forested landscape facilitate the production of maps of forest attributes such as basal area (BA) and tree density (TD). Because imputation methods can efficiently predict multiple response variables simultaneously, they may be usefully applied to map several structural attributes at the species-level. We compared several approaches for imputing the response variables BA and TD, aggregated at the plot-scale and species-level, from topographic and canopy structure predictor variables derived from discrete-return airborne LiDAR data. The predictor and response variables were associated using imputation techniques based on normalized and unnormalized Euclidean distance, Mahalanobis distance, Independent Component Analysis (ICA), Canonical Correlation Analysis (aka Most Similar Neighbor, or MSN), Canonical Correspondence Analysis (aka Gradient Nearest Neighbor, or GNN), and Random Forest (RF). To compare and evaluate these approaches, we computed a scaled Root Mean Square Distance (RMSD) between observed and imputed plot-level BA and TD for 11 conifer species sampled in north-central Idaho. We found that RF produced the best results overall, especially after reducing the number of response variables to the most important species in each plot with regard to BA and TD. We concluded that RF was the most robust and flexible among the imputation methods we tested. We also concluded that canopy structure and topographic metrics derived from LiDAR surveys can be very useful for species-level imputation.
Published by Elsevier Inc.

Keywords: Forestry; *k*-NN imputation; LiDAR remote sensing; Mapping; Random forest

1. Introduction

Imputation algorithms can provide predictions of timber volume (Mäkelä & Pekkarinen, 2004), basal area (Franco-Lopez et al., 2001; LeMay & Temesgen, 2005), stems per hectare (LeMay & Temesgen, 2005), timber yield (Maltamo & Eerikäinen, 2001), and species-level forest inventory data (Temesgen et al., 2003). The development of *k*-nearest neighbor (*k*-NN) imputation techniques has been driven by the widespread availability of moderate resolution, multi-spectral satellite imagery, particularly from Landsat (Tomppo, 1991; Nilsson, 1997; Katila & Tomppo, 2001; Tomppo et al., 2002; Tomppo & Halme, 2004). Imputation methods have grown especially popular for their ability to relate simultaneously multiple attributes of interest from relatively expensive forest inventories to relatively inexpensive satellite data, thus greatly enhancing the efficiency by which forest inventory may be applied towards characterizing entire forest landscapes at a reasonable cost (McRoberts et al., 2002; Ohmann & Gregory, 2002; Kim & Tomppo, 2006).

The inclusion of predictor variables derived from high-resolution remotely sensed data may improve imputed estimates of forest characteristics. For example, Tuominen & Pekkarinen (2004) demonstrated that including independent variables derived from high-resolution aerial photography improved

* Corresponding author. Tel.: +1 208 883 2327; fax: +1 208 883 2318.
E-mail addresses: ahudak@fs.fed.us (A.T. Hudak), ncrookston@fs.fed.us (N.L. Crookston), jevans02@fs.fed.us (J.S. Evans), dehall@fs.fed.us (D.E. Hall), falk4587@uidaho.edu (M.J. Falkowski).

0034-4257/$ - see front matter. Published by Elsevier Inc.
doi:10.1016/j.rse.2007.10.009

estimates of stem volume considerably, as compared to other studies utilizing data from only moderate resolution sensors such as Landsat. However, the errors attained by Tuominen & Pekkarinen (2004) upon including aerial photography variables were still high (relative RMSE = 58%), leading them to suggest that the inclusion of LiDAR-derived metrics may improve the quality of predictions from imputation algorithms. Indeed, Maltamo et al. (2006) demonstrated that a most similar neighbor (MSN) imputation algorithm including both LiDAR data and aerial photography textural information as predictor variables reduced the error of stem volume predictions by 23%, as compared to predictions attained when using aerial photography texture alone.

Active LiDAR remote sensing is now widely regarded as having greater potential for characterizing highly variable forest canopy structure than passive optical imagery, including aerial photos (Lefsky et al., 2001; Koukoulas & Blackburn, 2005). Relationships between canopy structure attributes characterized with LiDAR, and stand structure attributes characterized in field plots, appear strong and non-asymptotic in temperate coniferous (Means et al., 1999, 2000; Lefsky et al., 2005a,b), temperate deciduous (Lefsky et al., 1999b), and tropical evergreen (Drake et al., 2002a) forests. LiDAR is sensitive to canopy structure variation even in high biomass coniferous (Lefsky et al., 1999a) and broadleaf (Drake et al., 2002b) forests, where passive optical sensors saturate. Relationships between LiDAR canopy structure measures and field measures of above-ground biomass vary little between boreal coniferous, temperate coniferous, and temperate deciduous forest biomes (Lefsky et al., 2002). These relationships do appear stronger in coniferous than in deciduous forests due to the conical architecture of conifer trees, which allows for greater penetration of LiDAR pulse energy into and through the canopy. However, LiDAR penetration into closed-canopy broadleaf forests is sufficient to distinguish a successional sequence of young, intermediate, mature, and old-growth stand conditions with statistical significance (Harding et al., 2001). Conversely, sufficient LiDAR pulse energy is reflected off the top of deciduous canopies during leaf-off conditions to distinguish individual tree crowns (Brantberg et al., 2003).

The utility of LiDAR in all forest types is not limited as much by technology as by availability and cost. LiDAR data are primarily acquired at large expense from airborne platforms over specific project areas that are typically much smaller than the spatial extent at which lower cost satellite image datasets are routinely acquired. Cost and unfamiliarity with handling LiDAR datasets have limited the operational use of LiDAR until fairly recently, as LiDAR remote sensing applications progress from the research to the operational realm. We know of no published reports of attempts to map stand structural attributes at the species-level using only LiDAR-derived predictor variables. Most demonstrations of species-level mapping have used hyperspectral imagery (Clark et al., 2005), often with upwards of 200 spectral bands, while LiDAR systems operate at a single wavelength. On the other hand, within the spatial extent of a single two-dimensional image pixel, LiDAR surveys can provide literally hundreds of three-dimensional points. The canopy height profile derived from these LiDAR points could be termed a "structural signature" with at least comparable information content to a hyperspectral signature derived from an image pixel.

Hudak et al. (2006) used stepwise multiple regression and best subsets regression to predict total plot-level basal area (BA; m^2/ha) and tree density (TD; trees/ha) from satellite image and airborne LiDAR data. They found that predictor variables derived from the LiDAR data explained 89% and 87% of the variation in total plot-level BA and TD, respectively. Forest managers, however, more often desire maps of basic structural attributes for species of interest. Therefore, the goal in this analysis was to impute simultaneously plot-level BA and TD by species. In this paper, we evaluate nine methods for imputing plot-level BA and TD of 11 conifer species occurring in the mixed-conifer forests of north-central Idaho, USA.

2. Background

2.1. Imputation versus regression

It is helpful to explain nearest neighbor (NN) imputation, a form of nonparametric regression, by comparing it to parametric regression. The objective in either case is to predict response variables measured intermittently across the landscape at plots (e.g., forest inventory data) from predictor variables measured continuously across the landscape and then partitioned into contiguous pixels (e.g., remotely sensed imagery). Both predictor and response variables are measured at field sample plot units to comprise the *reference set*. Only predictor variables are measured in pixel units across the entire landscape to comprise the *target set*. The empirical relationships between predictor and response variables for units in the reference set are used to predict the response variables for units in the target set. Importantly, the sample plots for collecting reference set units across the landscape are assumed to characterize the entire range of variability in the predictor variables. It is also assumed that the field plot and LiDAR samples characterize the entire reference set units, and that the LiDAR samples characterize the entire target set units. Ideally, reference set units and target set units are the same size.

