Landscape-scale parameterization of a tree-level forest growth model: a *k*-nearest neighbor imputation approach incorporating LiDAR data

Michael J. Falkowski
Michigan Technological University, mjfalkow@mtu.edu

Andrew T. Hudak
Rocky Mountain Research Station, ahudak@fs.fed.us

Nicholas L. Crookston
Rocky Mountain Research Station, ncrookston@fs.fed.us

Paul E. Gessler
University of Idaho, paulg@uidaho.edu

Edward H. Uebler
USDA Forest Service, euebler@fs.fed.us

See next page for additional authors

Authors
Michael J. Falkowski, Andrew T. Hudak, Nicholas L. Crookston, Paul E. Gessler, Edward H. Uebler, and Alistair M.S. Smith

Landscape-scale parameterization of a tree-level forest growth model: a *k*-nearest neighbor imputation approach incorporating LiDAR data

Michael J. Falkowski, Andrew T. Hudak, Nicholas L. Crookston, Paul E. Gessler, Edward H. Uebler, and Alistair M.S. Smith

Abstract: Sustainable forest management requires timely, detailed forest inventory data across large areas, which is difficult to obtain via traditional forest inventory techniques. This study evaluated *k*-nearest neighbor imputation models incorporating LiDAR data to predict tree-level inventory data (individual tree height, diameter at breast height, and species) across a 12 100 ha study area in northeastern Oregon, USA. The primary objective was to provide spatially explicit data to parameterize the Forest Vegetation Simulator, a tree-level forest growth model. The final imputation model utilized LiDAR-derived height measurements and topographic variables to spatially predict tree-level forest inventory data. When compared with an independent data set, the accuracy of forest inventory metrics was high; the root mean square difference of imputed basal area and stem volume estimates were 5 $m^2 \cdot ha^{-1}$ and 16 $m^3 \cdot ha^{-1}$, respectively. However, the error of imputed forest inventory metrics incorporating small trees (e.g., quadratic mean diameter, tree density) was considerably higher. Forest Vegetation Simulator growth projections based upon imputed forest inventory data follow trends similar to growth projections based upon independent inventory data. This study represents a significant improvement in our capabilities to predict detailed, tree-level forest inventory data across large areas, which could ultimately lead to more informed forest management practices and policies.

Résumé : L'aménagement durable des forêts requiert des données appropriées et détaillées d'inventaire forestier sur de grandes superficies, ce qui est difficile à obtenir par le biais de techniques traditionnelles d'inventaire forestier. Cette étude évalue des modèles d'imputation basés sur les *k* plus proches voisins incorporant des données lidar pour prédire des mesures d'inventaire à l'échelle de l'arbre (hauteur, diamètre à hauteur de poitrine et espèce des arbres individuels) dans une aire d'étude de 12 100 ha du nord-est de l'Oregon, aux États-Unis. L'objectif premier est de fournir des données spatialement explicites pour paramétrer un modèle de croissance forestière à l'échelle de l'arbre, le «Forest Vegetation Simulator». Le modèle final d'imputation utilise des mesures de hauteur et des variables topographiques dérivées du lidar pour prédire spatialement des données d'inventaire forestier à l'échelle de l'arbre. Lorsqu'elles ont été comparées à un fichier indépendant de données, la précision des mesures d'inventaire forestier était élevée: l'erreur quadratique moyenne des estimations imputées de surface terrière et de volume étaient respectivement de 5 $m^2 \cdot ha^{-1}$ et 16 $m^3 \cdot ha^{-1}$. Cependant, l'erreur des mesures imputées d'inventaire forestier qui tiennent compte des petits arbres (p. ex. le diamètre moyen quadratique et la densité des arbres) était considérablement plus élevée. Les projections de croissance du «Forest Vegetation Simulator» basées sur des données imputées d'inventaire forestier suivent une tendance similaire aux projections basées sur des données indépendantes d'inventaire. Cette étude représente une amélioration importante de nos capacités à prédire des données détaillées d'inventaire forestier à l'échelle de l'arbre sur de grandes superficies, ce qui pourrait éventuellement mener à des pratiques et des politiques d'aménagement forestier mieux fondées.

[Traduit par la Rédaction]

Introduction

Over the last few decades, increasing concerns over the potential impacts of climate change, biodiversity loss, and large-scale disturbances, such as insects and wildland fire, coupled with forest commodity needs, have created a need for land managers to efficiently and precisely quantify multiple resources in forested ecosystems (Lund 2004). To effectively manage forested ecosystems in a sustainable manner, the condition of forested ecosystems must be characterized and monitored across multiple spatial extents (e.g., stand, watershed, region). In an ideal situation, land manag-

Received 1 May 2009. Accepted 2 November 2009. Published on the NRC Research Press Web site at cjfr.nrc.ca on 29 January 2010.

M.J. Falkowski.[1] School of Forest Resources and Environmental Science, Michigan Technological University, Houghton, MI 49931, USA.
A.T. Hudak and N.L. Crookston. USDA Forest Service, Rocky Mountain Research Station, Forestry Sciences Laboratory, 1221 South Main Street, Moscow, ID 83843, USA.
P.E. Gessler and A.M.S. Smith. Department of Forest Resources, University of Idaho, Moscow, ID 83843, USA.
E.H. Uebler. USDA Forest Service, Malheur National Forest, John Day, OR 97845, USA.

[1] Corresponding author (e-mail: mjfalkow@mtu.edu).

doi:10.1139/X09-183

ers would possess detailed forest inventory data quantifying the size, species, and condition of every tree within every management unit across an entire forest (Temesgen et al. 2003). Such tree-level information could be summarized and analyzed to characterize forest status and condition across any spatial extent. In addition, tree-level forest inventory data could be used to parameterize individual tree-based forest growth models, such as the Forest Vegetation Simulator (FVS; Crookston and Dixon 2005), so that the future status of forested ecosystems can be projected for the purpose of forest planning. Such an approach would provide a means to evaluate the efficacy and ecological impacts of alternative management decisions across multiple spatial and temporal extents.

Directly measuring every tree over large areas is not practical given time and funding constraints, thus sampling theory is employed to estimate forest composition and structure across large spatial extents (Kohl 2004). *k*-Nearest Neighbor (*k*-NN) imputation is one approach commonly used to extrapolate forest inventory data collected at discrete sampling locations to larger areas (Barrett and Fried 2004; McRoberts et al. 2007). In their simplest form, *k*-NN imputation algorithms assign forest inventory data collected at discrete sampling locations to unsampled areas based upon the statistical similarity or statistical distance (e.g., Euclidian distance, Mahalanobis distance) between sampled and unsampled areas, where statistical similarity is determined based upon covariates available across the entire area of interest (e.g., remotely sensed data). *k*-NN imputation approaches incorporating remotely sensed data have been used to predict timber volume (Mäkelä and Pekkarinen 2004), basal area (Franco-Lopez et al. 2001; LeMay and Temesgen 2005), tree density (LeMay and Temesgen 2005), and timber yield (Maltamo and Eerikainen 2001) across large spatial extents. Typically, *k*-NN imputation relies on medium resolution satellite data (e.g., Landsat data) for predicting forest characteristics. However, recent research has demonstrated that incorporating predictor variables from high resolution remotely sensed data improves estimates of forest characteristics. For example, Tuominen and Pekkarinen (2005) demonstrated that including texture features derived from high resolution aerial photographs in a *k*-NN algorithm reduced stem volume error estimates by 26%, while Maltamo et al. (2006) demonstrated that a *k*-NN algorithm including both LiDAR data and textural information derived from aerial photography reduced the error of stem volume predictions by 23% in comparison with estimates attained when using aerial photography texture alone. In a recent study, Hudak et al. (2008) demonstrated that a *k*-NN imputation approach incorporating a suite of LiDAR-derived metrics produced reasonable ($r = 0.80$) species-wise predictions of basal area across a mixed conifer forest in north Idaho, USA.

Although imputation has been effective for predicting stand-level forest inventory metrics, to date predicting tree-level forest inventory data via imputation has been limited to a few studies. Fehrmann et al. (2008) employed *k*-NN imputation to produce nonspatial predictions of individual tree characteristics, such as single tree biomass, while a recent study conducted by Wallerman and Holmgren (2007) predicted plot-level forest inventory data to unsampled areas via an imputation model incorporating LiDAR and optical satellite data. Other studies have used imputation to predict a suite of stand summary statistics (e.g., stems per hectare by species, basal area, and volume) that can be used to parameterize the FVS forest growth model (e.g., Temesgen et al. 2003). However, there is still a need to further develop methods to predict tree-level forest inventory data, with a specific focus on parameterizing forest growth models, such as FVS. The continual development of LiDAR remote-sensing technology coupled with recent advances in *k*-NN imputation algorithms could improve the accuracy of spatially explicit predictions of tree-level forest inventory data across large spatial extents.

Research objective and hypotheses

This paper presents a novel methodology for predicting tree-level forest inventory data across a 12 100 ha study area in northeastern Oregon, USA. Specifically, we employ a recently developed imputation approach (randomForest imputation; Crookston and Finley (2008)) incorporating LiDAR-derived predictor variables to generate 'virtual' forest inventory data across the entire study area. The primary impetus of this study is to parameterize the FVS model with the imputed forest inventory data so that forest growth can be modeled across the entire study area. Upon completion of the current study, the imputed tree-level forest inventory data will aid in basic forest management decision making. In addition to parameterizing FVS, the imputed data could be used for many purposes, including forest commodity assessment, carbon accounting, and wildlife habitat modeling, among others.

We evaluate four multivariate imputation models relating plot-level forest structure (basal area and basal area-weighted tree diameter) and species composition (forest type) to a suite of LiDAR height metrics and digital elevation model (DEM) variables (Table 1). Since the LiDAR height metrics characterize current forest structure and the DEM variables characterize biophysical gradients that potentially influence forest species composition, an imputation model incorporating these variables should produce accurate predictions of both forest species composition and forest structure. After developing the initial plot-level imputation models, we apply them to unsampled areas to generate a 'virtual' forest inventory data set consisting of tree-level forest inventory data in a format that can be used to parameterize the FVS forest growth model. The following null research hypotheses are tested to determine if the virtual forest inventory data are equivalent to coincident field-based forest inventory data:

> $H_0$1: Forest inventory metrics from the virtual forest inventory are not significantly different from forest inventory metrics from a coincident, independent forest inventory.
>
> $H_0$2: The sampling error of the virtual forest inventory is not significantly different from the sampling error of a coincident, independent forest inventory.
>
> $H_0$3: Species composition of the virtual forest inventory is not significantly different from species composition measured in a coincident, independent forest inventory.

