Geospatial Forest Inventory Models for Michigan

Ram K. Deo, Robert E. Froese (PI) and Michael J. Falkowski

School of Forest Resources and Environmental Science
Michigan Technological University, Houghton, MI 49931, U.S.A.

30 September 2011

Acknowledgement:

This material is based upon work supported by the Department of Energy under award number DE-EE-0000280.

Michigan Tech
Michigan Technological University

Executive Summary

Strategic decisions concerning the timing and location of forest operations require accurate information about the state and rate of change of forest growing stock. Regional forest management would benefit from spatially explicit inventory data across the entire forest area. Due to cost considerations in detailed field inventories, remotely sensed data have often been employed to extend forest inventories through both space and time. The purpose of this study was to generate spatially explicit predictions of growing stock inventory, growth, mortality and removals of above ground biomass across the entire state of Michigan, U.S.A. This is achieved by coupling spectral information from Landsat TM images and other geospatial layers with the inventory data from Forest Inventory and Analysis (FIA) program of U.S. Forest Service in a kNN imputation procedure. The predictor layers in the imputation process included normalized difference vegetation index (NDVI), land cover classes, basal area weighted height (BAWHT), and digital elevation model (DEM) while volume, growth, mortality and removal data from 6,702 sample plots of FIA were utilized for both model calibration and validation. A Random Forest imputation algorithm was implemented to develop spatially explicit forest inventory information across the Entire State of Michigan via the integration of geospatial data with forest inventory plot measurements. Generally, state wide spatial inventory was unbiased but demonstrated relatively low precision. Inclusion of BAWHT as an explanatory variable was found to significantly improve the prediction accuracy. Three levels of comparison were made to evaluate the accuracy of the inventory estimates: at the plot, stand and county levels. The validation procedure confirmed that county level estimates were better than stand level which in turn was better than plot level estimates. The statistical modeling approach employed (imputation) achieved acceptable accuracies when compared to other studies.

Introduction

Forest inventory and monitoring are crucial to develop sound forest management plans, which in turn help in achieving desired ecological, economic, and social objectives. The strategic decisions concerning the timing and location of forest operations are typically dependent on assessment and mapping of forest variables. Two types of forest inventories can be distinguished based on the extent of survey: strategic inventories focus on large area estimates for a large number of attributes while management inventories focus on small area estimates to guide operational forest management (McRoberts *et.al.,* 2007). The number of variables measured during forest inventory is usually high, and new variables are further derived through computations (Tomppo *et.al.*, 2002). Regional forest planning requires spatially explicit inventory of forest attributes across the entire area of interest. Since exhaustive field measurements at the landscape or regional level are prohibitively expensive, remotely sensed data are often coupled with limited forest inventory data to extend the measurements through both space and time.

Imputation Methods

K nearest neighbor (kNN) imputation can be used to develop spatially explicit inventories by coupling sparse sample plot data with continuous scale remote sensing data partitioned into pixel units. The kNN imputation is often applied to (i) supply missing data to complete a data set for subsequent analyses or (ii) to estimate sub-population totals (Reese *et.al.*, 2002; Stage and Crookston, 2007). The idea that motivates kNN imputation methods is that two records with similar X-values (predictor variables) should have similar Y-values (dependent variables) (Eskelson *et.al.*, 2009). The kNN imputation is an approach commonly used to extrapolate forest inventory data collected at discrete sampling locations to progressively larger spatial extents. The value imputed to a location can be a value measured at another sample plot location, or an average value computed from multiple sample plot locations (Eskelson *et.al.*, 2009). In a forestry context, the kNN imputation exploits the association between auxiliary variables that are inexpensive to measure over the entire area of interest (such as remote sensing and geospatial data), and forest attributes (which are expensive to measure) of interest measured at discrete sampling locations within the area of interest (Crookston and Finley, 2008). The process involves integrating forest plot inventory data with spatially explicit geo-information on land cover, topography, climate, among others, which are often derived from remote sensing data.

The kNN imputation is a two phase sampling procedure in which the first phase involves identifying and obtaining spatially explicit auxiliary layers (geospatial predictors) and the second phase involves a detailed sample plot inventory for the required forest parameters (Falkowski *et.al.*, 2010; Falkowski, 2008; Moeur and Stage, 1995). In the procedure, a reference dataset is first produced to generate a model which is then generalized over a continuous target dataset to predict attributes of interest in un-inventoried areas (Hudak *et.al.*, 2008). The reference set consists of field measured sample plot data and corresponding pixel values from geo-referenced raster layers while target set is comprised of only the predictor variables for the total area of interest in the form of pixel values (Bernier *et.al.,* 2010; Hudak et.al., 2008). The kNN imputation algorithm associates one or more of the sample plot data in the reference set to unsampled areas based on the spectral similarity between sampled and unsampled areas. The spectral similarity between a target and reference pixels can be determined by calculating spectral space distance based on covariate (auxiliary variables) characteristics available from geospatial layers for both target and reference sets (Falkowski, *et.al.,* 2010; LeMay and Temesgen, 2005; McRoberts *et.al.*,

2007). The spectral distance (i.e. nearness of a target and reference pixels) can be measured through several algorithms such as Euclidian distance, Mahalanobis distance, and random forest proximity and target locations can then be imputed with response variables from the nearest neighbors in the reference dataset.