Regression predictions can be represented geometrically as a curve where the predictions fall along the curve while the deviations of the observations from the curve represent residual error. This error is usually quantified as the sum of the squared differences between observed and predicted values. Taking the square root of the mean of this sum is the common regression model error statistic, the Root Mean Square Error (RMSE),

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - p}} \quad (1)$$

where:

y_i = observed
$\hat{y}_i$ = predicted
n = number of observations
p = number of parameters

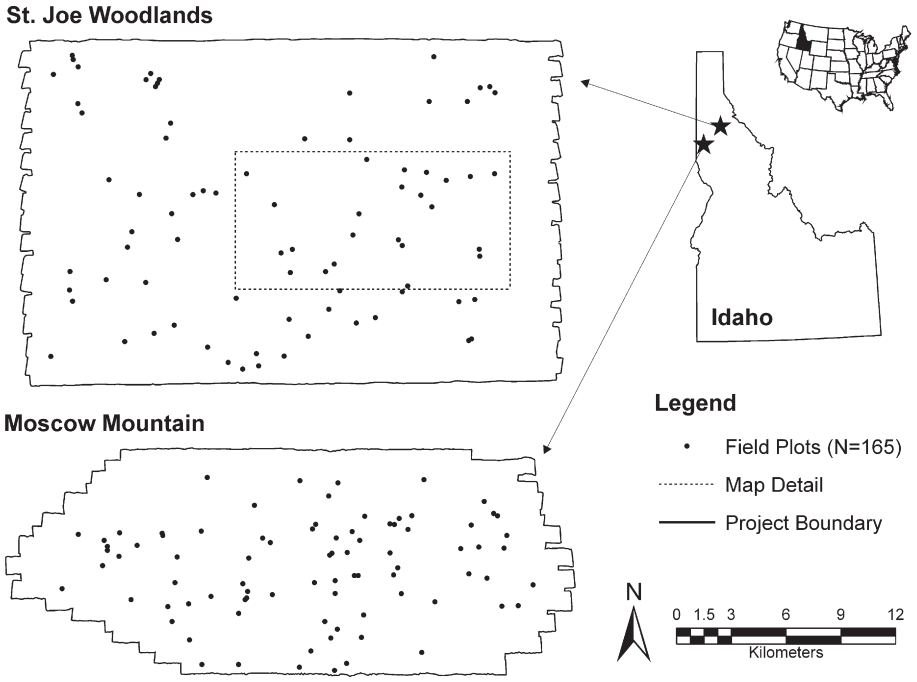

Fig. 1. Moscow Mountain and St. Joe Woodlands study areas in north-central Idaho, USA.

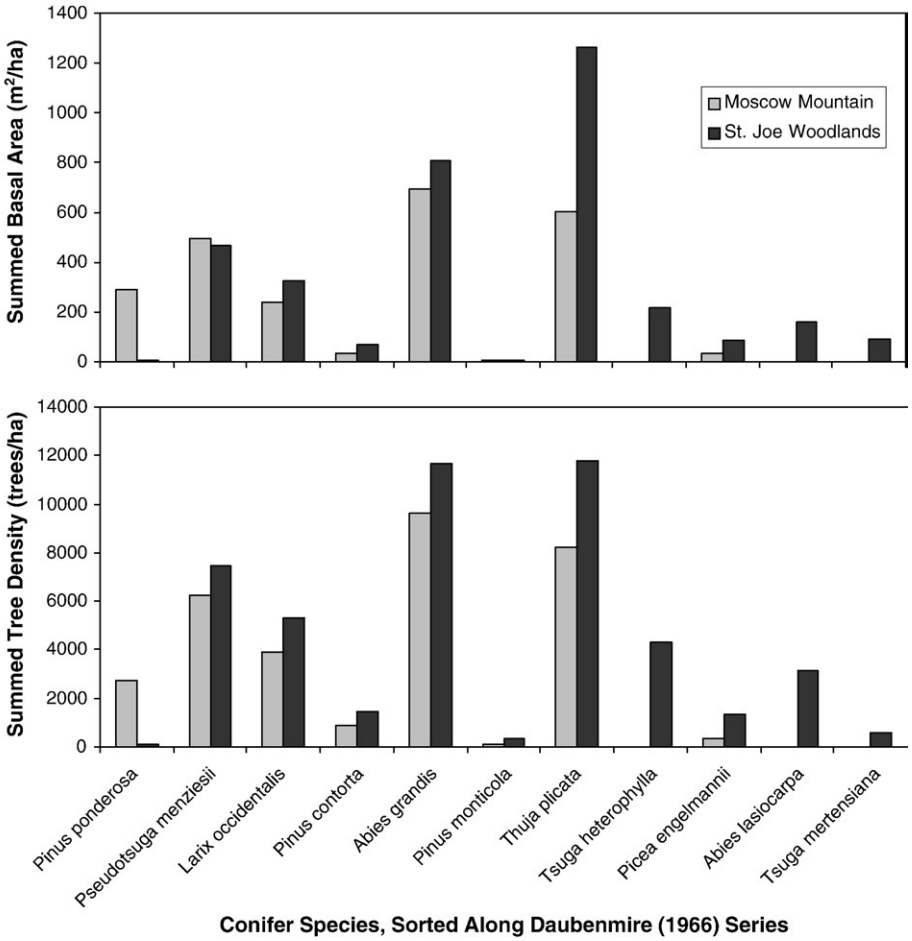

Fig. 2. Cumulative a) basal area and b) tree density measured in 165 field plots at the Moscow Mountain and St. Joe Woodlands study areas, for 11 conifer species of interest. The tree species are organized along the Daubenmire (1966) Series, to illustrate compositional differences along the warm/dry → cool/wet temperature/moisture gradient that spans the two study areas.

Imputations do not share the same mix of error components as regression predictions. Imputation errors are almost always greater than regression errors because the errors do not result from a least-squares minimization like in regression (Stage & Crookston, 2007). Imputations are instead calculated from the pool of observations and include a distance component. The difference between imputed and observed values provides a measure of model error analogous to RMSE that can be similarly quantified as the sum of squared differences. This statistic is the Root Mean Squared Difference (Stage & Crookston, 2007). To standardize the RMSD and enable comparison of imputation models across multiple response variables with varying measurement units, the RMSD was divided by the standard deviation of the observations. This scaled RMSD statistic can be used to compare and evaluate alternative imputation models just as the RMSE statistic is often used to compare and evaluate alternative regression models, provided in both cases that the residual variance is homogeneous.

A predicted response using regression can differ from any of the observed responses. On the other hand, an imputed response using a single nearest neighbor will be the response from the most similar reference observation based on whichever reference unit has the highest degree of similarity to the target unit with respect to the predictor variables. Imputation with a single nearest neighbor produces imputations with similar variance structure to that of the observations (Moeur & Stage, 1995; Franco-Lopez et al., 2001; McRoberts et al., 2002). Depending on user objectives, this can be an advantage over the higher accuracies attainable by predicting response variables via regression, or using a large number of nearest neighbors via imputation, which reduces the variance in the predictions relative to the observations.

2.2. Random Forest

The Random Forest (RF) method requires some elaboration because it differs fundamentally from traditional imputation methods. A RF randomly and iteratively samples the data and variables to generate a large group, or forest, of classification and regression trees (CART). The classification output from RF represents the statistical mode of many decision trees, hence achieving a more robust model than a single classification tree produced by a single model run (Breiman, 2001). The iterative nature of RF affords it a distinct advantage over the other methods considered in this analysis, as this effectively bootstraps the data for more robust predictions. Random subsets of predictor variables allow derivations of variable importance measures and prevent problems associated with correlated variables and overfitting (Breiman, 2001). Variable importance can be assessed by two measures (Liaw & Wiener, 2002). One provides a measure of accuracy, by quantifying the degree to which inclusion of a variable in the model decreases the mean squared error. The other importance measure is the Gini index, a measure of node impurity, or the degree to which a variable produces terminal nodes in the forest of classification trees. Splitting a node on a variable causes the Gini index for the two descendent nodes to be less than the parent node. Summing these decreases in the Gini index for a variable across the forest of classification trees provides a measure of variable importance (Breiman et al., 2006).

Strictly speaking, RF is not an imputation method, but a classification method. The RF analog to multivariate distance (as with the other imputation methods) is defined as one minus the proportion of shared terminal nodes in the forest of classification trees. The RF method can be summarized as follows: Each continuous response variable is discretized, and reference observations are then classified with respect to the other response variables and the predictor variables. The resulting classification trees are then concatenated to calculate the proportion of shared terminal nodes across all the response variables for the purpose of identifying "nearest neighbors." Although the response variables are discretized to enable RF, many classes can be defined so as to approximate continuous variable distributions. Also, RF can be employed in an unsupervised manner, with the response variables excluded from the model, and still do a reasonable job of identifying nearest neighbors. However, including the response variables in the model does improve classification accuracy.

3. Methods

3.1. Study areas

The Moscow Mountain (32,708 ha) and St. Joe Woodlands (55,684 ha) study areas are situated in north-central Idaho, USA (Fig. 1). Both areas are comprised of actively managed mixed-conifer forests. Elevation ranges are 777–1518 m at Moscow Mountain and 638–2005 m at St. Joe Woodlands. The drier conditions at Moscow Mountain produce more open canopy structure than in the wetter St. Joe Woodlands. Twelve conifer species have been described by Daubenmire (1966) along a temperature/moisture gradient, ranging from *Pinus ponderosa* at the warm/dry end (common on southern or western aspects, especially in the Moscow Mountain area) to *Pinus albicaulis* at the cool/wet end (not sampled in our study areas but occurs rarely at the highest elevations in St. Joe Woodlands) (Fig. 2).

3.2. Response variables

Field plots ($N=165$) were established in locations following a stratified random sampling design. The stratification layers were defined by intersecting three elevation strata from a 30-m × 30-m USGS digital elevation model (DEM), three solar insolation strata (HEMI, 2000), and nine middle infrared corrected normalized difference vegetation index (NDVIc) strata (Nemani et al., 1993) calculated from an 18 August 2002 Landsat ETM+ image. The NDVIc incorporates middle infrared band 5 into the NDVI formula and has been shown to be a better indicator of leaf area index than NDVI in mixed-conifer forests of northern Idaho (Pocewicz et al., 2004). The goal was to establish one sample plot in each of the 81

resulting strata for each study area, to insure that the plots were distributed well across these three stratification variables.