Table 1. LiDAR metrics used as auxiliary variables in the imputation model.

Metric name	Metric description
HMIN	Minimum height
HMAX	Maximum height
HRANGE	Range of heights
HMEAN*	Mean height
HMEDIAN	Median height
HMODE	Modal height
NMODES	Number of modes
HSTD	Standard deviation of heights
HVAR	Variance of heights
HSKEW	Skewness of heights
HKURT	Kurtosis of heights
HCV	Coefficient of variation of heights
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
CANOPY*	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
STRATUM2*,†	Percentage of vegetation returns >1 m and ≤2.5 m
STRATUM3†	Percentage of vegetation returns >2.5 m and ≤10 m
STRATUM4*	Percentage of vegetation returns >10 m and ≤20 m
STRATUM5†	Percentage of vegetation returns >20 m and ≤30 m
STRATUM6	Percentage of vegetation returns >30 m
TEXTURE	Standard deviation of non-ground returns >0 m and ≤1 m
INSOL	Solar insolation ($W \cdot m^{-2}$)
ELEVATION‡	Elevation (m)
SLPPCT‡	Slope (%)
ASPECT	Aspect (°)
TRASP	Transformed aspect (Roberts and Cooper 1989)
SCOSA	Percent slope × cos(aspect) transformation (Stage 1976)
SSINA	Percent slope × sin(aspect) transformation (Stage 1976)
FLOWD	Flow distance to stream (Tarboton 1997)
CTI‡	Compound topographic index (Tarboton 1997)

*Selected as an important variable for imputing basal area.

†Selected as an important variable for imputing basal area weighted DBH.

‡Selected as an important variable for imputing forest species composition.

Background information

The forest vegetation simulator

In the United States, the FVS is a widely applied forest growth model used to aid in forest management decision making (Dixon 2003). The FVS is an empirically driven model that operates at the individual tree level, providing summary statistics of initial stand conditions as well as stand-level projections of future forest growth and conditions (Crookston and Dixon 2005). The FVS has the capability to model growth across a wide array of forest species compositions and structures (i.e., single to mixed species, even-aged to uneven-aged stands and single- to multi-story stands; Dixon 2003). In addition, through the use of model variants (currently 22 unique variants exist) the FVS can be used to predict growth across many forest types in the United States and Canada (Crookston and Dixon 2005).

The FVS is parameterized with standard tree-level forest inventory data consisting of tree-lists quantifying required (species and diameter at breast height) and optional state variables (tree count, diameter growth, height, height growth, and crown ratio) for each tree within a plot, stand, or other management unit (Dixon 2003). The model can also incorporate information quantifying the slope, elevation, aspect, and site potential or habitat type at the sample point, plot, or stand levels (Crookston and Dixon 2005). After the input data are read, the model performs a self-calibration procedure during which its internal growth models are adjusted to mimic growth rates apparent within the inventory data when provided by the user (Crookston and Dixon 2005). Following calibration, the FVS provides summary statistics of initial stand conditions and then projects forest growth and other dynamics into the future. As the model runs, forest growth and yield projections are adjusted to account for user-specified forest management activities, such as harvesting. Growth projections can also be adjusted to account for tree mortality, tree regeneration, and fire, as well as the impact of parasites and pathogens if so desired.

Although the FVS is primarily used to forecast forest growth and yield based on different silvicultural treatments, it has also been used to evaluate trends in wildlife habitat quality (Eng 1997; Wilson 1997; Maffei and Tandy 2002), to assess the impacts of forest policy on future forests (Cousar et al. 1997), to evaluate fire hazard and potential fire behavior (Fulé et al. 2004), and to gauge future forest conditions (Atkins and Lundberg 2002). Despite the diversity of applications, there are limitations to the FVS model. For example, since it is empirically driven, predictions of future forest conditions are only valid if future climate conditions do not deviate far from current conditions. In addition, because the FVS is not a process-based model, its efficacy for evaluating forest growth under dynamic rates of physiological or biogeochemical processes is limited. However, current work is underway to incorporate the influence of future climate projections into FVS growth predictions (Crookston and Dixon 2005). In addition, research conducted by Milner et al. (2003) demonstrated that the FVS could be coupled with a process-based model (Stand-Bio-Geochemical Cycles; Milner and Coble 1995) to simulate biogeochemical and physiological influences on forest growth and yield.

The FVS could potentially be used to predict future forest conditions across large spatial extents. However, as previously mentioned, the FVS model requires tree-level forest inventory data for parameterization, which are difficult to collect continuously across large areas. As a result, the use of the FVS has typically been limited to stand-level or multistand-level studies. However, the continual development of imputation algorithms and remote-sensing technology that precisely characterize the vertical and horizontal structure of vegetation (e.g., LiDAR and RADAR) may provide a means to obtain spatially continuous predictions of tree-level forest inventory data across entire landscapes.

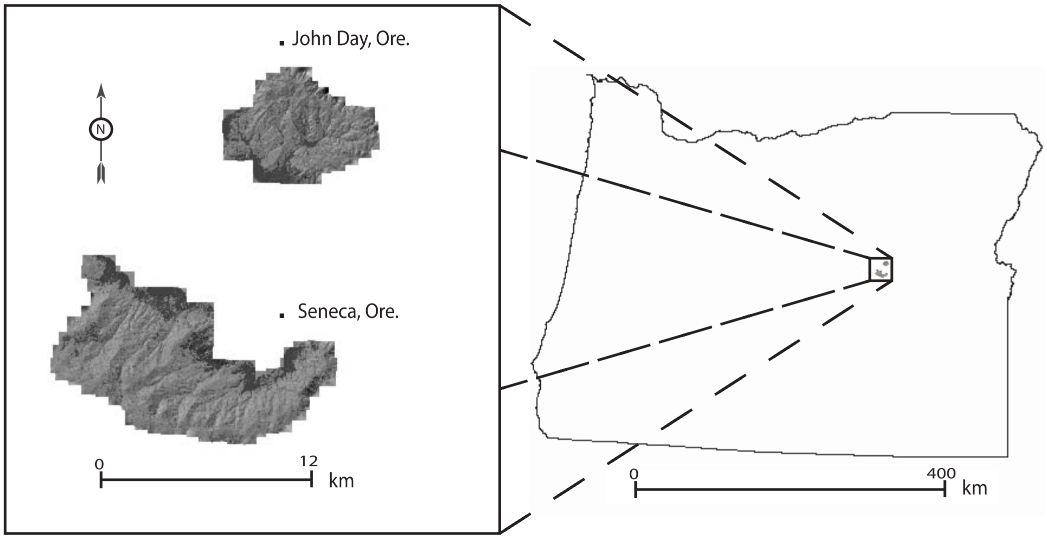
k-NN imputation

In forest inventory and assessment, *k*-NN imputation is typically used to predict forest inventory attributes in uninventoried areas based upon a two-phased sampling design. In the first phase, ‘auxiliary variables’ that can be easily measured across the entire landscape of interest are obtained (e.g., remotely sensed data). The second phase involves a detailed inventory of ‘variables of interest’ (e.g., forest inventory data) at discrete sampling locations within the study area (Moeur and Stage 1995). This procedure produces two separate data sets; a reference data set containing both auxiliary variables and variables of interest measured at each sampling location, and a target data set composed only of auxiliary variables measured across the entire population of interest. The goal of imputation is to predict the variables of interest (i.e., forest inventory data) in unsampled areas. To achieve this, the reference data are used as a training set to characterize the relationship between the auxiliary variables and the variables of interest. Missing attributes within the target data set are then estimated by imputing them from the nearest neighbors within the reference data set, where nearness is measured in terms of the statistical similarity or distance (e.g., Euclidian distance, Mahalanobis distance) between auxiliary variables in the reference and target data sets. When $k = 1$, the missing target value is simply taken from the nearest neighbor in the reference data set, and when k is >1 other methods, such as a weighted average, are used to calculate the values of a missing observation from the *k*-selected neighbors (Crookston and Finley 2008).

k-NN imputation via the random forest proximity matrix

Many different approaches have been developed to quantify the statistical distance between target and reference observations. Typically, these approaches determine the distance between observations based on the Euclidian distance (or weighted variants of the Euclidian distance) between reference and target observations (Crookston and Finley 2008). Although any distance metric could be used, the current study employs a novel *k*-NN imputation distance metric that quantifies the statistical distance between reference and target observations based on a proximity matrix calculated via the randomForest (RF) classification and regression tree algorithm. This distance metric was selected because it has produced reliable predictions of stand-level basal area and tree density across a similar study area in the Inland Northwest, USA (Hudak et al. 2008). Although the RF imputation method has recently been explained in detail by Crookston and Finley (2008) as well as by Hudak et al. (2008), the novelty of this approach warrants a brief review of its functional approach. In its native form, the RF algorithm develops classification or prediction rules by growing an ensemble (>100 to >1000) of classification or regression trees from random subsets of training data, while randomly permuting independent variables at each node (see Breiman (2001), Prasad et al. (2006), and Lawrence et al. (2006) for detailed descriptions of the RF algorithm). In addition to predicting or classifying new observations, the RF algorithm calculates the proximity of every observation by classifying each observation via each tree within the ensemble. The proximity of a pair of observations is increased by one every time they end up in the same terminal node after classification. The final proximity values are divided by the total number of trees in the ensemble to calculate the overall proximity between each observation. For example, if a pair of observations ends up in the same terminal node 75 times, and there are 100 total trees in the ensemble, the proximity of these two observations equals 0.75. Subtracting one from the final proximity values is analogous to calculating the statistical distance between each observation in the data set; a high proximity equals a small statistical distance and vice versa (Breiman 2001; Crookston and Finley 2008). Crookston and Finley (2008) developed a *k*-NN imputation approach that incorporates the RF proximity matrix when searching for similar neighbors. To facilitate *k*-NN imputation a few modifications have been made to the original RF algorithm. One important modification extends the RF algorithm to allow for multivariate imputation (i.e., to impute the values of multiple response variables simultaneously). This is achieved by growing a separate ensemble of trees for each response variable in the model. The final proximity (i.e., statistical distance) of each observation is calculated by joining the proximity matrices from each ensemble of trees.