Imputation mapping is a promising technique, with potential for generating spatially explicit, border-to-border information on forest composition (Grossmann, *et.al.*, 2009). The kNN method is non-parametric (i.e. there is no assumption of distributional characteristics of the variables) and can estimate multiple forest variables simultaneously and is a simple but powerful tool to extend a wide range of field data to landscapes (Haapanen *et.al.*, 2002; LeMay and Temesgen, 2005; Katila and Tomppo, 2002). The method can also preserve the covariance structure of forest variables and thus produce maps that appear very realistic in terms of their spatial pattern (Haapanen *et.al.*, 2002; Holmstrom, 2003; LeMay and Temesgen, 2005; Moeur and Stage, 1995). Nearest neighbors techniques have been shown to be useful for predicting multiple forest attributes from forest inventory and remote sensing data such as Landsat imagery (McRoberts, 2009).

Imputation mapping is impacted by a wide array of factors including the selection of explanatory variables. Examples of such factors include raster images from satellite sensors as explanatory layers, type of distance metric (for finding nearest neighbors), and the number of nearest neighbors (i.e. value of *k*) from a reference set to be considered in the imputation (Kalila and Tomppo, 2001). The type of distance metric and the number of neighbors (*k*) used in imputation mapping vary among applications (LeMay and Temesgen, 2005). When a single nearest neighbor is considered for imputing target locations, then simply the response variables of the nearest neighbor is assigned to the target points (Crookston and Finley, 2008; Falkowski, 2010; Hudak *et.al.*, 2008), and in such a case the natural variation of the forest variables is retained in the prediction but accuracy of prediction reduces (Katila and Tomppo, 2001; Makela and Pekkarinen, 2004). When more than one nearest neighbors are used to impute missing parameters at target locations, the prediction accuracy improves but more bias is introduced (McRoberts *et.al.,* 2002). The bias introduced with increasing values of *k* can be reduced to some extent by undertaking weighted average of the *k* neighbors (Kalila and Tomppo, 2001).

The pixel-wise estimates of any forest parameter can be made as the weighted average of the parameter (*v*) measured in *k* nearest sample plots (*j*). For any target pixel (*p*) in a feature space, the value of a parameter (*v*) can be expressed as in the equation 1. The weights assigned to each of the *k* samples are generally proportional to the inverse squared Euclidean distance (*d*) between the pixel to be estimated (*p*) and the reference plot-pixel, *j* (see equation 2) (Haapanen, *et.al.*, 2002; Nilsson, 2002; Reese *et.al.*, 2002; Tomppo *et.al.*, 2002). The estimate of variable (*v*) for pixel *p* is

$$\hat{v}_p = \sum_{j=1}^{k} w_{j,p} \times v_{j,p} \quad \text{.....1}$$

Where

$$w_{j,p} = \frac{1}{d_{j,p}^2} \Bigg/ \sum_{i=1}^{k} \frac{1}{d_{i,p}^2} \quad \text{.....2}$$

such that $d_{1,p} \leq d_{2,p} \leq \ldots\ldots \leq d_{k,p}$ and

$d_{j,p}$ = feature space distance from pixel *p* to plot *j* and

$v_{j,p}$ = variables for the plot with distance $d_{j,p}$

Haapanen and Ek (2001) have described the kNN algorithm in a straightforward way as:

1. For each target pixel calculate the Euclidean distance to all reference pixels
2. Rank the *k* nearest plots based on the Euclidean distance
3. Calculate weighted average of the desired forest parameters of the *k* nearest plots
4. Proceed to the next target pixel.

The kNN imputation simultaneously gives estimates for more than one response variables (Bernier, *et.al.,* 2010); in fact, the plot identifiers (IDs) of reference set are imputed to the missing locations based on the similarity of explanatory variables and the values of the response variables corresponding to the IDs are assigned to that location (Falkowski, 2010). The imputation with a single nearest neighbor produces output with similar variance structure to that of the sample plots measurements. Imputation error is evidently greater than ordinary least square regression because values assigned to each target unit using single neighbor imputation are the original values from reference plot data (Hudak *et.al.*, 2008). The root mean square difference (RMSE) calculated from the difference between imputed and observed values provide a measure of model error.