Plot centers were geolocated (Trimble Pro-XR) by logging a minimum of 150 points that were subsequently differentially corrected using online base station files, then averaged for a final three-dimensional (3D) point position with ±0.8 m horizontal and ±1.1 m vertical accuracy. The fixed-radius plots were 0.04 ha (0.1 acre) at Moscow Mountain and 0.08 ha (0.2 acre) at St. Joe Woodlands. Within each plot, all trees with diameters at breast height (dbh) ≥12.7 cm (5 in.) were measured. Eleven plots at Moscow Mountain lacked trees by this definition but were included in this analysis. At the opposite end of the structure gradient is the old-growth stand condition, which was not selected among the stratified random sample plots due to its rarity in both study areas. Therefore, a supplementary old-growth plot was added to the sampling design in each study area. The center of each old-growth plot was randomly located within the old-growth stand to minimize subjectivity.

We calculated plot-level BA and TD by species from the tree diameter and tally data measured in the sample plots. This amounted to plot-level BA and TD for 11 conifer species (Fig. 2), five deciduous species that occurred infrequently (*Acer glabrum*, *Betula occidentalis*, *Populus balsamifera* (=*P. trichocarpa*), *Populus tremuloides*, and *Salix* spp.), one unknown species (for failure to record on the datasheets), and plot-level BA and TD totals (summed for all species in the plot), for a total of 36 response variables.

3.3. Predictor variables

The candidate predictor variables used in this analysis were derived from an airborne LiDAR survey. Horizons, Inc. (Rapid City, SD) acquired the data during the summer of 2003, using an ALS40 discrete-return system operating at 1064 nm, a pulse rate of 20 kHz, and flown at an altitude of 2438 m above mean terrain. The nominal post spacing was 1.95 m with a 30 cm footprint and up to three returns collected per laser pulse. The data were delivered in the form of unclassified point data. Evans and Hudak (2007) developed the Multiscale Curvature Classification algorithm to classify the returns as either ground or non-ground. The classified ground returns were interpolated into a 2 m DEM, from which several topographic variables were then derived (Table 1).

Subtracting the 2-m DEM from the unclassified LiDAR returns produced a canopy height layer normalized for topography. Returns classified as ground returns equaled 0 m in height by definition. Returns greater than 0 m in height were considered non-ground returns. Returns greater than 1 m in height were considered vegetation returns. To build empirical relationships with the field data, a large variety of candidate predictor variables were calculated from the LiDAR returns within every plot footprint: vegetation density was calculated as the percentage of total returns that were vegetation returns; percentages of returns within six defined canopy height strata were calculated; and 19 distributional statistics were calculated from the height and intensity values of the vegetation returns. In total, 60 candidate predictor variables were generated for modeling (Table 1).

Table 1
Sixty candidate predictor variables, with twelve selected variables indicated in bold

Variable	Description
EAST	UTM Easting (meters)
NORTH	UTM Northing (meters)
ELEV	Elevation (meters)
SLP	Slope (degrees)
TSRAI	Topographic Solar Radiation Aspect Index (Roberts & Cooper, 1989)
SCOSA	Percent slope × cos(Aspect) transformation (Stage, 1976)
SSINA	Percent slope × sin(Aspect) transformation (Stage, 1976)
INSOL	Solar insolation (HEMI, 2000)
CRR	Canopy relief ratio (Pike & Wilson, 1971)
HMIN	Heights minimum
HMAX	Heights maximum
HRANGE	Heights range
HMEAN	Heights mean
HAAD	Heights average absolute deviation
HMAD	Heights median absolute deviation
HSTD	Heights standard deviation
HVAR	Heights variance
HSKEW	Heights skewness
HKURT	Heights kurtosis
HCV	Heights coefficient of variation
H05PCT	Heights 5th percentile
H10PCT	Heights 10th percentile
H25PCT	Heights 25th percentile
H50PCT	Heights 50th percentile (median)
H75PCT	Heights 75th percentile
H90PCT	Heights 90th percentile
H95PCT	Heights 95th percentile
HIQR	Heights interquartile range
IMIN	Intensity minimum
IMAX	Intensity maximum
IRANGE	Intensity range
IMEAN	Intensity mean
IAAD	Intensity average absolute deviation
IMAD	Intensity median absolute deviation
ISTD	Intensity standard deviation
IVAR	Intensity variance
ISKEW	Intensity skewness
IKURT	Intensity kurtosis
ICV	Intensity coefficient of variation
I05PCT	Intensity 5th percentile
I10PCT	Intensity 10th percentile
I25PCT	Intensity 25th percentile
I50PCT	Intensity 50th percentile (median)
I75PCT	Intensity 75th percentile
I90PCT	Intensity 90th percentile
I95PCT	Intensity 95th percentile
IIQR	Intensity interquartile range
DENSITY	Canopy cover (Vegetation returns/Total returns × 100)
STRATUM0	Percentage of ground returns = 0 m
STRATUM1	Percentage of non-ground returns >0 m and <=1 m in Height
STRATUM2	Percentage of vegetation returns >1 m and <=2.5 m in height
STRATUM3	Percentage of vegetation returns >2.5 m and <=10 m in height
STRATUM4	Percentage of vegetation returns >10 m and <=20 m in height
STRATUM5	Percentage of vegetation returns >20 m and <=30 m in height
STRATUM6	Percentage of vegetation returns >30 m in height
TEXTURE	Standard deviation of non-ground returns >0 m and <=1 m
PCT1	Percentage 1st returns
PCT2	Percentage 2nd returns
PCT3	Percentage 3rd returns
NOTFIRST	Percentage 2nd or 3rd returns

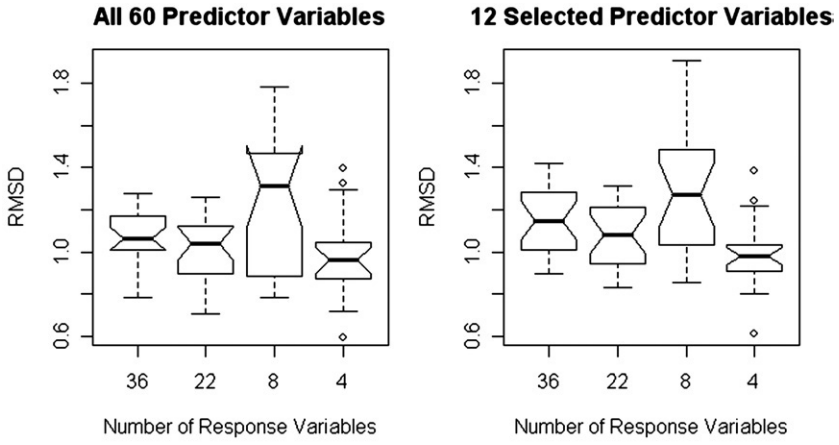

Fig. 3. Scaled RMSD distributions for imputing plot-level BA and TD of 11 conifer species from RF models imputing either 36, 22, or 8 response variables, or from an RF2 model imputing 4 response variables, using either a) All 60 predictor variables, or b) 12 selected predictor variables. Non-overlapping notches between boxplots suggest that the medians significantly differ (Chambers et al., 1983).