Fig. 1. Damon study area – LiDAR canopy cover, hill shade composite.



Furthermore, the algorithm also allows the user to grow each tree ensemble with a different number of trees and (or) with a unique set of predictor variables. This functionality is useful when certain response variables are more important than others (i.e., should be weighted higher in the imputation) or when specific predictor variables explain variation in one response variable but not in others (Crookston and Finley 2008).

Methods

Study area

This study was conducted within the Damon study area (~12 100 ha), which is within the Shirttail and Van Aspen subwatersheds of the Blue Mountain Ranger District in the Malheur National Forest near Seneca, Oregon, USA (44.14°N, –118.97°W; Fig. 1). The area is an uplifted plain composed of sedimentary and metasedimentary rocks covered by volcanic ash soils originating from the Mount Mazama, eruption which occurred around 7677 ± 150 years BP (Zdanowicz, et al. 1999). Aspects vary across the region, but in general range from north to east and from southeast to southwest across the northern and southern portions of the study area, respectively. Slopes are primarily less than 30%; however, slopes do reach 50% in a few areas. Precipitation ranges from 40 to 65 cm per year, primarily occurring as snow in the winter months. In general, forest stands are dominated by *Pinus ponderosa* var. *scopulorum*, as a result of historic fire regimes. However, *Abies grandis* var. *idahoensis* and *Pseudotsuga menziesii* var. *glauca* also occur on north-facing aspects and in higher elevations as well as in areas that historically favored the removal of large *P. ponderosa* via logging. Minor amounts of *Pinus contorta* var. *latifolia* and *Larix occendentalis* Nutt. can be found throughout the study area, while *Populus tremuloides* Michx. occurs throughout riparian areas across the Damon study area.

Sampling design and data collection

Forest inventory data were collected within 88 forest stands encompassing the full range of forest structure and species composition present across the Damon study area. Two separate forest inventories were conducted within each of the 88 stands: a variable-radius plot inventory and a fixed-radius plot inventory. For the variable-radius plot inventory, a total of 641 variable-radius plots were inventoried following a systematic sampling design, where the spacing between plots and total number of plots within each stand was dependent upon stand density and stand variability. The variable-radius plot inventory was designed to produce an estimate of total stand basal area within 20% of one standard deviation in each stand. Once located, a relaskop was used to determine which trees were within the variable-radius plot boundary. Diameter at breast height (DBH), height, species, crown ratio, and other standard inventory metrics were recorded for every tree or snag selected by the relaskop. Seedlings and saplings were also tallied and measured within each plot on a smaller fixed-radius plot; elevation, slope, aspect, habitat type, and forest type were recorded at every plot within every stand. For the fixed-radius plot inventory, one 0.04 ha fixed-radius inventory plot was installed within each of the 88 stands (i.e., a total of 88 fixed-radius plots). The location of the fixed-radius plot was randomly selected from the pool of variable-radius plots within each stand. The aforementioned inventory metrics were recorded within each of the 88 fixed-radius plots. Seedlings and saplings were also tallied and measured on a subplot within each fixed-radius plot. During the inventory, a Trimble GeoXT global positioning system was used in conjunction with a differential correction procedure to accurately measure the position of each forest inventory plot. For the purpose of this study, the fixed-radius plot inventory is used as a reference data set for imputation model development and tree-level forest inventory data prediction, while the variable-radius plot inventory data set is used as a validation data set to evaluate the accuracy of the imputed (i.e., 'virtual') forest inventory data.

LiDAR acquisition and processing

Discrete return LiDAR data were acquired 15–16 September 2007 across the Damon study area by Watershed Sciences (Corvallis, Oregon). The sensor operated at 1084 nm.

Fig. 2. *k*-Nearest neighbor virtual forest inventory schematic. An imputation model relating stand-level forest inventory data (e.g., basal area and forest type) to LiDAR metrics (Table 1) is developed. New virtual forest inventory plots (represented by black circles) are systematically located across a stand. LiDAR metrics are calculated within each new virtual inventory plot and are related to the nearest neighbor (in terms of the LiDAR metrics) within the reference data set via the imputation model. The tree records from this nearest neighbor are used as surrogate forest inventory data in the virtual forest inventory plot.

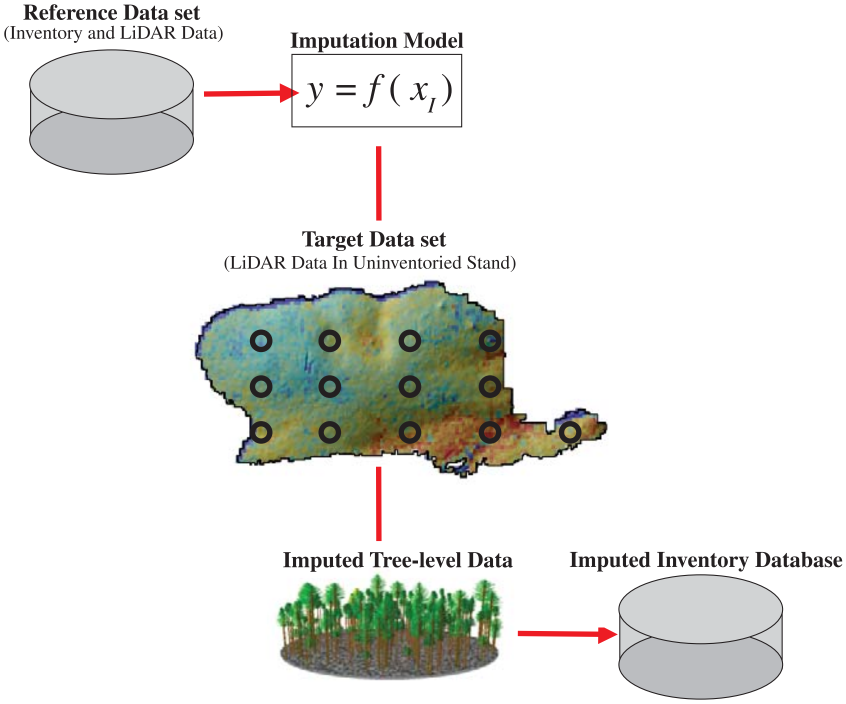

Reference Data set
(Inventory and LiDAR Data)
Imputation Model
$y = f(x_I)$
Target Data set
(LiDAR Data In Uninventoried Stand)
Imputed Tree-level Data
Imputed Inventory Database

Table 2. Imputation models and associated variables evaluated in this study.

Model	Y_1	Y_2	Y_3	Independent variables (X)
1	BA	Forest type	BA weighted DBH	LiDAR metrics for associated Y
2	BA	Forest type		LiDAR metrics for associated Y
3	BA	BA weighted DBH		LiDAR metrics for associated Y
4	BA			LiDAR metrics for associated Y

Note: BA, basal area.

The acquired LiDAR data had an average pulse density of 6.31 points·m^{-2} and an absolute vertical accuracy of 0.024 m. Once acquired, the raw LiDAR data points were classified as ground or non-ground returns using the Multiscale Curvature Classification algorithm (Evans and Hudak 2007). Following classification, a high resolution (1 m) DEM was interpolated from the ground returns, and the height above ground surface was calculated for all non-ground returns through DEM subtraction. Following processing, a variety of LiDAR-based height and topographic metrics (Table 1) that have proved useful for characterizing forest structure and species composition (Hudak et al. 2006, 2008) were calculated from the LiDAR returns coincident with each forest inventory plot within the reference and target forest inventory data sets.

Data analysis

Imputation model development

Imputation model development was conducted within the R statistical software program (R Development Core Team 2005) via the yaImpute R package (Crookston and Finley 2008). Prior to developing the imputation model, a variable selection procedure was employed to select the optimal LiDAR variables to use in the final imputation models. This process, which is implemented via the varSelRF R package (Diaz-Uriarte 2007), selects important predictor variables through an iterative, backwards variable elimination process designed to minimize the RF out-of-bag error rate without creating bias in the final model. Three separate variable selection procedures were run on the reference data set: one to select the optimal LiDAR variables for predicting plot-level basal area, another to select variables for basal area weighted tree diameter prediction, and another for selecting the best variables for predicting forest species composition.

Once the important variables were selected, a three-step process was employed to impute tree-level forest inventory via a *k*-NN imputation approach. First, a multivariate imputation model relating plot-level forest structure and species composition (*Y* variables; basal area, basal area weighted tree diameter, and forest species composition) to the selected plot-level LiDAR metrics (*X* variables; Table 1) was developed from the reference data set. Since there were three *Y* variables, three separate tree ensembles were grown within

Table 3. Evaluation statistics and equivalence tests for imputation models.