The plot level accuracy of imputation estimates are found to increase somewhat with higher values of *k* but also leads to over-prediction of forest growing stock (and hence bias) in areas having low stock and vice versa (Grossmann *et.al.*, 2009). Depending on the purpose of imputation, a lower value of *k* is set to retain the variation of field variables in the estimation and mapping while a higher value of *k* selected to minimize pixel level RMSE (Kalila and Tomppo, 2001; McRoberts *et.al.,* 2002). The errors of omission are found to decrease with increasing levels of *k*, but errors of commission increase for forest type predictions (Grossmann *et.al.*, 2009; Haapanen *et.al.*, 2002). Also by increasing the number of sample plots in the reference set, the prediction accuracy can be expected to improve as better matches, in terms of predictor variables (X-values), would be found for the un-sampled areas (LeMay and Temesgen, 2005). Thus, to obtain a reliable estimate with the kNN method, it is important to have a practically large sample of field inventory plots representing all forest conditions available in the area of interest (Haapanen *et.al.*, 2002; Tomppo, 2006). The sample plots in the reference set are assumed to characterize the entire range of variability in the predictor variables i.e. the field plots and the corresponding geospatial data should characterize the entire area of study (Hudak *et.al.*, 2008). Choice of X and Y variables, distance measure and *k* all contribute to error, but no single choice gives best result for all applications, nor for all response variables within a given application, and models must be developed on a case-by-case basis (Eskelson *et.al.* 2009; Ohmann *et.al.*, 2011).

There are several popular variants of kNN and most of these methods define nearness in terms of weighted Euclidean distance. Previous researches suggest that Random Forest (RF) NN outperforms other NN methods for mapping forest attributes (Ohmann *et.al.* 2011). The RF algorithm offers a novel nearest neighbor distance metric for imputation problems and can handle both categorical and continuous data

simultaneously (Breiman, 2001). The algorithm is based on classification and regression tree (CART) technique that has achieved excellent results in classifying remotely sensed data (Falkowski *et.al.* 2009).

RF is a type of 'ensemble learning' that generates many trees and aggregates their results. The two well-known methods of ensemble learning are boosting and bagging. In boosting, successive trees give extra weights to elements incorrectly predicted by earlier trees and final prediction is based on weighted voting of each. In bagging, each tree is independently constructed using a bootstrap sample of the data set and final prediction is based on a simple majority vote of prediction (Liaw and Wiener, 2002). The RF algorithm begins with the selection of many bootstrap samples and a classification tree is fitted to each of the bootstrap samples (see Figure 1) such that each node of a tree is split using the best among a subset of predictor variables selected randomly at that node for the binary partitioning (Breiman and Cutler, 2004; Cutler *et.al.*, 2007; Liaw and Wiener, 2002). RF can be thought of as a special case of bagging: in the case of bagging, best split at a node is chosen from among all predictors but in case of RF the best split is chosen from a random sample (subset) of predictors (Liaw and Wiener, 2002). Hence, every tree is different owing to two factors: first, at each node, a best split is chosen from a random subset of the predictors rather than all of them; second, every tree is build using a bootstrap sample. At each step in fitting a classification tree, an optimization is carried out to select a node, a predictor variable, and a cut-off value (for numerical variables) that result in the most homogeneous subgroups for the data, as measured by the Gini index (Falkowski *et.al.,* 2009). The splitting process continues until further subdivision no longer reduces the Gini index. Every time a split of a node is made on variable *m* the Gini impurity criterion for the two descendent nodes is less than the parent node. After the trees are fully-grown, each is used to predict the OoB observations (about one-third of the cases are left out of the bootstrap sample). The predicted class of an observation is calculated by majority vote of the OoB predictions for that observation (Cutler *et.al.*, 2007). The error rates are computed for each observation using the OoB predictions and then averaged over all observations. Thus out-of-bag (OoB) observations are used to estimate the prediction accuracy and variable importance for an element, and there is no need for cross-validation or a separate test set to get an unbiased estimate of error (Breiman and Cutler, 2004; Cutler *et.al.*, 2007; Falkowski *et.al.*, 2009; Liaw and Wiener, 2002; Pang *et.al.*, 2006). The RF algorithm is strictly non-parametric, flexible and robust with respect to non-linear and noisy relations among input features and class labels (Breiman, 2001; Cutler *et.al.*, 2007; Falkowski *et.al.*, 2009).