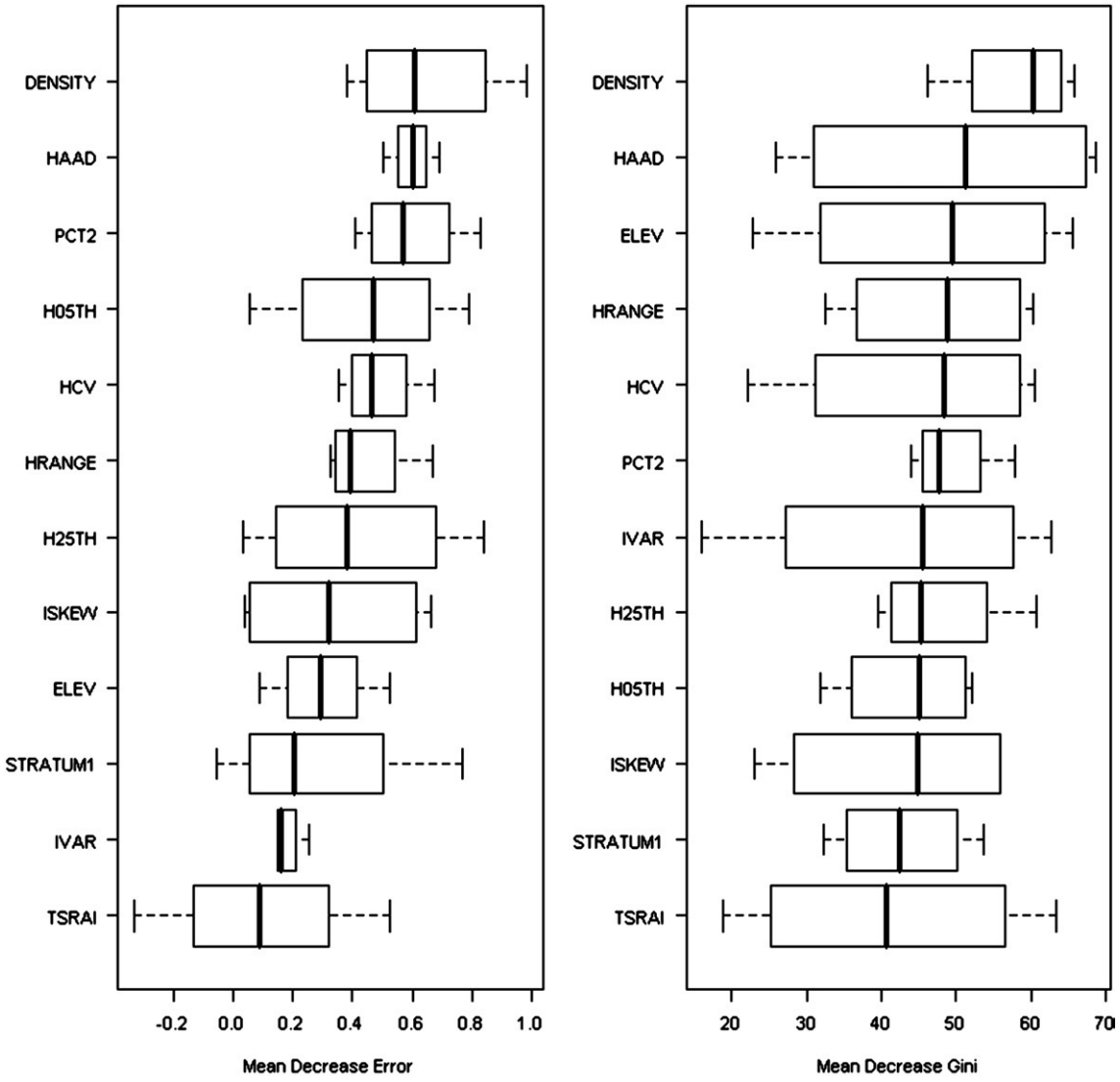

Fig. 4. Variable importance graphs for the 12 selected predictor variables as quantified by a) mean decrease in the mean squared error or b) mean decrease in the Gini index using the RF2 method. Boxplots represent the distribution of importance values for imputing plot-level BA and TD of the 11 conifer species of interest. The predictor variables are sorted in descending order from top to bottom based on median importance values.

In preparation for mapping, target unit rasters were made at a 30-m × 30-m resolution across both study areas. The 60 candidate predictor variables were generated from the LiDAR points within each target unit as they were within each plot footprint for the reference observations.

3.4. Model building and evaluation

The tree species with both commercial and non-commercial value in our study area were all conifers. We used plot-level BA and TD for the 11 conifer species encountered, amounting to 22 response variables, to compare and evaluate all models consistently with the intention of minimizing the mean scaled RMSD across these 22 response variables.

Many of the predictor variables in our starting list of 60 candidates were too highly correlated to include in the same model. Given RF's ability to bootstrap the available data, we used it to prune the predictor variables down to a parsimonious subset. A stepwise looping procedure was used to iterate RF, discarding the least important of the candidate variables at each iteration, based on the Gini index, until only a single predictor variable remained. This process is analogous to backwards stepwise multiple regression. Combinations of predictor variables that produced particularly low mean scaled RMSD for the 22 response variables of primary interest were saved for further consideration, in a process similar to best subsets regression.

3.5. Imputation methods

Imputation was accomplished in this study using the yaImpute package in R (Crookston and Finley, in review). Many imputation methods can be used for associating target and reference units. The following eight methods are the most common in the literature and were employed in this study to characterize the multivariate relationships between predictor and response variables that enable NN imputation:

1) EUC — Euclidean distance is computed in a multivariate predictor variable space normalized by subtracting the mean and dividing by the standard deviation, for each predictor variable.

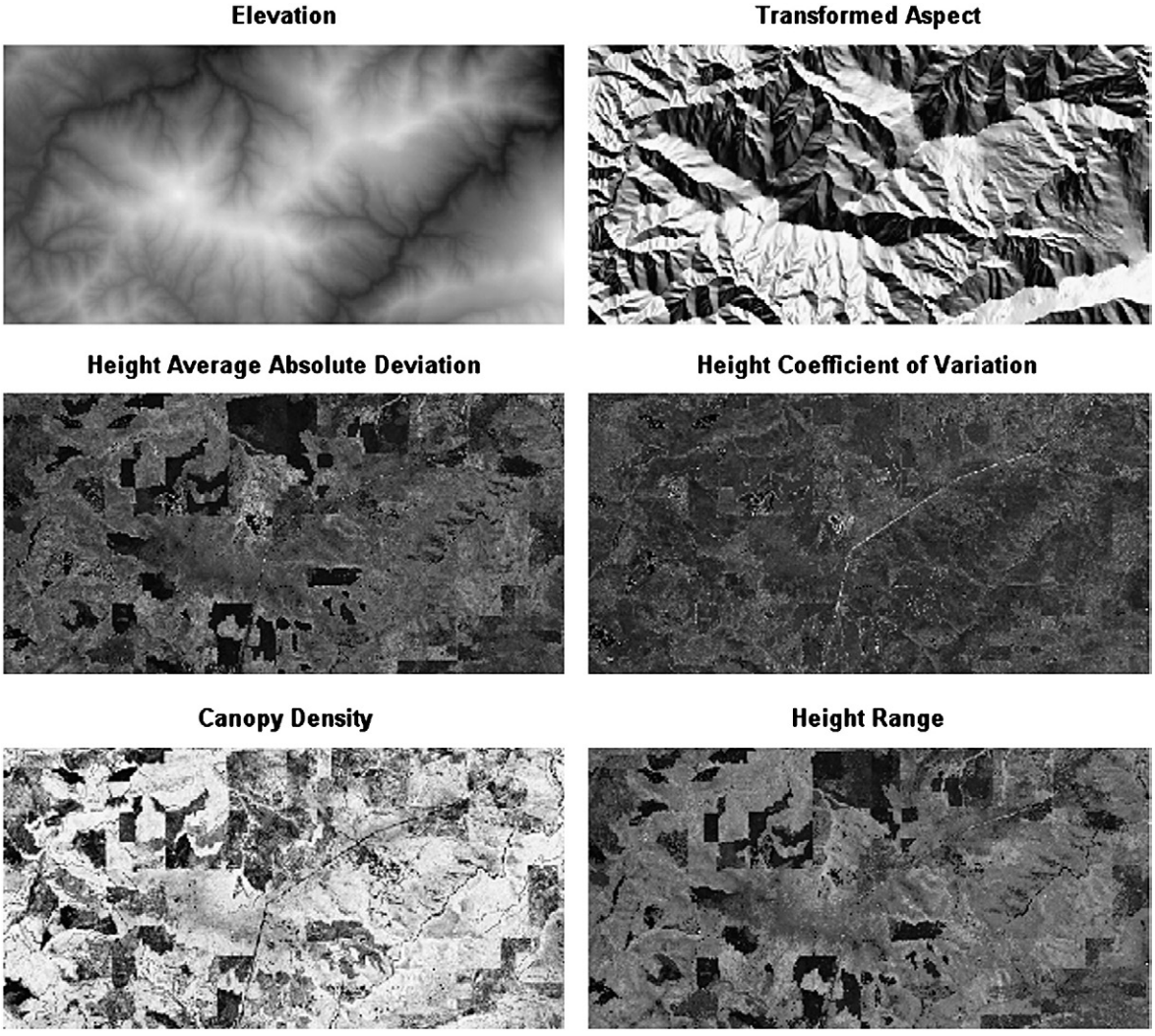

Fig. 5. Image subsets (15 km × 7.5 km) in the center of St. Joe Woodlands illustrating the six most important predictor variables selected for imputation. Larger values appear brighter.

2) RAW — Like Euclidean distance, but without normalization.
3) MAL — Mahalanobis distance is the dimensional components of Euclidean distance weighted by the inverse of the sample variance–covariance matrix (Mahalanobis, 1936).
4) ICA — Like Mahalanobis, but based on Independent Component Analysis where distance is computed in a projected space defined by components that are statistically independent and have assumed non-Gaussian distributions (Hyvarinen & Oja, 2000).
5) MSN — Most Similar Neighbor (Moeur & Stage, 1995) distance is computed in a projected canonical space.
6) MSN2 — Like MSN, but with canonical correlations weighted by their variances (Crookston et al., 2002).
7) GNN — Gradient Nearest Neighbor (Ohmann & Gregory, 2002) distance is computed using a projected ordination of predictors based on canonical correspondence analysis.
8) RF — Distance in a Random Forest is calculated as one minus the proportion of classification trees where a target observation is in the same terminal node as a reference observation.