Inventory metric	*r*	RMSD	EI slope (%)	EI intercept (%)
Model 1 (Y_{BA}, Y_{wtDBH}, Y_{ForTyp} = $f(X_{\mathrm{BA}}$, X_{wtDBH}, $X_{\mathrm{ForTyp}})$)				
BA ($m^2 \cdot ha^{-1}$)	0.82	5.51	20.43	13.87*
Total volume ($m^3 \cdot ha^{-1}$)	0.87	15.59	21.04	8.75*
Stand density index	0.83	136.58	16.20*	16.45*
Tree density				
Total	0.14	1830.82	105.33	24.89
>53 cm	0.70	11.31	64.21	44.01
40–53 cm	0.45	15.20	69.19	56.33
23–40 cm	0.62	63.49	42.63	19.29
18–23 cm	0.36	49.17	73.10	35.09
8–18 cm	0.49	170.23	71.45	58.81
0–8 cm	0.09	1824.93	108.15	24.49
QMD (cm)	0.41	6.79	81.93	15.44*
Weighted DBH (cm)	0.65	9.21	56.51	5.47*
Overstory DBH (cm)	0.69	7.95	52.33	6.31*
BA sampling error	0.23	2.07	96.91	25.41
	Accuracy	**Kappa**		
Forest type	64.07%	23.25		
Model 2 (Y_{BA}, Y_{ForTyp} = $f(X_{\mathrm{BA}}$, $X_{\mathrm{ForTyp}})$)				
BA ($m^2 \cdot ha^{-1}$)	0.78	6.24	18.84*	18.70*
Volume ($m^3 \cdot ha^{-1}$)	0.77	20.49	26.44	17.19*
Stand density index	0.80	153.95	18.70*	20.51
Tree density				
Total	0.18	1526.90	103.04	32.19
>53 cm	0.12	41.78	104.40	355.55
40–53 cm	0.01	42.78	112.03	171.26
23–40 cm	0.04	87.03	133.38	17.39*
18–23 cm	0.25	81.83	142.20	28.87
8–18 cm	0.10	276.56	131.03	58.40
0–8 cm	0.05	1393.29	128.12	77.78
QMD (cm)	0.27	7.34	94.11	14.6*
Weighted DBH (cm)	0.17	17.29	103.65	33.57
Overstory DBH (cm)	0.18	19.26	102.90	42.26
BA sampling error	0.08	2.28	114.89	12.5*
	Accuracy	**Kappa**		
Forest type	66.67%	27.49		
Model 3 (Y_{BA}, Y_{wtDBH} = $f(X_{\mathrm{BA}}$, $X_{\mathrm{wtDBH}})$)				
Basal area ($m^2 \cdot ha^{-1}$)	0.85	5.14	16.67*	13.97*
Volume ($m^3 \cdot ha^{-1}$)	0.86	15.99	23.68	7.37*
Stand density index	0.87	118.76	14.39*	13.69*
Tree density				
Total	0.29	1412.73	91.60	18.83*
>53 cm	0.74	11.11	62.09	55.08
40–53 cm	0.40	15.86	70.28	54.26
23–40 cm	0.65	61.62	41.57	16.7*
18–23 cm	0.47	46.16	51.76	36.00
8–18 cm	0.46	172.90	56.12	60.24
0–8 cm	0.21	1436.03	99.78	21.30
QMD (cm)	0.32	7.14	91.57	8.24*
Weighted DBH (cm)	0.70	8.27	48.46	6.48*
Overstory DBH (cm)	0.68	7.85	51.18	6.79*
BA sampling error	0.15	2.13	107.71	19.7*
	Accuracy	**Kappa**		
Forest type	69.23%	12.89		

Table 3 (*concluded*).

Inventory metric	*r*	RMSD	EI slope (%)	EI intercept (%)
Model 4 ($Y_{BA} = f(X_{BA})$)				
Basal area ($m^2 \cdot ha^{-1}$)	0.88	4.76	17.21*	13.67*
Volume ($m^3 \cdot ha^{-1}$)	0.89	13.85	18.74*	8.28*
Stand density index	0.88	118.16	16.05*	14.78*
Tree density				
Total	0.23	1330.26	100.21	31.43
>53 cm	0.62	10.43	64.34	50.77
40–53 cm	0.55	14.59	50.28	52.78
23–40 cm	0.64	62.77	44.90	14.33*
18–23 cm	0.39	49.13	65.26	41.18
8–18 cm	0.38	177.50	61.71	55.90
0–8 cm	0.09	1329.35	114.46	33.86
QMD (cm)	0.23	6.92	97.92	8.7*
Weighted DBH (cm)	0.59	9.51	59.72	5.7*
Overstory DBH (cm)	0.63	8.39	55.71	7.87*
BA sampling error	0.17	2.12	104.84	17.02*
	Accuracy	**Kappa**		
Forest type	68.18%	31.06		

Note: EI is the interval at which the metrics become equivalent. RMSD, root mean squared difference; BA, basal area; QMD, quadratic mean diameter; DBH, diameter at breast height.

*Statistically equivalent.

the yaImpute packages RF imputation mode. Each ensemble consisted of 3000 bootstrap replicates (i.e., classification and regression trees). Furthermore, only the LiDAR metrics that were selected as being important for a particular *Y* variable were used to generate the tree ensemble for that *Y* variable. For the second step, the final imputation model was applied to the target data set (i.e., the variable-radius plot forest inventory data set). In addition to imputing plot-level *Y* variables to each variable-radius plot location, this step determined which forest inventory plot in the reference data set is closest, in terms of statistical distance, to each variable-radius plot location in the target data set. Finally, tree-level inventory data from the reference data set were used as surrogate tree-level forest inventory for the closest plot (in terms of the statistical distance) in the target data set (Fig. 2). This process produced a virtual forest inventory data set for each of the 88 stands surveyed via the variable-radius plot inventory. Four separate imputation models were evaluated for imputing tree-level forest inventory data following the three-step process outlined above: a full imputation model and three reduced imputation models (Table 2).

Model evaluation and hypothesis testing

The accuracy of the imputed tree-level forest inventory data was determined through a comparison with forest inventory data measured during the validation inventory. Numerous stand-level forest inventory metrics (e.g., basal area, total volume, tree density by DBH class, quadratic mean diameter), as well as stand-level sampling error for basal area, were calculated from both the imputed and validation forest inventory data sets. These metrics were compared via Pearson's correlation coefficients (*r*) and root mean squared difference (RMSD; Stage and Crookston 2007). In addition, the first two hypotheses ($H_0$1 and $H_0$2; equivalence of forest inventory metrics and sampling errors) were tested via statistical equivalence tests, which were used to test the null hypothesis of no significant difference between the two forest inventory data sets. Specifically, a regression-based equivalence test (Robinson et al. 2005) was employed to test for intercept equality (i.e., the mean of imputed forest inventory metrics are equal to the mean of validation forest inventory metrics across the entire population) and for slope equality to 1 (i.e., if the pairwise (between-stand) forest inventory metrics are equal, the regression will have a slope of 1). The region of equivalence was set to ±20% (of the mean) for the intercept (b_0) and to ±20% for the slope (b_1). The null hypothesis of dissimilarity between the imputed and validation inventory metrics was rejected if the interval of equivalence (±20%) contained two joint one-sided 97.25% confidence intervals ($\alpha = 0.05$) for the slope or intercept. The accuracy of forest species composition from the imputed data ($H_0$3) was determined via the overall accuracy and Cohen's Kappa statistics (Cohen 1960; Congalton and Green 1999).

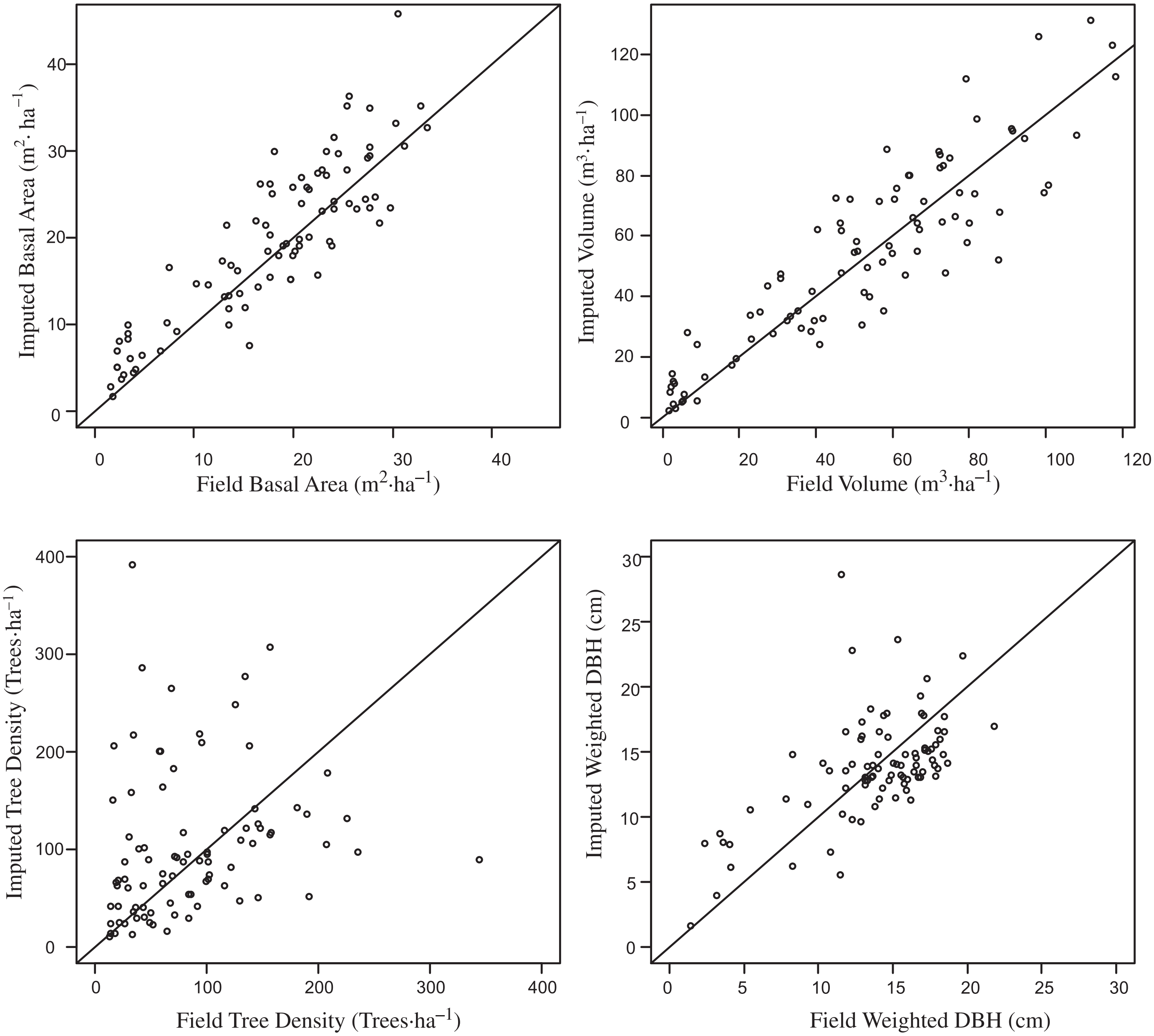
FVS growth projection comparison

To further assess the performance of the virtual forest inventory data, the FVS was parameterized with the forest inventory data from the best-performing imputation model. Forest growth was then projected in 10-year increments for 90 years with data from the selected imputation model as well as with data from the validation forest inventory. To determine if the growth projections from both data sets followed similar trends, basal area projections were compared within each of the 88 stands via *r*, RMSD, and equivalence tests.