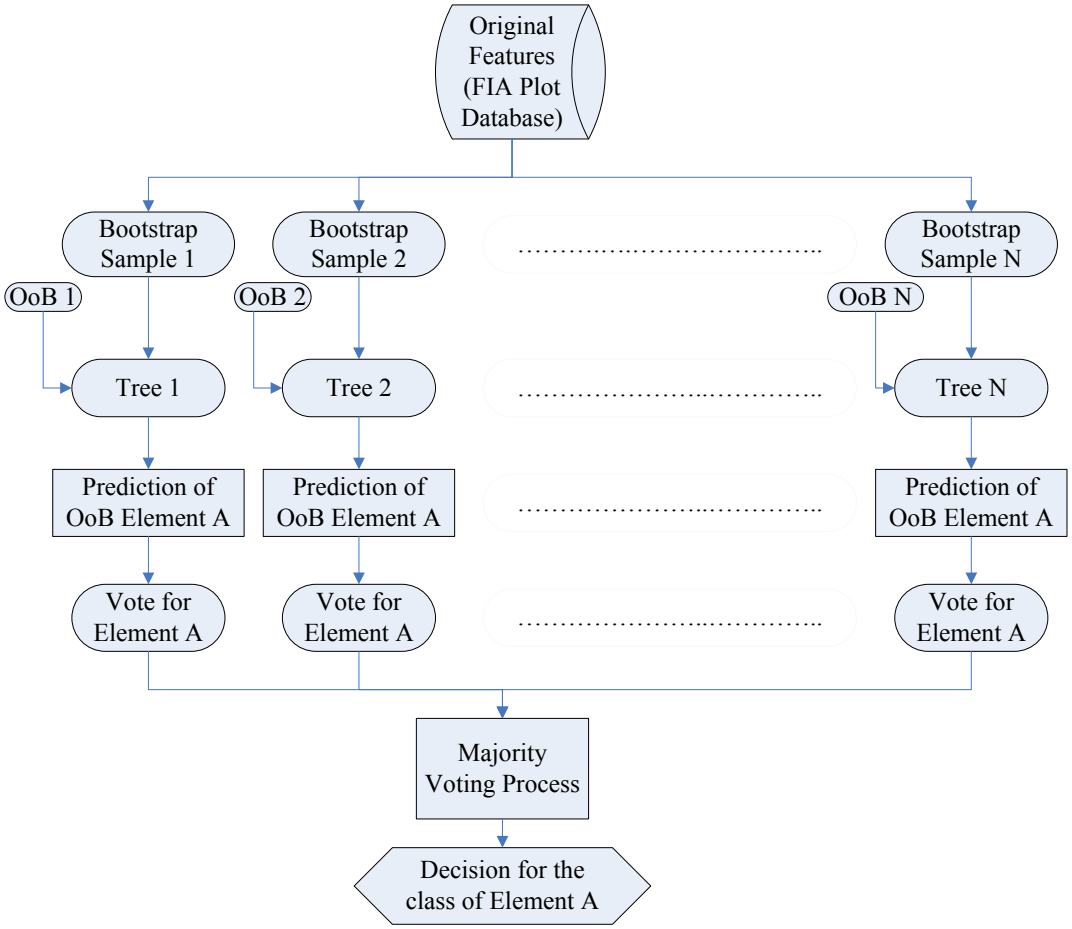

Figure 1. A simple schematic of random forest algorithm (adapted from Yang *et.al.* 2008)

In RF, nearness is defined by one minus the proportion of trees where target observation is in the same terminal node as a reference observation (Breiman, 2001; Crookston and Finley, 2008; Liaw and Wiener, 2002). The intuition is that similar observations should be in the same terminal nodes more often than dissimilar ones (Liaw and Wiener, 2002).

The RF described by Breiman (2001) has following characteristics:

- Its accuracy is as good as Adaboost and sometimes better
- It's relatively robust to outliers and noise
- It's faster than bagging and boosting
- It gives useful internal estimates of error, strength, correlation and variable importance
- It's simple and easily parallelized.

Data Sources

Forest Inventory and Analysis Program

The Forest Inventory and Analysis (FIA) program of the US Forest Service has a nationwide periodic inventory system for national, regional, and state-level assessment of forest resources to describe status and change in forest resources (McRoberts, 2000). The FIA program has collected and compiled data for

several attributes under the permanent plot design, established especially after 1999 when FIA program shifted from a periodic inventory to an annual inventory system (Woudenberg *et.al.* 2010). FIA permanent ground plots are designed to cover 1-acre sample area (however, not all trees on the acre are measured) such that each plot is representative of 6,000 acre hexagonal area on ground and consists of a national standard fixed radius four sub-plots as shown in Figure 2 (Woudenberg *et.al.* 2010; Burkman, 2005). The annual inventory scheme of FIA program covers upto 20 percent of the total plots (called panel) in a state each year (hence, generally all plots in a state are visited once in five years and have five panels). The sampling intensity of this program is obviously designed for large area estimation such as entire states or regions.