For any of the methods listed above, any number (k) of reference observations can be chosen to impute target units. The k-NN prediction for a continuous variable is then calculated from the k-nearest neighbors as an average, with the option of inverse distance weighting (where "distance" is measured in multivariate predictor variable space, not in geographic space). Increasing k effectively shifts the prediction toward the sample mean, or alters the shape in the distribution of predictions toward normal, which is unrealistic when the distribution of observations is non-normal or skewed (as is often the case). Increasing k reduces the imputation error (up to a point), but can also be unproductive to the extent that it reduces pure error. Pure error is the component of imputation error, apart from measurement error, that describes the variability in response values with the same predictor values around the true mean of the response values (Stage & Crookston, 2007), which is an important consideration for comparing the utility of imputation models.

It is ultimately up to the user to decide whether s/he is more interested in predictive accuracy or reproducing similar variance structure in the imputations as in the observations. To increase accuracy, increasing k is advised up to the point where additional nearest neighbors provide no significant accuracy improvement (Muinonen et al., 2001). Franco-Lopez et al. (2001) recommended that if variance in the imputations similar to variance in the observations is desired, then a single nearest neighbor k is appropriate (McRoberts et al., 2002).

A variable reduction strategy was employed to reduce the number of response variables down to the two individual species having the largest plot-level BA and TD values (continuous variables) in each plot, along with the corresponding names of these two species (categorical variables). These names were treated as factors by RF to streamline the subsequent classification. This reduced the number of response variables from 36 to four and proved advantageous in this analysis, both for predictor variable pruning in the procedure just described, and as a ninth imputation method (RF2) in its own right.

Response variable values were imputed to target units across both study areas at a 30-m × 30-m resolution.

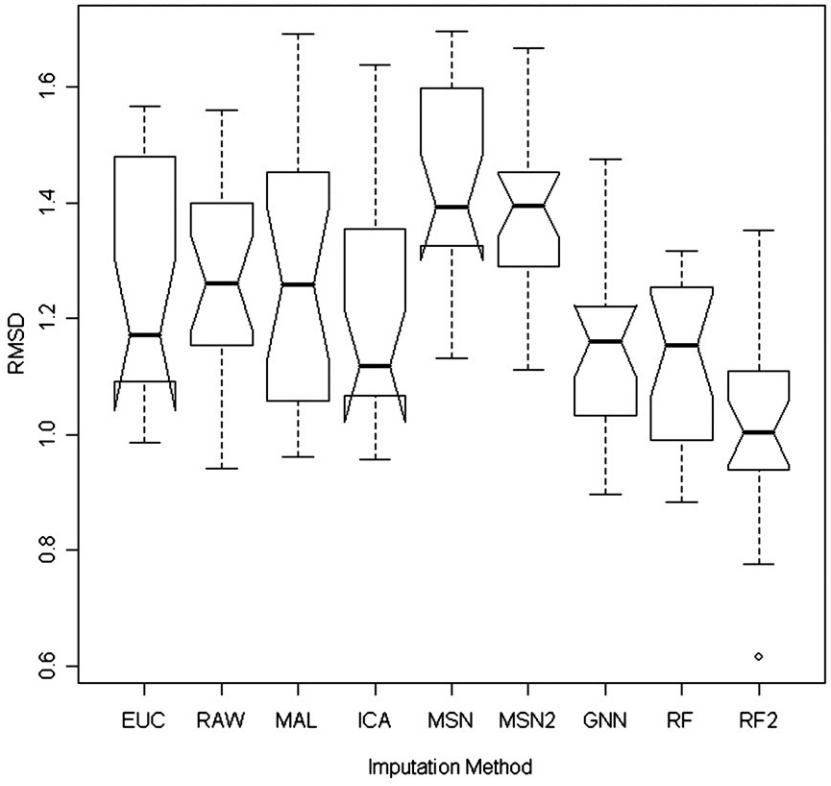

Fig. 6. Scaled RMSD distributions for imputing plot-level BA and TD of 11 conifer species with nine imputation methods, using 12 selected predictor variables. Non-overlapping notches between boxplots suggest that the medians differ significantly (Chambers et al., 1983).

4. Results

4.1. Variable selection

The random element of RF causes the results to vary slightly after each run. This variation can be stabilized by increasing the number of classification trees built for each response variable of interest. After some experimentation, we found that generating 300 classification trees per response variable produced repeatable results without an undue amount of processing time.

We considered how including different sets of response variables in the imputation models affected the scaled RMSD. We compared the full set of 36 response variables, the 22 coniferous variables of primary interest (plot-level BA and TD of 11 conifer species), eight variables consisting of the plot-level BA and TD of the four most prevalent conifer species (PSME, LAOC, ABGR, and THPL; Fig. 2), or four variables (maximum plot-level BA and TD by species in each plot, along with the corresponding species names). Results showed similar trends in the distribution of scaled RMSD whether all 60 or only 12 predictor variables were used in the RF2 method (Fig. 3). Considering only the four response variables as defined for the RF2 method produced the lowest scaled RMSD. Basing the imputation models on the same 22 response

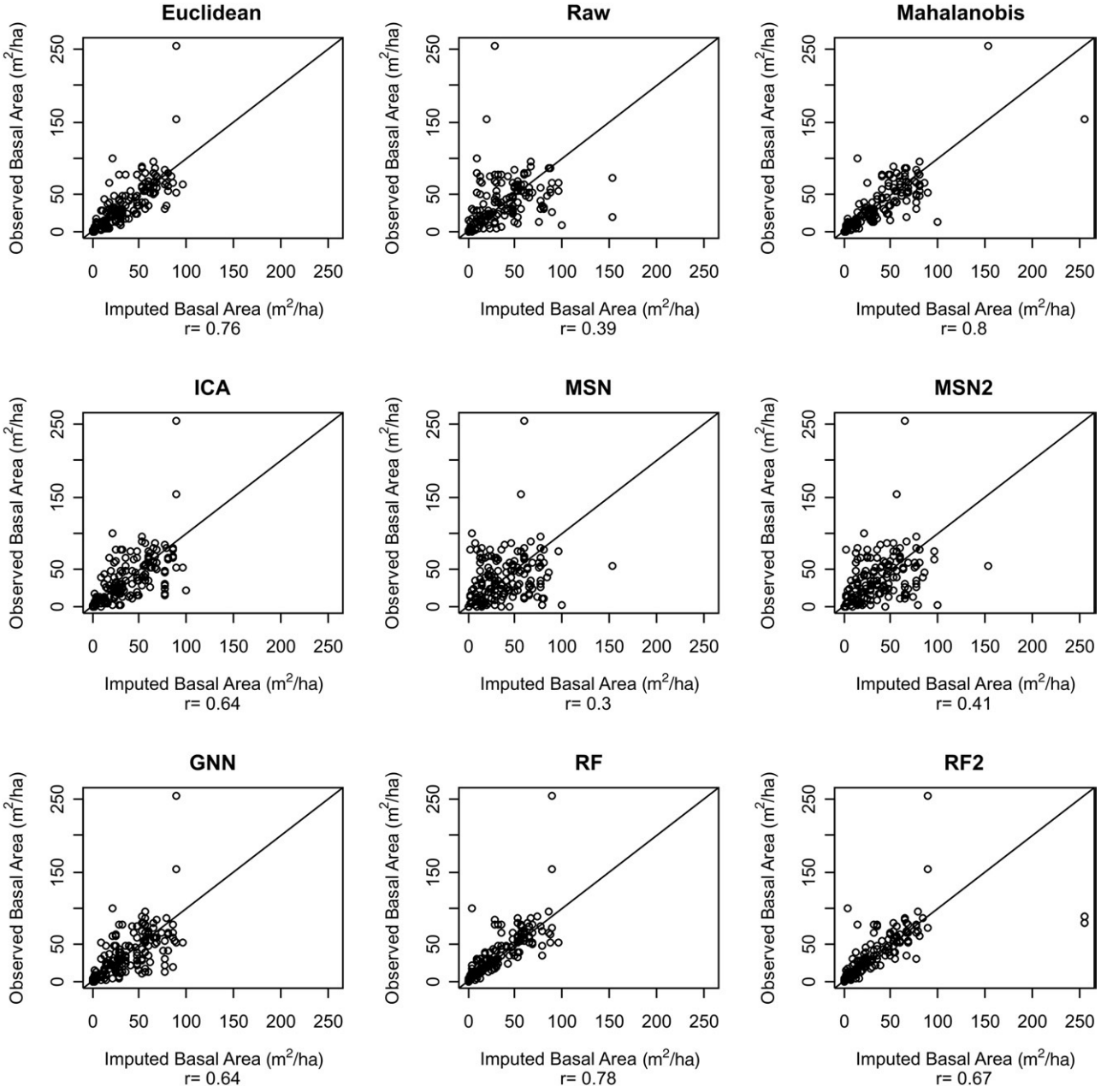

Fig. 7. Scatterplots of imputed versus observed plot-level BA from nine imputation methods. The two largest observations in each scatterplot represent the two old-growth plots. Pearson correlations are provided below each graph.

variables used for evaluation produced only slightly lower scaled RMSD than basing the imputation models on all 36 response variables. Basing the imputation models on only the eight most prevalent conifer variables produced the highest scaled RMSD (Fig. 3).