Landscape-level prediction and growth projection

The best-performing imputation model was also employed to predict tree-level forest inventory data across the entire

Fig. 3. Scatter plots of imputed metrics (from Model 4) versus field forest inventory metrics.



Damon study area. To achieve this, the LiDAR point data were summarized in 20 m bins (i.e., grid cells) across the study area, and the best-performing imputation model was then applied to each grid cell. Following this process, every 20 m grid cell within the Damon study area contained tree-level forest inventory data, which could be used to estimate various forest inventory metrics. The FVS was also parameterized to spatially predict forest growth within each 20 m grid cell across the entire Damon study area.

Results

Variable importance

The final imputation models were developed from 10 of the 40 original candidate LiDAR metrics (Table 1). The variable selection procedure employed herein rated canopy cover, height of the 95th percentile, mean height, and pulse density within strata two and four as the most important variables for imputing basal area. The pulse density within strata two, three, and five, as well as the height of the 95th percentile were selected as important variables for the basal area weighted DBH metric. Forest species composition was best explained by three LiDAR DEM metrics: the compound topographic index, elevation, and percent slope (Table 1).

Imputation model accuracy and statistical equivalence

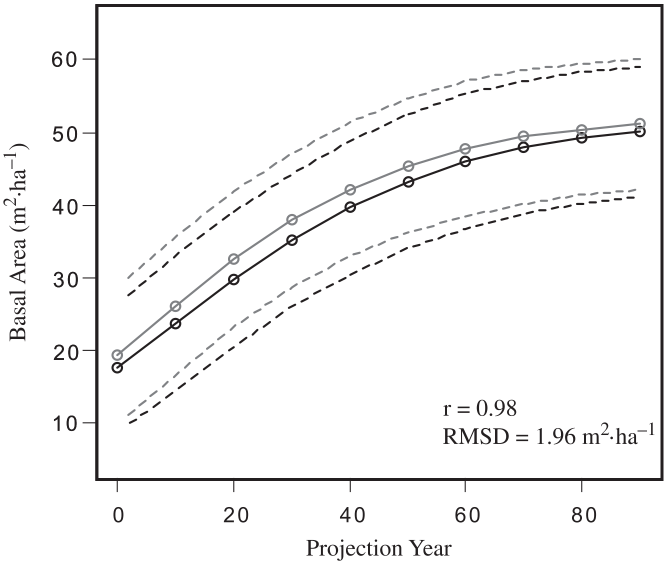
The full imputation model (Model 1; Table 3) produced estimates of basal area, total volume, and stand density index (SDI) that were strongly correlated ($r > 0.8$) with the validation inventory metrics, whereas the density of trees >53 cm DBH and between 23 and 40 cm (DBH), basal area weighted DBH, and the DBH of overstory trees exhibited moderate correlations ($r > 0.5$). All other inventory metrics were weakly correlated ($r < 0.5$) with the validation forest inventory metrics. In terms of statistical equivalence of the means, basal area, total volume, SDI, quadratic mean diameter (QMD), weighted DBH, and the DBH of overstory trees were statistically equivalent to the validation inventory data. However, except for SDI, the pairwise equivalence test (i.e., slope equivalence to 1) indicated that none of the inventory metrics were equivalent at the ±20% equivalence level. Furthermore, the sampling error of the virtual forest inventory was not statistically equivalent to the sampling error of the validation forest inventory data set. The forest species composition of the imputed forest inventory had an overall accuracy of 64.07% and a Kappa value of only 23.25 (Table 3).

The second model (Model 2; Table 3) also produced estimates of basal area, total volume, and SDI that were strongly correlated ($r > 0.77$) with the validation inventory metrics. However, all other forest inventory metrics were weakly correlated ($r < 0.27$) with the validation forest inventory metrics. The statistical equivalence test of the means indicated that basal area, volume, the density of trees 23–40 cm (DBH), and QMD were equivalent at the ±20% level, while the pairwise equivalence test indicated that basal area and SDI were also equivalent at the ±20% level. The sampling error from Model 2 was not equivalent to the sampling error of the validation forest inventory, and in terms of forest species composition, Model 2 had overall accuracy and a Kappa value of 66.67% and 27.49, respectively (Table 3).

In terms of correlations, the third model (Model 3; Table 3) was similar to Model 1. Basal area, volume, and SDI were strongly correlated ($r > 0.85$) with the independent data. The density of trees >53 cm and 23–40 cm (DBH), basal area weighted DBH, and the DBH of overstory trees exhibited moderate correlations (r = 0.65–0.74), whereas other forest inventory metrics displayed weak correlations ($r < 0.47$). At the ±20% equivalence level, the means of basal area, volume, SDI, the density of trees larger than 53 cm and trees 23–40 cm (DBH), QMD, weighted DBH, and the DBH of overstory trees were equivalent to the validation forest inventory data. However, the pairwise equivalence test indicated that only basal area and SDI were equivalent to the validation inventory data at the ±20% equivalence level. The mean sampling error for Model 3 was equivalent to the sampling error for the validation inventory; however, the pairwise estimates of sampling error were not equivalent at the ±20% equivalence level. Overall accuracy and the Kappa value for the species composition of the imputed forest inventory data were 69.23% and 12.89, respectively (Table 3).

Imputed forest inventory data from the fourth model (Model 4; Table 3) produced the most accurate results when compared with the independent forest inventory data set. Specifically, basal area, volume, and SDI exhibited strong correlations ($r > 0.88$), whereas the density of trees >23 cm (DBH), basal area weighted DBH, and the DBH of overstory trees were moderately correlated (r = 0.55–0.64) with the independent forest inventory data set. All other forest inventory metrics displayed weak correlations ($r < 0.39$). In terms of statistical equivalence, the mean and slope equivalence tests indicated that basal area, volume, and SDI were equivalent to the validation forest inventory data at the ±20% equivalence level. Furthermore, the mean equivalence test indicated that the density of trees 18–23 cm (DBH), QMD, basal area weighted DBH, the DBH of overstory trees, and the sampling error were equivalent to the validation forest inventory data. In terms of forest species composition, the imputed data had an overall accuracy of 68.18% and a Kappa value of 31.06 (Table 3; Fig. 3).

Fig. 4. Average Forest Vegetation Simulator basal area growth projections from the imputed (black) and validation (grey) inventory data sets of the 88 stands studied. Solid line is the projected data and dashed lines are ±1 standard deviation.



Growth projection comparison

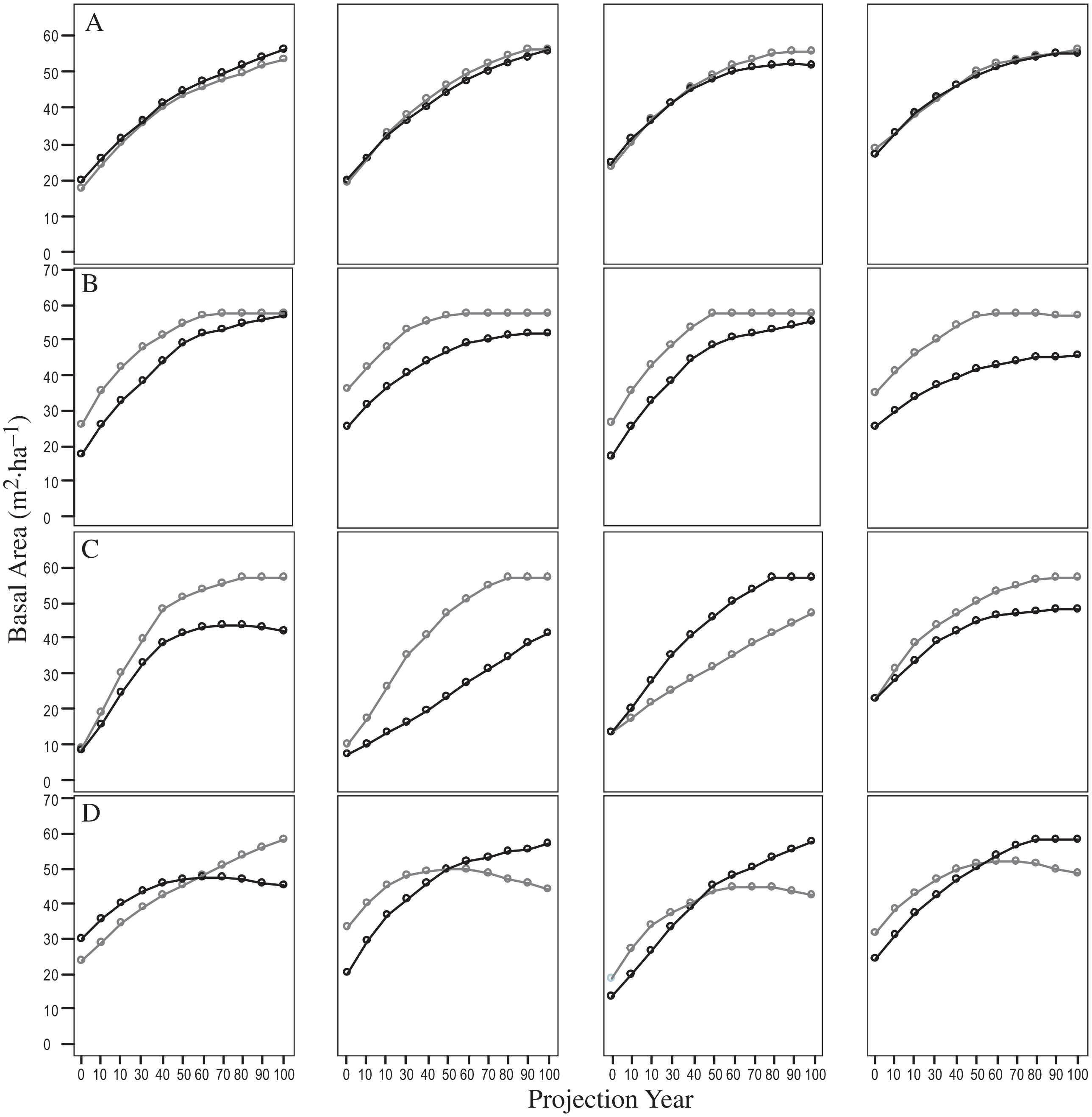
In general, the FVS growth projections from the imputed forest inventory data (from Model 4) and the validation forest inventory data followed similar trends. Specifically, correlations between basal area projections were greater than 0.91 (mean, minimum, and maximum basal area correlations are 0.96, 0.91, and 0.99, respectively), and RMSDs were less than 6.54 $m^2 \cdot ha^{-1}$ (mean, minimum, and maximum basal area RMSDs are 1.97, 0.24, and 6.54 $m^2 \cdot ha^{-1}$, respectively).