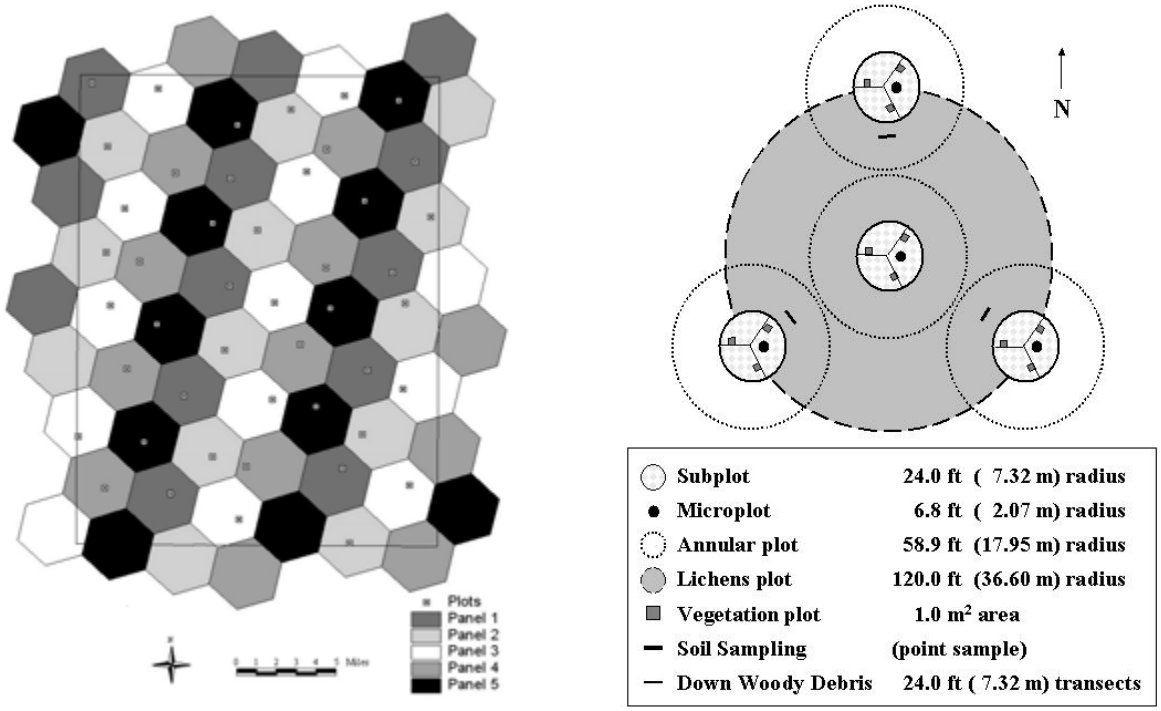

Figure 2. Distribution of one FIA sample plot per hexagon of 6,000 acre on ground (left) and standard plot design for the measurement of different forest components

Remote Sensing Products

The development of remote sensing sensor systems, both satellite-borne and airborne, and GPS devices are facilitating the enhanced use of remotely sensed data in forest inventories. Landsat Thematic Mapper (TM) has archived the longest data record since 1972 and has a long history of widespread use and acceptance (Powell *et.al.,* 2010). These data products are available for free at the global archive of U.S. Geological Survey online server (http://glovis.usgs.gov/) in a standard processing format at a spatial resolution compatible to the size of FIA field inventory plots (see Figure 3).

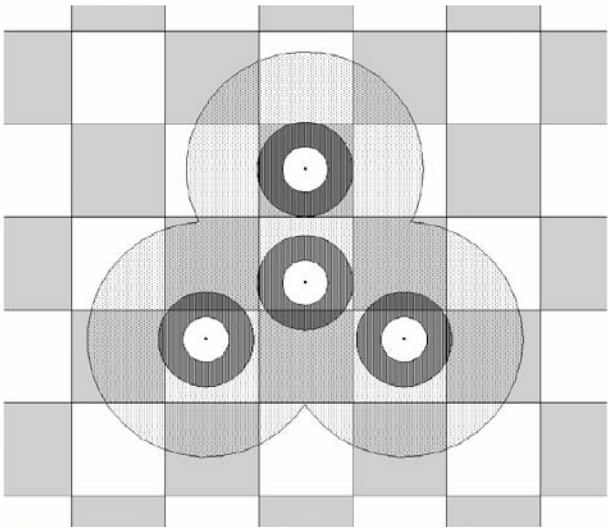

Figure 3. The FIA ground plot geometry versus 30 m TM pixels. The dark grey circles represent the area of locational error due to GPS errors. The larger grey circles represent the potential locational error due to image registration (Hoppus *et.al.* 2000).