Based on the preceding results (Fig. 3), we used the RF2 method of considering four response variables to select predictor variables. The random element of RF2 caused the stepwise variable selection results to vary considerably between runs, since RF2 reran after discarding each variable, one by one, beginning with 60 and ending with 1. In other words, the predictor variable ranks, based on their Gini importance values, changed every time RF2 reran. Therefore, we looped the stepwise procedure through 1000 iterations and recorded predictor variable subsets that produced low mean scaled RMSD values averaged across our 22 response variables of interest. More than 60,000 RF2 iterations helped us to select 12 predictor variables with consistently high importance values (Fig. 4), six of which are illustrated in Fig. 5. We also inspected a correlation matrix of the 12 selected predictor variables to check that the maximum correlation between any two predictor variables was limited to 0.9 to guard against undue redundancy.

4.2. Imputation methods

Eight of the nine imputation methods compared were based on the 12 selected predictor variables and all 36 response variables. The ninth method (RF2) included the variable reduction transformation to reduce the 36 response variables to four. Thus, all nine methods evaluated were based on different sets of response variables than the 22 response variables used for evaluation. The variable reduction transformation of RF (RF2) produced the lowest distribution of scaled RMSD values (Fig. 6). The next best methods were ICA, GNN, and RF. The worst methods were MSN and MSN2, and in between were the EUC, RAW, and MAL methods (Fig. 6).

Basing the imputation models on a subset of response variables does not limit the user from subsequently imputing values of additional response variables. For instance, total plot-level BA was imputed using the RF2 method and compared

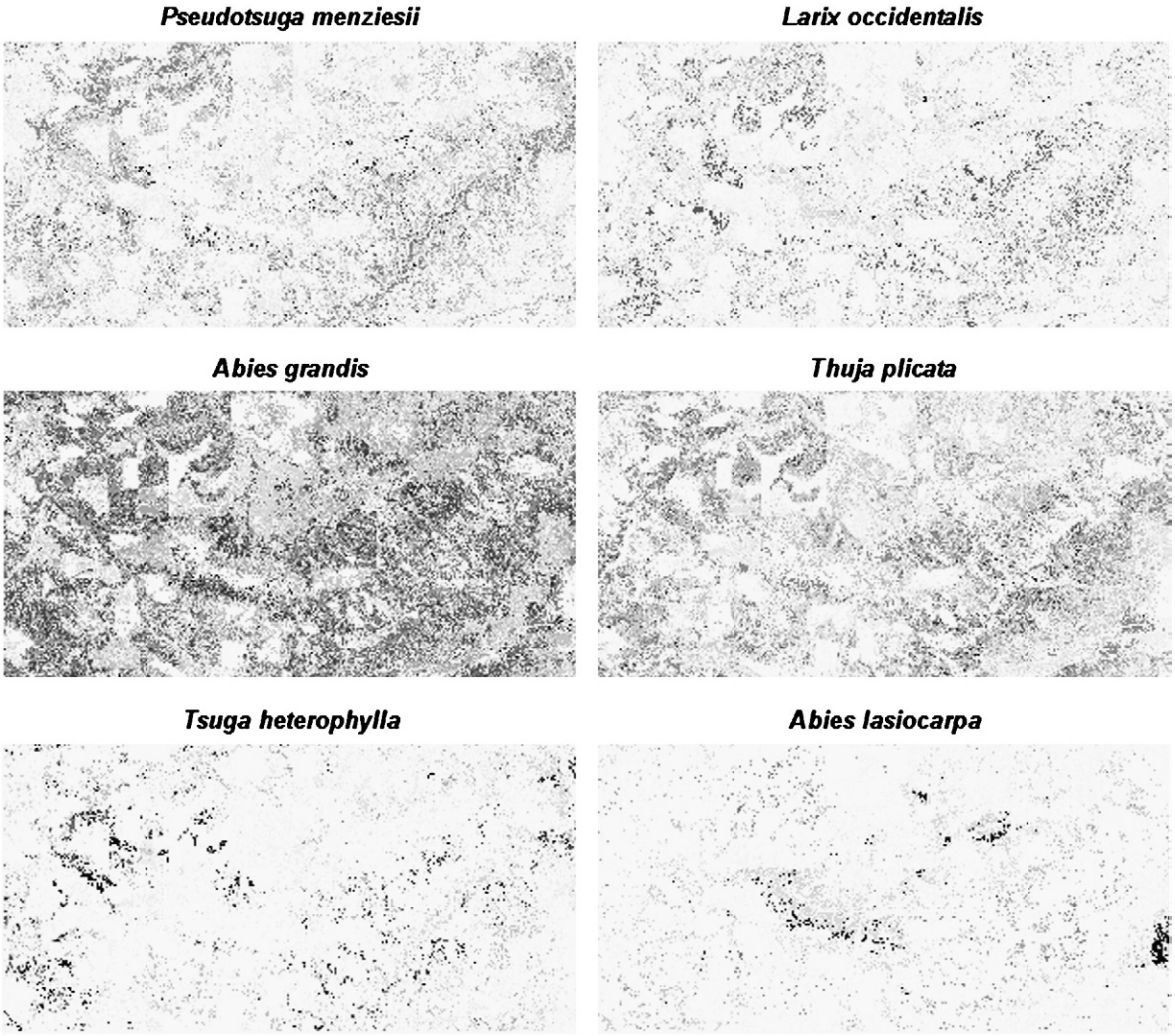

Fig. 8. Image subsets (15 km × 7.5 km) in the center of St. Joe Woodlands illustrating TD (trees/ha) imputed using the RF2 method for the six most prevalent conifer species. Zero values appear white.

favorably to results from the eight other methods (Fig. 7). Overall, total plot-level BA imputed using the MAL and RF methods correlated most highly to observed total plot-level BA. Results for total plot-level TD (not shown) were very similar.

We graphed histograms of imputed values and the shape of these distributions from all nine methods closely matched the shape in the distribution of observations. We also experimented using more than a single nearest neighbor in the nine methods, and the shape of these distributions grew progressively more rounded as k increased. The higher frequency of imputations near the mean as k increased also increased accuracy, as measured by the Pearson correlations between observations and imputations. These experiments confirmed our previously stated assumption that as k increases, so does accuracy, but at the expense of an altered variance structure, as exhibited by the changed shape in the distribution of the imputations relative to the observations.

We also tested for heteroscedasticity in the RF2 imputation residuals. First, we examined scatterplots of observed versus imputed values for several response variables, which revealed no obvious evidence of heterogeneous variance. Next, we sorted the 165 observations and divided them into 11 bins with 15 observations per bin. Across these 11 bins, we then graphed the standard deviation of the RF2 imputation residuals against the mean of the observations, repeating for ten response variables (plot-level BA and TD for PSME, LAOC, ABGR, THPL, and the sum of all species) having a sufficient number of values to populate the 11 bins. These plots showed that the residual variance was indeed homogeneous across these ten variables.

4.3. Maps

Because the RF2 method produced the lowest mean scaled RMSD over our 22 response variables of interest, we used RF2 to map them. Imputed maps of plot-level TD for six conifer species are shown in Fig. 8. (Similar maps of plot-level BA by species appear virtually the same but are not shown because the outlying old-growth plots caused plot-level BA to scale less conveniently for illustration purposes.) For clarity, we reduced the size of the map extent in Fig. 8 to a smaller region within St. Joe Woodlands that well represents the environmental gradients (Figs. 1 and 5). *Abies grandis* was the most broadly distributed species, followed by *Thuja plicata* and *Pseudotsuga menziesii*. *Larix occidentalis* and *Pinus ponderosa* were often observed in recent harvest units, where larger individuals are often left as seed trees for regeneration of these commercially-valuable species. *Pinus monticola* is represented mostly by young trees regenerating in isolated patches, which are scattered due to historic decimation by white pine blister rust or logging. *Pinus contorta* is broadly but thinly distributed. *Tsuga heterophylla* and *Picea engelmannii* are not managed explicitly, but develop naturally in the shaded understory beneath other species. Distributions of *Abies lasiocarpa* and even more so, *Tsuga mertensiana*, were restricted to higher elevations in St. Joe Woodlands, only a portion of which is included in Fig. 8. For brevity, we did not include similar maps for Moscow Mountain because only St. Joe Woodlands includes all 11 conifer species surveyed (Fig. 2).