Discussion

Forest inventory metrics

The most accurate model (Model 4) was developed based on one *Y* variable (basal area) and four *X* variables (CANOPY, H95PCT, HMEAN, STRATUM2, and STRATUM4; see Table 1). In terms of correlation coefficients and RMSD statistics, Model 4 produced accurate estimates of basal area, total volume, SDI, as well as the density and DBH of large trees. However, imputed estimates of small-tree density and QMD, which incorporates small-tree diameters, were not accurate when compared with the validation forest inventory data set. The statistical equivalence test employed herein indicates that basal area, volume, and stand density calculated from the imputed forest inventory data were statistically equivalent to the same inventory metrics calculated from the validation inventory data set (i.e., we rejected $H_0$1, the null hypothesis of dissimilarity); all other in-

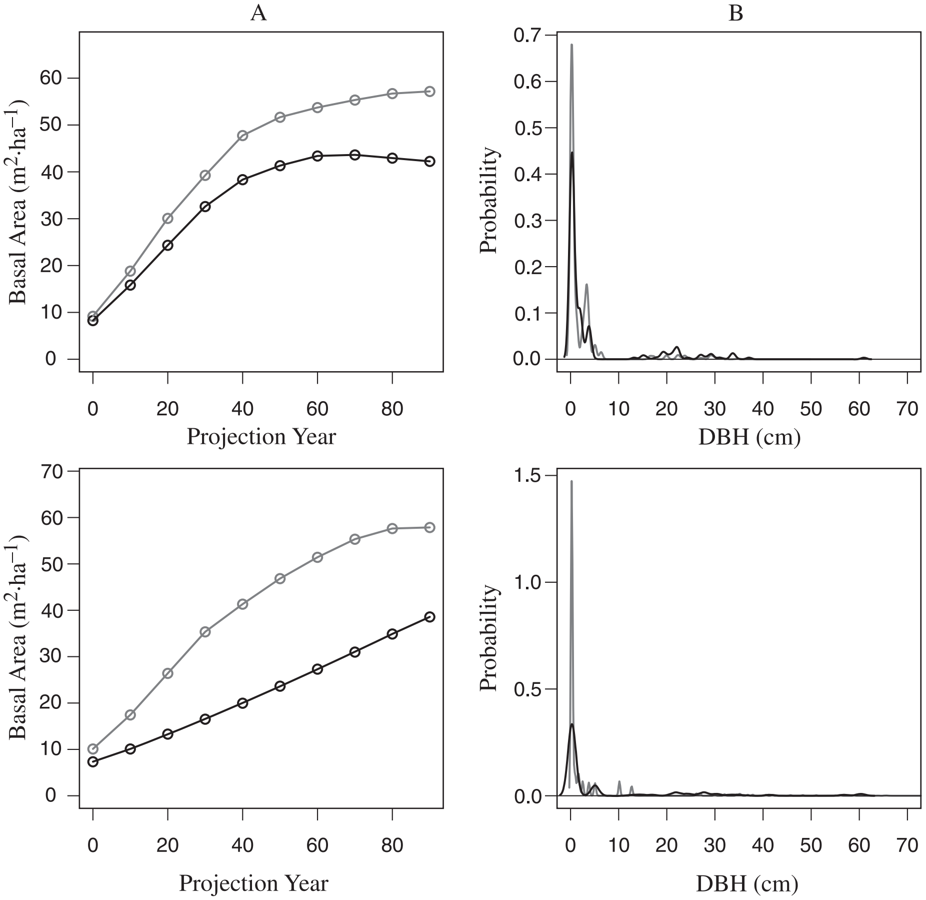
Fig. 5. Forest Vegetation Simulator basal area growth projections from the imputed (black lines) and validation (grey lines) inventory data sets of 16 stands.



ventory metrics were not equivalent at the ±20% equivalence level (i.e., we failed to reject $H_0$1, the null hypothesis of dissimilarity). These results suggest that the LiDAR metrics or modeling strategy presented herein cannot sufficiently characterize tree density or DBH, especially when small trees are incorporated into the inventory metrics.

Compared with other *k*-NN imputation studies, the imputation model presented in the current study produced comparable estimates of most forest inventory metrics. For example, Maselli et al. (2005) developed an imputation algorithm from Landsat data and attained a correlation of 0.72 and a RMSE of 3.65 $m^2 \cdot ha^{-1}$ when imputing basal area. Temesgen et al. (2003) compared tree-list predictions in mixed species, uneven-aged stands from four separate *k*-NN algorithms with different distance metrics and found that each of the algorithms tested produced similar results (stems per hectare RMSE = 302–631, basal area RMSE = 17–28 $m^2 \cdot ha^{-1}$, volume RMSE = 92.9–280 $m^3 \cdot ha^{-1}$). Reese et al. (2002) implemented a NN algorithm that included Landsat-derived auxiliary variables and attained an average RMSE of ∼120 $m^3 \cdot ha^{-1}$ when predicting stem volume. In a separate study, Holmström et al. (2001) utilized predictor variables from aerial photographs and achieved RMSEs of 49.4 and 26.8 $m^3 \cdot ha^{-1}$ for plot-level and stand-level esti-

Fig. 6. Forest Vegetation Simulator (FVS) basal area growth projections displaying errors of divergence for two different stands (figure rows). Column A: FVS basal area growth projections. Column B: probability density function of tree diameters. In both columns the black and grey lines correspond to the imputed and validation inventory data sets, respectively.

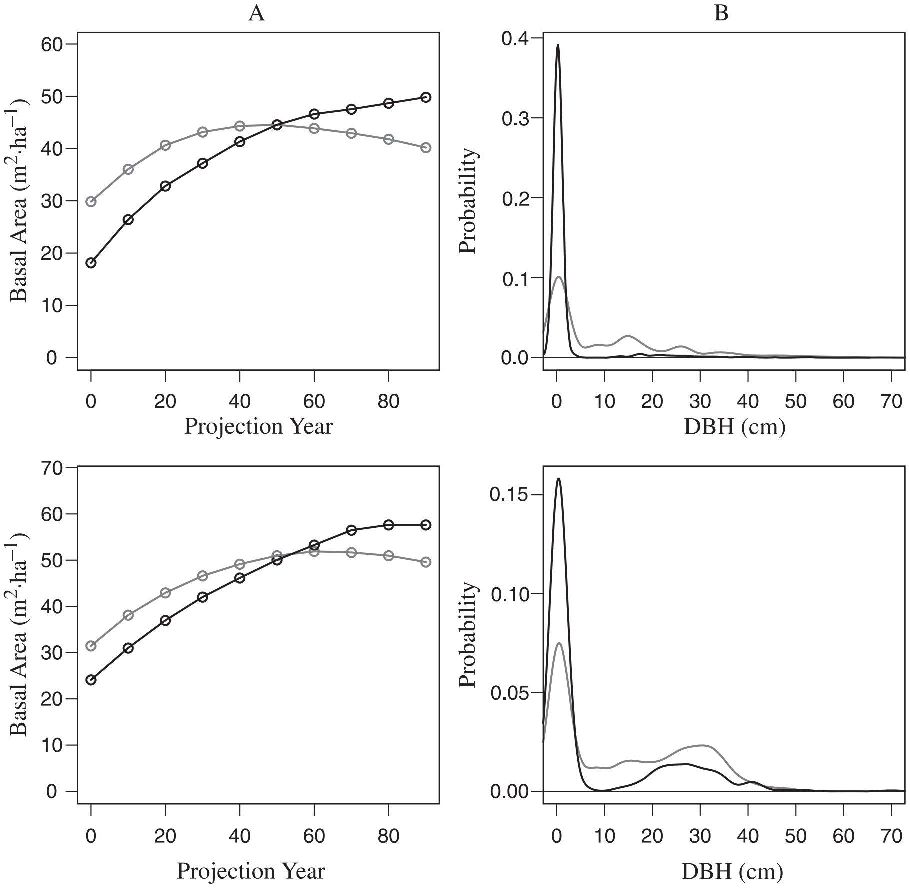


mates of stem volume, respectively. Tuominen and Pekkarinen (2005) also used predictor variables from aerial photography and achieved a relative RMSE of 58%. Studies that incorporated LiDAR-derived predictor variables produced more accurate results. For example, Maltamo et al. (2006) integrated LiDAR with aerial photography to generate stem volume imputations with a 5.89% RMSE, whereas Hudak et al. (2008) used LiDAR and reported correlation coefficients of 0.76 and 0.78 when imputing tree density and basal area, respectively. Wallerman and Holmgren (2007) developed an imputation model that integrated LiDAR information and optical data from the SPOT sensor and reported RMSEs of 45 $m^2 \cdot ha^{-1}$ and 209 $stems \cdot ha^{-1}$ when imputing basal area and tree density, respectively.

Sampling error

Although the mean sampling error of the virtual forest inventory is equivalent (±20%) to the mean sampling error of the validation forest inventory, the pairwise comparisons are not equivalent at the ±20% level (i.e., we fail to reject H_02, the null hypothesis of dissimilarity). Further analysis indicates that the difference between the virtual and validation forest inventory sampling errors is ±15 $m^2 \cdot ha^{-1}$ in seven of the 88 stands — four stands exhibit negative differences greater than 15 $m^2 \cdot ha^{-1}$ and three stands exhibit positive differences greater than 15 $m^2 \cdot ha^{-1}$. The positive differences occur in small stands with highly variable forest structure (i.e., stands with widely spaced individual trees or clumps trees). Overestimating sampling error in stands with high structural variability is not surprising given that small shifts in the location of an imputation target (i.e., plot location) would produce drastically different intrastand estimates of plot-level forest inventory parameters. Although the methodology presented herein attempted to impute to exact reference plot locations (i.e., field plot locations), global positioning system measurement errors could offset the imputation targets enough to introduce significant differences in sampling error estimates when stand conditions are highly variable. This becomes a problem in small stands because there are fewer targets to account for the high variability in forest structure. The negative differences in sampling error occur in relatively homogeneous, closed-canopy stands with basal areas greater than 160 $m^2 \cdot ha^{-1}$. Since only five of the 88 stands analyzed in this study have basal areas greater than 160 $m^2 \cdot ha^{-1}$, there is only a small pool of reference plots for the algorithm to choose from when imputing to areas with high basal areas. This results in the same reference plot being imputed to multiple target locations within stands with high basal areas, ultimately reducing sampling error estimates.