Objectives

In this study we mapped regional scale growing stock woody volume and net-growth by coupling four geospatial layers (auxiliary variables), namely normalized difference vegetation index (NDVI) derived from Landsat TM imageries, Digital Elevation Model (DEM), land-cover types (IFMAP) and Basal Area Weighted Height (BAWHT), with forest inventory data of US Forest Service's Forest Inventory and Analysis (FIA) program. We applied the RF method of k-NN imputation to predict forest inventory information of interest across the entire State of Michigan. The RF technique was chosen because it has shown better performance in the prediction of forest inventory parameters as compared to other methods (Breidenbach *et.al.,* 2010; Hudak *et.al.* 2008). The technique is intended to create detailed maps of volume and net-growth with useful precision and a high degree of automation.

The purpose of this study was to generate spatially explicit predictions of growing stock volume, and growth across 47 counties in the northern parts of Michigan, USA. The specific research questions were:

1. How can geospatial and FIA data be efficiently integrated to predict the distribution of the desired forest inventory parameters at a regional scale in a spatially explicit manner?
2. How does the accuracy of forest attribute estimates from imputation technique vary from plot to stand to county level?
3. How reliable are the geospatial data for estimation of the inventory parameters of interest?

Methods

Study Area

The northern Lower Peninsula and entire Upper Peninsula of Michigan, comprising 47 counties, was the area of interest for this research (see Figure 4). The study area was delineated so as to avoid the counties with sparse forest distribution in terms of canopy density, particularly from the southern Michigan. The growing stock volume, growth, mortality and removal data (in standard format) from the extensive FIA data base were the basic response variables considered in the analysis for estimation and mapping purpose in this study. The FIA database has inventory records available at both plot and county level.

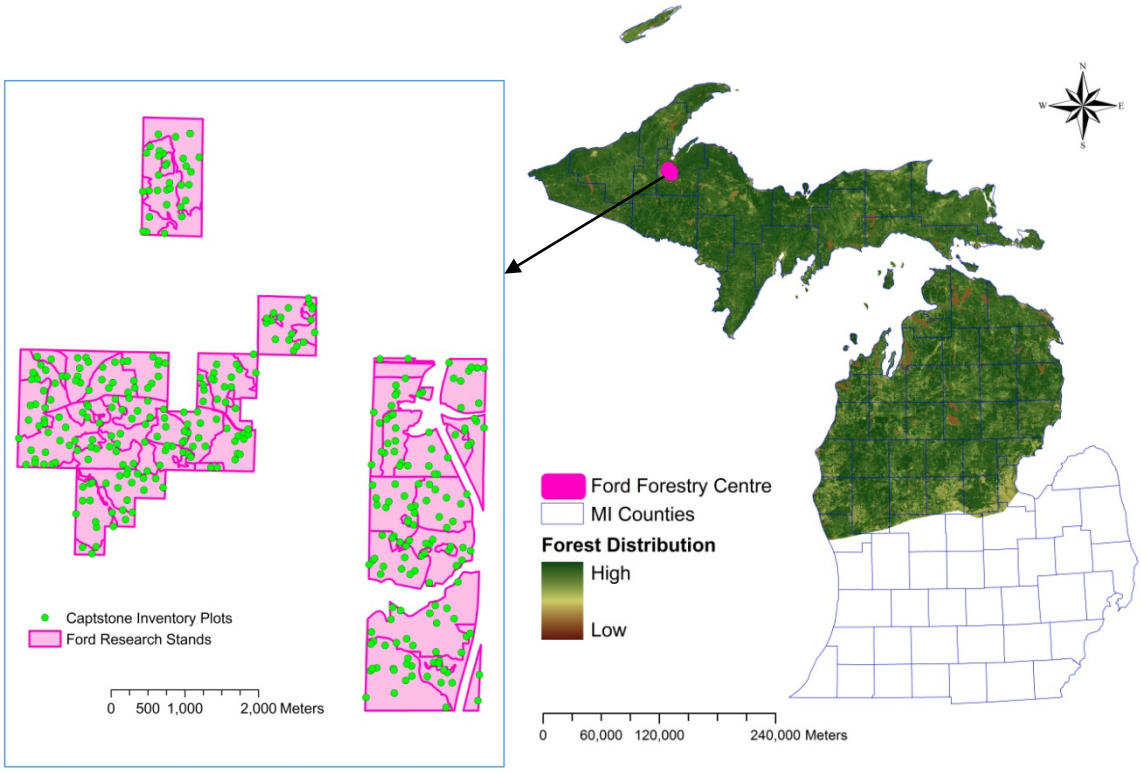

Figure 4. Study area in Michigan, U.S.A.