5. Discussion

5.1. Predictor variables

Following the example of Hudak et al. (2006), satellite image reflectance bands (Advanced Land Imager) were initially considered as predictor variables. However, these variables were unhelpful in explaining variation beyond the variation that could be explained by the LiDAR metrics (Table 1), so they were simply dropped from consideration. Multispectral or hyperspectral imagery could prove more useful for discriminating between coniferous and deciduous species, although the structural differences between coniferous and deciduous species may be captured just as well, or perhaps even better, by LiDAR metrics as employed in this study. In particular, DENSITY, a physically-based measure of canopy density that cannot be derived from passive optical imagery, was the strongest predictor variable (Fig. 4).

Hudak et al. (2006) used UTM Easting and Northing in their final models, with the justification that these variables captured the warm-dry to cool-moist gradient that spans the two study areas. Although these same variables were included in this analysis, they were not among the final 12 selected predictor variables because TSRAI proved a better predictor. We found that TSRAI, which is essentially an aspect transformation, captured the fine-scale topographic variation that largely drives community composition in these study areas in a way that simple geographic coordinates cannot. The effect of aspect on vegetation community composition is so important that we considered it necessary to include a topographic variable relating to aspect among the selected predictor variables. The number of predictor variables selected (12) closely matches the number of predictor variables selected by Hudak et al. (2006) to predict total plot-level BA (10 variables) and total plot-level TD (12 variables) using multiple linear regression models.

5.2. Imputation versus regression

Pearson correlations between observed and imputed total plot-level BA and TD using the RF method were 0.78 (Fig. 7) and 0.76, respectively. These correlations were, respectively, less than or comparable to the Pearson correlations reported by Hudak et al. (2006) between the same observations and cross-validated, regression-based predictions (0.90 for total plot-level BA, 0.74 for total plot-level TD). To assess prediction bias, imputed versus regression-based total plot-level BA maps were compared by aggregating the 30-m × 30-m pixels from the maps to the stand-level, and then validating these stand-level aggregates with independent stand exam data. The multiple linear regression model of Hudak et al. (2006) strongly overpredicted total plot-level BA, as did another multiple linear regression model based on a different set of LiDAR-derived predictors. On the other hand, the imputation model in this study only weakly overpredicted total plot-level BA, as did another imputation model employing the RF method and a different set of LiDAR predictors (Hudak et al., in press).

5.3. Imputation methods

Pruning the number of predictor variables from 60 to 12 with the RF2 method produced a much more parsimonious subset of predictor variables for imputation (Fig. 4). The RF and RF2 methods can help inform the decision of which predictor variables to select in that both rows (observations) and columns (predictor variables) in the data frame are bootstrapped to provide an objective evaluation of predictor variable importance. Pruning the number of predictor variables with RF2 also achieved lower scaled RMSD for some of the other methods applied, particularly GNN (Fig. 6). The lower scaled RMSD of GNN relative to MSN and the other methods besides RF and RF2 (Fig. 6) supports the decision of Ohmann and Gregory (2002) to base the GNN method upon canonical correspondence analysis. Canonical correspondence analysis assumes that species response to environmental gradients is nonlinear (Austin et al., 1994), not linear, as is the less realistic assumption of canonical correlation analysis underlying the MSN method (Jongman et al., 1997; Ohmann & Gregory, 2002). However, there is nothing inherent in the MSN method that would preclude transformation of the predictor or response variables so as to create linear relationships among variables as is often required (Crookston et al., 2002), yet is time consuming.

We did not consider any categorical predictor variables in this analysis. With regard to the response variables, only the RF2 method treated the categorical species names as factors, which may largely explain the better RF2 results relative to the other methods. Although the RF method did not treat species names as factors, it still produced results that were generally superior to results from the other imputation methods tested (Figs. 6–7).

5.4. Sampling and scale

The 30-m × 30-m map resolution used in this study is customary for imputation models that employ 30-m × 30-m Landsat imagery (McRoberts et al., 2002; Ohmann & Gregory, 2002; Tomppo et al., 2002). The 900 m^2 area of the square target units was slightly larger than the ~810 m^2 area of the round 0.08 ha reference units characterized at St. Joe Woodlands, and over twice the ~405 m^2 area of the round 0.04 ha reference units characterized at Moscow Mountain. While the size of our target units and reference units matched fairly closely, 20-m × 20-m (400 m^2) may have been a more appropriate map resolution at Moscow Mountain, or 25-m × 25-m (625 m^2) for both study areas. We do recommend that the size of the response variable observation units match that of the predictor variable observation units as closely as possible to minimize any potential scale discrepancies. We chose to map at 30-m × 30-m resolution to facilitate comparison of our imputed maps to regression-based maps of the same study areas predicted from LiDAR and 30-m × 30-m ALI satellite imagery (Hudak et al., 2006). However, it is likely that discrete-return LiDAR surveys with sample densities equal to or higher than ours can characterize plot-level structural attributes at higher resolution than 30-m × 30-m (Reutebuch et al., 2005). The minimum density of LiDAR returns (samples) among our 165 field plots was ~0.335 m^{-2}, in a non-forested plot that produced only ground returns. This translates into ~300 samples per 30-m × 30-m target unit. Moreover, this calculation only considers the horizontal plane. Multiple-return LiDAR systems are capable of collecting more data in complex forest canopies that have the structural elements to stimulate multiple returns. This is a dynamic capability that passive optical image systems lack, with the data volume per image pixel constrained to not exceed the sum of the number of spectral bands.

6. Conclusions

We have demonstrated that plot-level structural attributes of individual tree species can be simultaneously imputed from predictor variables derived solely from LiDAR data. Useful predictor variables included LiDAR metrics calculated not just from the canopy height distributions, but also the distributions of intensity values of the vegetation returns, and most importantly, measures of vegetation density calculated within vertical canopy strata. Since LiDAR cannot differentiate species based on spectral variation, we conclude that this is possible because LiDAR can characterize detailed canopy structure variation and fine-scale topographic variation that constrains the distribution of species assemblages. Imputation of species-level structural attributes holds more promise for unbiased forest inventory than parametric linear regression-based methods. Models based on machine-learning algorithms such as RF represent a flexible and robust alternative to traditional imputation methods. We conclude that LiDAR may become an important forest inventory data source when combined with appropriately designed sample plots in the field and with appropriate modeling tools.

Acknowledgements

This research was funded through the Sustainable Forestry component of Agenda 2020, a joint effort of USDA Forest Service Research and Development and the American Forest and Paper Association. Additional funding was provided by the Rocky Mountain Research Station. Research partners included Potlatch, Inc. and Bennett Lumber Products, Inc. Curtis Kvamme, K.C. Murdock, Kasey Prestwich, Jacob Young, Bryn Parker, Stephanie Jenkins, Tessa Jones, and Jennifer Clawson collected field data. We also thank Dennis Ferguson, Kristi Coughlon, Rudy King, and four anonymous reviewers for their exceptionally thorough reviews.

References

Austin, M. P., Nicholls, A. O., Doherty, M. D., & Meyers, J. A. (1994). Determining species response functions to an environmental gradient by means of a β-function. *Journal of Vegetation Science*, *5*, 215−228.

Brantberg, T., Warner, T. A., Landenberger, R. E., & McGraw, J. B. (2003). Detection and analysis of individual leaf-off tree crowns in small footprint, high sampling density lidar data from the eastern deciduous forest in North America. *Remote Sensing of Environment*, *85*, 290−303.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5−32.

Breiman, L., Cutler, A., Liaw, A., & Wiener, M. (2006). *Breiman and Cutler's random forests for classification and regression.* http://CRAN.R-project.org/ [R package version 4.5–16].

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical methods for data analysis.* Belmont, CA: Wadsworth and Boston, MA: Duxbury.

Clark, M. L., Roberts, D. A., & Clark, D. B. (2005). Hyperspectral discrimination of tropical rain forest tree species at leaf to crown scales. *Remote Sensing of Environment*, *96*, 375−398.

Crookston, N. L., & Finley, A. (in review). yaImpute: An R package for *k*-NN imputation. Journal of Statistical Software. http://forest.moscowfsl.wsu.edu/gems/yaImputePaper.pdf. Package URL: http://cran.r-project.org/src/contrib/Descriptions/yaImpute.html

Crookston, N. L., Moeur, M., & Renner, D. (2002). *Users guide to the most similar neighbor imputation program version 2. RMRS-GTR-96.*Ogden, UT: USDA Forest Service Rocky Mountain Research Station 35 p.

Daubenmire, R. (1966). Vegetation: Identification of typal communities. *Science*, *151*, 291−298.

Drake, J. B., Dubayah, R. O., Clark, D. B., Knox, R. G., Blair, J. B., Hofton, M. A., et al. (2002). Estimation of tropical forest structural characteristics using large-footprint lidar. *Remote Sensing of Environment*, *79*, 305−319.