Fig. 7. Forest Vegetation Simulator (FVS) basal area growth projections displaying errors of convergence followed by divergence for two different stands (figure rows). Column A: FVS basal area growth projections. Column B: probability density function of tree diameters. In both columns the black and grey lines correspond to the imputed and validation inventory data sets, respectively.



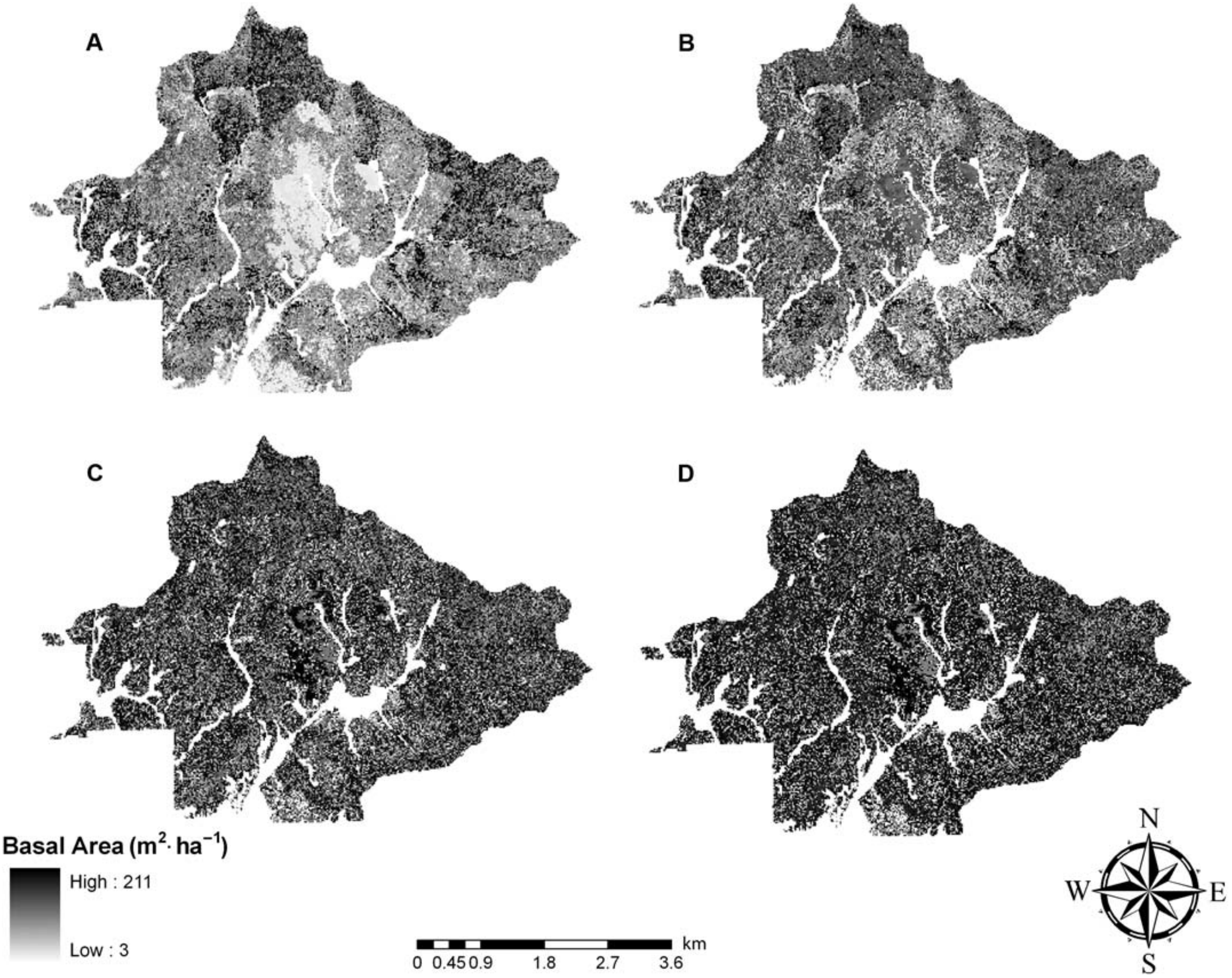
Species composition

When compared with the validation forest inventory data, the imputation models displayed species composition accuracies ranging between 64% and 69% for seven different forest type classes. However, the Kappa values were quite low (≤31%), indicating that the overall accuracies were less than 31% better than a random classification (i.e., a classification arrived at by pure chance). Furthermore, models that included LiDAR topographic metrics (Models 1 and 2) had the lowest overall accuracies and had Kappa values below 30%. In this case, we failed to reject the null hypothesis of species composition dissimilarity ($H_0$3). This result indicates that the forest types present across the Damon study area cannot be accurately classified based upon topographical variables alone. This is not surprising given the relatively narrow range of environmental conditions and forest types found across the study area. Current forest species composition throughout the study area is more likely a function of disturbance history (i.e., logging and fire) as opposed to environmental gradients. Employing LiDAR-derived topographic metrics to predict forest species composition may produce higher classification accuracies in study areas with stronger environment gradients and less disturbance history. In addition, integrating remotely sensed data collected by spectral sensors (e.g., Landsat and SPOT) with the LiDAR metrics presented herein may improve species composition accuracies across similar study areas.

FVS growth projections

Results of the growth projection comparison between forest inventory data from the virtual and validation forest inventory data sets demonstrates that the growth projections follow similar trends in most of the 88 stands analyzed is this study. In general, stand basal area projections from the imputed and validation inventory data sets were highly correlated, had low RMSDs, and followed similar trends when averaged across all stands (Fig. 4). Nineteen of the 88 basal area projections had RMSD statistics >3 m^2·ha^{-1}. Eleven of these 19 stands had imputed tree density errors that were greater than 500 trees·ha^{-1}, while nine had imputed basal area errors greater than 5 m^2·ha $^{-1}$. Only four of these 19 stands had species composition imputation errors. A visual comparison of the FVS growth projections within each stand revealed four unique scenarios: (*i*) no imputation error; near perfect agreement between projections across all time steps (Fig. 5A); (*ii*) an imputation error at time zero resulting in offset growth projections (Fig. 5B); (*iii*) no imputation error at time zero; however, growth projections diverge (Fig. 5C); and (*iv*) an imputation error at time zero with growth projections intersecting midway through the projection time series (Fig. 5D). The latter two types of errors (i.e., diverging or intersecting growth projections) are caused by significant differences in the density of trees by size class between the imputed and field-measured forest inventory data. For example, diverging growth projections typically occur in stands

Fig. 8. Spatial FVS basal area growth projections. Four time steps across the north section of the Damon study area: A is 2007, B is 2037, C is 2067, and D is 2097.



with larger basal areas (<10 m^2·ha^{-1}), with a large difference in the density of small trees (seedlings) between the imputed and field-measured forest inventory data. This results in an over- or under-estimation in basal area at the end of the growth projections (Fig. 6). On the other hand, intersecting growth projections typically occur in stands with moderate basal areas (>20 m^2·ha^{-1}) exhibiting imputation errors in both the density of small and large trees (Fig. 7). These results suggest that in the forest types present within the Damon study area, FVS growth projections are more influenced by errors in forest structure than by errors in species composition. Future research should focus upon developing improved methods for characterizing the density and size of small trees in the forest understory. The best imputation methodology presented herein can also be employed to create spatially continuous predictions of tree-level forest inventory data, which can in turn be used to spatially parameterize FVS at landscape scales. Figure 8 presents four time periods from a spatial FVS growth projection executed across the north section of the Damon study area.

Conclusions

This study presents a novel methodology for predicting tree-level forest inventory data in unsampled areas via an imputation modeling procedure incorporating LiDAR-derived predictor variables. The imputation methodology presented herein proved to be an effective approach to generate ‘virtual’ forest inventory data from LiDAR metrics across the Damon study area. Most forest inventory metrics calculated from the imputed data had high accuracies when compared with independent forest inventory data. Furthermore, most FVS growth projections followed similar trends. This study represents a significant improvement in our capabilities to predict the size and species of every tree within every management unit across an entire forest. The imputed, tree-level forest inventory data will be used in conjunction with FVS to evaluate various alternative management decisions across the Damon study area. Specifically, a project focused upon evaluating the efficacy of fuels reduction treatments is now underway. In addition to evaluating management decisions, imputed tree-level forest inventory data could be used for a variety of applications, including forest commodity assessment, carbon accounting, wildlife habitat assessment, among others.

Acknowledgements

This work was primarily supported through Agenda 2020's sustainable forestry initiative via a grant provided by the USDA Forest Service Rocky Mountain Research Station, Moscow Forest Sciences Laboratory (RJVA-11222063-299). The authors also acknowledge additional sources of funding that supported this work, including the USDA Forest Service Rocky Mountain Research Station Missoula Fire Sciences Laboratory (RJVA-11222048-140), the NASA Synergy program, and the University of Idaho's Geospatial Laboratory for Environmental Dynamics. Data acquisition was funded by the USDA Forest Service's Pacific Northwest Research Station and the Malheur National Forest.

References

Atkins, D., and Lundberg, R. 2002. Analyst hazards when assessing fire, insect, and disease hazard in Montana using FIA data with FVS: or alligators we didn't see coming. *In* Second Forest Vegetation Simulator Conference, 12–14 February 2002, Fort Collins, Colorado. *Compiled by* N.L. Crookston and R.N. Havis. USDA For. Serv. Proc. RMRS-P-25, Ogden, Utah. pp. 83–90.