The geospatial predictor layers considered in the study were Landsat TM derived NDVI, Land Cover data from 2001, digital elevation model (DEM), and basal area weighted canopy height (BAWHT). All these data layers are publicly available and have the same spatial resolution of 30 m. An independent inventory dataset across Michigan Tech's forestland holdings (1,840 ha at Ford Forestry Centre) was also utilized in this study to compare the results of estimates at the plot and stand-level via field inventory and imputation techniques.

Geospatial Data Preparation

NDVI raster

The following Landsat 5 Thematic Mapper (TM) images as in Table 1 were downloaded from the Global Visualization Viewer (http://glovis.usgs.gov/) operated by the Earth Resource Observation and Science Centre (EROS) of the USGS. The images are already processed at the source to Level 1T (L1T) that provides systematic radiometric and geometric accuracy by incorporating ground control points and also employing a Digital Elevation Model (DEM) for topographic accuracy (http://edcsns17.cr.usgs.gov/helpdocs/landsat/product_descriptions.html#terrain_l5_l1t).

Table 1: Landsat imageries used in the study for the derivation of NDVI raster

SN	WRS-2 Path/ Row	Lat/ Long	Acquisition Date	Scan Time	UTM Zone	Sun elevation	Earth-Sun Distance* (d)
1.	20/ 29	44.6/ -82.7	2007-06-11	16:09:52	17	62.86	1.01536
2.	20/ 30	43.2/ -83.2	2008-05-28	16:04:19	17	61.93	1.01355
3.	20/ 31	41.8/ -83.7	2008-07-15	16:03:22	17	61.07	1.01646
4.	21/ 28	46.0/ -83.8	2008-07-06	16:08:37	17	60.11	1.01670
5.	21/ 29	44.6/ -84.3	2008-07-06	16:09:01	16	60.86	1.01670
6.	21/ 30	43.2/ -84.8	2006-06-15	16:15:11	16	63.57	1.01577
7.	21/ 31	41.8/ -85.3	2007-07-20	16:16:21	16	61.46	1.01616
8.	22/ 28	46.0/ -85.3	2007-06-25	16:21:35	16	61.91	1.01652
9.	22/ 29	44.6/ -85.8	2007-06-09	16:22:16	16	62.78	1.01513
10.	22/ 30	43.2/ -86.3	2007-06-09	16:22:40	16	63.59	1.01513
11.	22/ 31	41.8/ -86.8	2008-07-13	16:15:47	16	61.35	1.01655
12.	23/ 28	46.0/ -86.9	2006-07-15	16:27:24	16	59.79	1.01646
13.	23/ 29	44.6/ -87.4	2007-08-03	16:27:47	16	57.21	1.01471
14.	23/ 30	43.2/ -87.9	2007-08-03	16:28:11	16	58.08	1.01471
15.	24/ 27	47.4/ -87.9	2009-08-31	16:29:14	16	47.23	1.00946
16.	24/ 28	46.0/ -88.4	2007-06-23	16:33:59	16	62.00	1.01642
17.	24/ 29	44.6/ -88.9	2010-07-17	16:30:58	16	59.95	1.01635
18.	24/ 30	43.2/ -89.4	2010-07-01	16:31:25	16	62.61	1.01667
19.	25/ 27	47.4/ -89.4	2009-06-03	16:33:57	16	60.05	1.01433
20.	25/ 28	46.0/ -89.9	2009-06-03	16:34:21	16	60.92	1.01433
21.	25/ 29	44.6/ -90.5	2007-08-17	16:39:59	15	53.81	1.01244
22.	26/ 28	46.0/ -91.5	2007-07-07	16:46:10	15	60.93	1.01669

(* earth-sun distance in astronomical units for Day of the Year).

The images used were captured in between the years 2006 and 2010, ranging from June to August, in the peak of growing season of tree species; only the images from June-Aug were considered to reduce the impact of seasonal and phenological variation. The best quality images with zero percent cloud cover were selected for each scene and area of interest (AOI) were determined for each based on visual inspection of false color composite (FCC) in Erdas Imagine 2010 software. The procedure and parameters suggested by Chander *et.al.* (2009) were applied for the conversion of calibrated digital numbers (DNs) to absolute units of at-sensor spectral radiance and top-of-atmosphere (TOA) reflectance. The models for this conversion were built in Erdas Imagine's Modeler. Thus reflectance images were generated and the red and near infrared bands (namely band 3 and 4) were used for the computation of Normalized

Difference Vegetation Index (NDVI) for each of the scenes using model builder in Erdas Imagine 2010. The individual NDVI images were then mosaicked using MosaicPro tool in Erdas Imagine 2010. The overlap function was set to 'feather' and color correction across the NDVI images was made using illumination equalization and image dodging across images (ERDAS Desktop 2010, online help).