Drake, J. B., Dubayah, R. O., Knox, R. G., Clark, D. B., & Blair, J. B. (2002). Sensitivity of large-footprint lidar to canopy structure and biomass in a neotropical rainforest. *Remote Sensing of Environment*, *81*, 378−392.

Evans, J. S., & Hudak, A. T. (2007). A multiscale curvature algorithm for classifying discrete return lidar in forested environments. *IEEE Transactions on Geoscience and Remote Sensing*, *45*, 1029−1038.

Franco-Lopez, H., Ek, A. R., & Bauer, M. E. (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.

Harding, D. J., Lefsky, M. A., Parker, G. G., & Blair, J. B. (2001). Laser altimeter canopy height profiles: Methods and validation for closed-canopy, broadleaf forests. *Remote Sensing of Environment*, *76*, 283−297.

Helios Environmental Modeling Institute (HEMI), LLC (2000). *The solar analyst 1.0 user manual.*

Hudak, A. T., Crookston, N. L., Evans, J. S., Falkowski, M. J., Smith, A. M. S., Gessler, P., et al. (2006). Regression modeling and mapping of coniferous forest basal area and tree density from discrete-return lidar and multispectral satellite data. *Canadian Journal of Remote Sensing*, *32*, 126−138.

Hudak, A. T., Crookston, N. L., Evans, J. S., Steigers, B., Taylor, R., Hemingway, H., et al. (in press). Aggregating pixel-level basal area predictions derived from LiDAR data to industrial forest stands in Idaho. *Proceedings of the 3rd Forest Vegetation Simulator Conference.*

Hyvarinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, *13*, 411−430.

Jongman, R. H. G., ter Braak, C. J. F., & van Tongeren, O. F. R. (Eds.). (1997). *Data analysis in community and landscape ecology* Wageningen, Netherlands: Pudoc.

Katila, M., & Tomppo, E. (2001). Selecting estimation parameters for the Finnish multisource National Forest Inventory. *Remote Sensing of Environment*, *76*, 16−32.

Kim, H. -J., & Tomppo, E. (2006). Model-based prediction error uncertainty estimation for *k*-nn method. *Remote Sensing of Environment*, *104*, 257−263.

Koukoulas, S., & Blackburn, G. A. (2005). Mapping individual tree location, height and species in broadleaved deciduous forest using airborne LIDAR and multi-spectral remotely sensed data. *International Journal of Remote Sensing*, *26*, 431−455.

Lefsky, M. A., Cohen, W. B., Acker, S. A., Parker, G. G., Spies, T. A., & Harding, D. (1999). Lidar remote sensing of the canopy structure and biophysical properties of Douglas-fir western hemlock forests. *Remote Sensing of Environment*, *70*, 339−361.

Lefsky, M. A., Cohen, W. B., Harding, D. J., Parker, G. G., Acker, S. A., & Gower, S. T. (2002). Lidar remote sensing of above-ground biomass in three biomes. *Global Ecology and Biogeography*, *11*, 393−399.

Lefsky, M. A., Cohen, W. B., & Spies, T. A. (2001). An evaluation of alternate remote sensing products for forest inventory, monitoring, and mapping of Douglas-fir forests in western Oregon. *Canadian Journal of Forest Research*, *31*, 78−87.

Lefsky, M. A., Harding, D., Cohen, W. B., Parker, G., & Shugart, H. H. (1999). Surface LIDAR remote sensing of basal area and biomass in deciduous forests of eastern Maryland, USA. *Remote Sensing of Environment*, *67*, 83−98.

Lefsky, M. A., Hudak, A. T., Cohen, W. B., & Acker, S. A. (2005). Geographic variability in lidar predictions of forest stand structure in the Pacific Northwest. *Remote Sensing of Environment*, *95*, 532−548.

Lefsky, M. A., Hudak, A. T., Cohen, W. B., & Acker, S. A. (2005). Patterns of covariance between forest stand and canopy structure in the Pacific Northwest. *Remote Sensing of Environment*, *95*, 517−531.

LeMay, V., & Temesgen, H. (2005). Comparison of nearest neighbor methods for estimating basal area and stems per hectare using auxiliary variables. *Forest Science*, *51*, 109−119.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*, 18−22.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Science of India*, *12*. (pp. 49−55).

Mäkelä, H., & Pekkarinen, A. (2004). Estimation of forest stand volumes by Landsat TM imagery and stand-level field-inventory data. *Forest Ecology and Management*, *196*, 245−255.

Maltamo, M., & Eerikäinen, K. (2001). The most similar neighbour reference in the yield prediction of *Pinus kesiya* stands in Zambia. *Silva Fennica*, *35*, 437−451.

Maltamo, M., Malinen, J., Packaln, P., Suvanto, A., & Kangas, J. (2006). Nonparametric estimation of stem volume using airborne laser scanning, aerial photography, and stand-register data. *Canadian Journal of Forest Research*, *36*, 426−436.

McRoberts, R. E., Nelson, M. D., & Wendt, D. G. (2002). Stratified estimation of forest area using satellite imagery, inventory data, and the k-nearest neighbors technique. *Remote Sensing of Environment*, *82*, 457−468.

Means, J. E., Acker, S. A., Fitt, B. J., Renslow, M., Emerson, L., & Hendrix, C. J. (2000). Predicting forest stand characteristics with airborne scanning lidar. *Photogrammetric Engineering and Remote Sensing*, *66*, 1367−1371.

Means, J. E., Acker, S. A., Harding, D. J., Blair, J. B., Lefsky, M. A., Cohen, W. B., et al. (1999). Use of large-footprint scanning airborne lidar to estimate forest stand characteristics in the Western Cascades of Oregon. *Remote Sensing of Environment*, *67*, 298−308.

Moeur, M., & Stage, A. R. (1995). Most similar neighbor: An improved sampling inference procedure for natural resource planning. *Forest Science*, *41*, 337−359.

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

Nemani, R., Pierce, L., Running, S., & Band, L. (1993). Forest ecosystem processes at the watershed scale: Sensitivity to remotely-sensed Leaf Area Index estimates. *International Journal of Remote Sensing*, *14*, 2519−2534.

Nilsson, M. (1997). Estimation of forest variables using satellite image data and airborne lidar. Ph.D. thesis, Swedish University of Agricultural Sciences, The Department of Forest Resource Management and Geomatics. Acta Universitatis Agriculturae Sueciae. Silvestria, 17.

Ohmann, J. L., & Gregory, M. J. (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.

Pike, R. J., & Wilson, S. E. (1971). Elevation relief ratio, hypsometric integral, and geomorphic area altitude analysis. *Geological Society of America Bulletin*, *82*, 1079−1084.

Pocewicz, A. L., Gessler, P. E., & Robinson, A. P. (2004). The relationship between effective plant area index and Landsat spectral response across elevation, solar insolation, and spatial scales, in a northern Idaho forest. *Canadian Journal of Forest Research*, *34*, 465−480.

Reutebuch, S. E., Anderson, H. -E., & McGaughey, R. J. (2005, September). Light detection and ranging (LIDAR): An emerging tool for multiple resource inventory. *Journal of Forestry*, 286−292.

Roberts, D. W., & Cooper, S. V. (1989). Concepts and techniques of vegetation mapping. *Land classifications based on vegetation: Applications for*

resource management. GTR-INT-257 (pp. 90−96). Ogden, UT: USDA Forest Service Intermountain Research Station.

Stage, A. R. (1976). An expression of the effects of aspect, slope, and habitat type on tree growth. *Forest Science*, *22*, 457−460.

Stage, A. R., & Crookston, N. L. (2007). Partitioning error components for accuracy-assessment of near-neighbor methods of imputation. *Forest Science*, *53*, 62−72.

Temesgen, H., LeMay, V. M., Froese, K. L., & Marshall, P. L. (2003). Imputing tree-lists from aerial attributes for complex stands of south-eastern British Columbia. *Forest Ecology and Management*, *177*, 277−285.

Tomppo, E. (1991). Satellite image-based national forest inventory of Finland. *International Archives of Photogrammetry and Remote Sensing*, *28*, 419−424.

Tomppo, E., & Halme, M. (2004). Using coarse scale forest variables as ancillary information and weighting of variables in k-NN estimation — A genetic algorithm approach. *Remote Sensing of Environment*, *92*, 1−20.

Tomppo, E., Nilsson, M., Rosengren, M., Aalto, P., & Kennedy, P. (2002). Simultaneous use of Landsat-TM and IRS-1C WiFS data in estimating large area tree stem volume and aboveground biomass. *Remote Sensing of Environment*, *82*, 156−171.

Tuominen, S., & Pekkarinen, A. (2004). Local radiometric correction of digital aerial photographs for multi source forest inventory. *Remote Sensing of Environment*, *89*, 72−82.