Barrett, T.M., and Fried, J.S. 2004. Inventory: modeling. 2004. *In* Encyclopedia of forest science. Elsevier, Amsterdam, Netherlands. pp. 426–433.

Breiman, L. 2001. Random forests. Mach. Learn. **45**(1): 5–32.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**(1): 37–46. doi:10.1177/001316446002000104.

Congalton, R., and Green, K. 1999. Assessing the accuracy of remotely sensed data: principles and practices. CRC/Lewis Press, Boca Raton, Fla., USA.

Cousar, P., Sessions, J., and Johnson, K.N. 1997. Individual stand projection under different goals to support policy analysis for the Sierra Nevada ecosystem project. *In* Proceedings: Forest Vegetation Simulator Conference, 3–7 February 1997, Fort Collins, Colorado. *Compiled by* R. Teck, M. Moeur, and J. Adams. USDA For. Serv. Gen. Tech. Rep. INT-373, Ogden, Utah. pp. 98–104.

Crookston, N.L., and Dixon, G.E. 2005. The forest vegetation simulator: a review of its structure, content and applications. Comput. Electron. Agric. **49**(1): 60–80. doi:10.1016/j.compag.2005.02.003.

Crookston, N.L., and Finley, A.O. 2008. YaImpute: an R package for *k*-NN imputation. J. Stat. Softw. **28**(10): 1–16.

Diaz-Uriarte, R. 2007. GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. BMC Bioinformatics, **8**(1): 328. doi:10.1186/1471-2105-8-328. PMID:17767709.

Dixon, G.E. 2003. Essential FVS: a user's guide to the forest vegetation simulator. Internal report. USDA For. Serv., For. Manage. Serv. Center, Fort Collins, Colorado.

Eng, H. 1997. Vegetation projection and analysis of the cumulative effects of timber harvest. *In* Proceedings: Forest Vegetation Simulator Conference, 3–7 February 1997, Fort Collins, Colorado. *Compiled by* R. Teck, M. Moeur, and J. Adams. USDA For. Serv. Gen. Tech. Rep. INT-373, Ogden, Utah. pp. 69–74.

Evans, J.S., and Hudak, A.T. 2007. A multiscale curvature algorithm for classifying discrete return lidar in forested environments. IEEE Trans. Geosci. Rem. Sens. **45**(4): 1029–1038. doi:10.1109/TGRS.2006.890412.

Fehrmann, L., Lehtonen, A., Kleinn, C., and Tomppo, E. 2008. Comparison of linear and mixed-effect regression models and a *k*-nearest neighbour approach for estimation of single-tree biomass. Can. J. For. Res. **38**(1): 1–9. doi:10.1139/X07-119.

Franco-Lopez, H., Ek, A, and Marvin E.B. 2001. Estimation and mapping of forest stand density, volume, and cover type using the *k*-nearest neighbors method. Remote Sens. Environ. **77**(3): 251–274. doi:10.1016/S0034-4257(01)00209-7.

Fulé, P.Z., Cocke, A.E., Heinlein, T.A., and Covington, W.W. 2004. Effects of an intense prescribed forest fire: Is it ecological restoration? Restor. Ecol. **12**(2): 220–230. doi:10.1111/j.1061-2971.2004.00283.x.

Holmström, H., Nilsson, M., and Stahl, G. 2001. Simultaneous estimations of forest parameters using aerial photograph interpreted data and the *k* nearest neighbor method. Scand. J. For. Res. **16**: 67–68. doi:10.1080/028275801300004424.

Hudak, A.T., Crookston, N.L., Evans, J.S., Falkowski, M.J., Smith, A.M.S., Gessler, P., and Morgan, P. 2006. Regression modeling and mapping of coniferous forest basal area and tree density from discrete-return lidar and multispectral satellite data. Can. J. Rem. Sens. **32**: 126–138.

Hudak, A.T., Crookston, N.L., Evans, J.S., Hall, D.E., and Falkowski, M.J. 2008. Nearest-neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. Remote Sens. Environ. **112**(5): 2232–2245. doi:10.1016/j.rse.2007.10.009.

Kohl, M. 2004. Inventory: multipurpose resource inventories. *In* Encyclopedia of forest sciences. Elsevier Academic Press, San Diego, Calif. pp. 403–409.

Lawrence, R.L., Wood, S.D., and Sheley, R.L. 2006. Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (RandomForest). Remote Sens. Environ. **100**(3): 356–362. doi:10.1016/j.rse.2005.10.014.

LeMay, V., and Temesgen, H. 2005. Comparison of nearest neighbor methods for estimating basal area and stems per hectare using auxiliary variables. For. Sci. **51**: 109–119.

Lund, H.G. 2004. Inventory: forest inventory and monitoring. 2004. *In* Encyclopedia of forest sciences. Elsevier Academic Press, San Diego, Calif. pp. 414–420.

Maffei, H., and Tandy, B. 2002. Methodology for modeling the spatial and temporal effects of vegetation management alternatives on late successional habitat in the Pacific Northwest. *In* Second Forest Vegetation Simulator Conference, 12–14 February 2002, Fort Collins, Colorado. *Complied by* N.L. Crookston and R.N. Havis. USDA For. Serv. Proc. RMRS-P-25, Ogden, Utah. pp. 69–77.

Mäkelä, H., and Pekkarinen, A. 2004. Estimation of forest stand volumes by Landsat TM imagery and stand-level field-inventory data. For. Ecol. Manage. **196**(2-3): 245–255. doi:10.1016/j.foreco.2004.02.049.

Maltamo, M., and Eerikainen, K. 2001. The most similar neighbour reference in the yield prediction of *Pinus kesiya* stands in Zambia. Silva Fenn. **35**: 437–451.

Maltamo, M., Malinen, J., Packalén, P., Suvanto, A., and Kangas, J. 2006. Nonparametric estimation of stem volume using airborne laser scanning, aerial photography, and stand-register data. Can. J. For. Res. **36**(2): 426–436. doi:10.1139/x05-246.

Maselli, F., Chirici, G., Bottai, L., Corona, P., and Marchetti, M. 2005. Estimation of Mediterranean forest attributes by the application of *k*-NN procedures to multitemporal Landsat ETM+ images. Int. J. Remote Sens. **26**(17): 3781–3796. doi:10.1080/01431160500166433.

McRoberts, R.E., Tomppo, E.O., Finley, A.O., and Heikkinen, J. 2007. Estimating areal means and variances of forest attributes using the *k*-nearest neighbors technique and satellite imagery. Remote Sens. Environ. **111**(4): 466–480. doi:10.1016/j.rse.2007.04.002.

Milner, K.S., and Coble, D.W. 1995. A mechanistic approach to predicting the growth and yield of stands with complex structures. *In* Proceedings of the Conference: Uneven-Aged Management: Opportunities, Constraints, and Methodologies, 29 April 1995, Missoula, Montana. *Edited by* K.S. O'Hara. University of Montana, Missoula, Mont. MFCES Misc. Publ. 56. pp. 144–166.

Milner, K.S., Coble, D.W., McMahan, A.L., and Smith, E.L. 2003. FVSBGC: a hybrid of the physiological model STAND-BGC and the forest vegetation simulator. Can. J. For. Res. **33**(3): 466–479. doi:10.1139/x02-161.

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

Prasad, A.M., Iverson, L.R., and Liaw, A. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems (N.Y., Print), **9**(2): 181–199. doi:10.1007/s10021-005-0054-1.

R Development Core Team. 2005. R: a language and environment for computing [online]. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org/ [accessed 7 July 2008].

Reese, H., Nilsson, M., Sandström, P., and Olsson, H. 2002. Applications using estimates of forest parameters derived from satellite and forest inventory data. Comput. Electron. Agric. **37**(1-3): 37–55. doi:10.1016/S0168-1699(02)00118-7.

Roberts, D.W., and Cooper, S.V. 1989. Concepts and techniques of vegetation mapping. *In* Land classifications based on vegetation: applications for resource management. *Edited by* D. Ferguson, P. Morgan, and F.D. Johnson. USDA For. Serv. Gen. Tech. Rep. INT-257. Ogden, Utah. pp. 90–96.

Robinson, A.P., Duursma, R.A., and Marshall, J.D. 2005. A regression-based equivalence test for model validation: shifting the burden of proof. Tree Physiol. **25**(7): 903–913. PMID:15870057.

Stage, A.R. 1976. An expression for the effect of slope, aspect and habitat type on tree growth. For. Sci. **22**: 457–460.

Stage, A.R., and Crookston, N.L. 2007. Partitioning error components for accuracy-assessment of near-neighbor methods of imputation. For. Sci. **53**: 62–72.

Tarboton, D.G. 1997. A new method for the determination of flow directions and contributing areas in grid digital elevation models. Water Resour. Res. **33**(2): 309–319.

Temesgen, H., LeMay, V.M., Froese, K.L., and Marshall, P.L. 2003. Imputing tree-lists from aerial attributes for complex stands of south-eastern British Columbia. For. Ecol. Manage. **177**(1-3): 277–285. doi:10.1016/S0378-1127(02)00321-3.

Tuominen, S., and Pekkarinen, A. 2005. Performance of different spectral and textural aerial photograph features in multi-source forest inventory. Remote Sens. Environ. **94**(2): 256–268. doi:10.1016/j.rse.2004.10.001.

Wallerman, J.P., and Holmgren, J. 2007. Data capture for forest management planning using sample plot imputation based on spatial statistics, laser scanner, and satellite image data. Remote Sens. Environ. **110**: 227–234.

Wilson, L. 1997. The use of the Forest Vegetation Simulator for the California spotted owl environmental impact statement. *In* Proceedings: Forest Vegetation Simulator Conference, 3–7 February 1997, Fort Collins, Colorado. *Compiled by* R. Teck, M. Moeur, and J. Adams. USDA For. Serv. Gen. Tech. Rep. INT-373, Ogden, Utah. pp. 64–68.

Zdanowicz, C.M., Zielinski, G.A., and Germani, M.S. 1999. Mount Mazama eruption; calendrical age verified and atmospheric impact assessed. Geology, **27**(7): 621–624. doi:10.1130/0091-7613(1999)027<0621:MMECAV>2.3.CO;2.