Land Cover Raster

The land cover 2001 raster, product of Integrated Forest Monitoring Assessment and Prescription (IFMAP) Project, was downloaded from MI Geographic Data Library <http://www.mcgi.state.mi.us/mgdl/?action=thm>. The land cover was derived by the IFMAP project via the classification of Landsat TM imageries collected between 1997-2001 over three seasons, spring (leaf-off), summer, and fall (senescence), to produce a dataset that can serve multiple functions. The land-cover rasters were available separately for the upper and lower peninsula of Michigan. The original raster had 32 classes; however, it was further reclassified into four broad categories (broadleaved forests, conifer forests, mixed forests and non-forest) for the purpose of this study.

DEM Raster

The digital elevation model (DEM) at 30 m spatial resolution (1 arc second) was downloaded for the study area from the National Map Seamless Server (URL: http://seamless.usgs.gov/website/seamless/viewer.htm) of USGS, Seamless Data Warehouse. The DEM rasters (in *.tiff* format) were downloaded in tiles and then mosaicked in the Erdas Imagine using NN resampling and overlay function.

BAWHT Raster

The basal area weighted height (BAWHT) in a raster format is developed by the Wood Hole Research Centre as a part of the National Biomass and Carbon Dataset (NBCD) 2000. The BAWHT raster (as *.tif* file) is freely available for download from http://atlas.whrc.org/NBCD2000. The raster is produced based on a mapping zone approach in which the conterminous U.S. is split into 66 eco-regionally distinct mapping zones and our study area is labeled as the zone 51. Digital numbers (DN) of the raster represent the average basal area weighted height in meters multiplied by 10. Thus, the average basal area weighted height in meters is DN/10. Development of the dataset is based on an empirical modeling approach that combined FIA sample plot data with high-resolution InSAR data acquired from the 2000 Shuttle Radar Topography Mission (SRTM) and optical remote sensing data acquired from the Landsat ETM+ sensor (Walker *et.al.*, 2007). The BAWHT for an FIA plot is calculated according to:

$$BAWHT = \frac{1}{BA_{plot}} \sum_{i=1}^{n} (BA_i \times ACTUALHT_i)$$

Where BA_i is the basal area (m^2) of the *i*th tree in the plot and is calculated according to:

$$BA = 0.00007854 \times DBH^2$$

(DBH is the diameter at breast height (cm) and $ACTUALHT_i$ is actual height of *i*th tree) and BA_{Plot} is the total basal area (m^2) for the plot and is calculated according to:

$$BA_{Plot} = \sum_{i=1}^{n} BA_i$$

The plot level BAWHT of the FIA database is integrated with the predictor rasters using regression tree modeling approach to derive spatially explicit basal area weighted height raster (Walker *et.al.*, 2007).

All the four predictor rasters and GIS layers used in the study were projected to Michigan GeoRef spatial reference (NAD_1983_Michigan_GeoRef_Meters; UTM_Zone_16N).

Mapping Model Development

Joining Geospatial Predictors to Field Inventory Data

The imputation procedure was implemented with a large number of FIA field sample plots in the reference set. The reason for including a large number of sample plots in the imputation process was to find a better match of nearest neighbors for the un-sampled areas based on the similarity of X values of the selected reference and the target pixels as suggested by LeMay and Temesgen (2005) and also with the hope that bigger sample size would better represent available forest conditions in the area of interest. Altogether 6,702 FIA inventory plots data (throughout the Michigan) were obtained for the reference set that included response variables such as growing stock volume, growth, mortality and removals on per acre basis (also by species group code).

In order to increase the sample size a compromise was needed in order to pass FIS security clearances. This was achieved by regrouping the NDVI and Land Cover rasters into broad classes instead of using in a continuous format. Because of the privacy and security restrictions of FIA, it was not possible to obtain field coordinates of the FIA plots. However, the generosity of FIA unit at the Northern Research Station made our life easy by attaching the pixel values of each predictor raster layers to the FIA field plots. But this task of linking the FIA field plot attributes to the raster values was possible only after retaining a limited number of unique values for the combinations of the raster layers and that needed grouping (classification) of the NDVI and Land Cover rasters. Since users of FIA data are could track an individual FIA plot on the ground (which is against the privacy requirements specified in the amendments to the U.S. Food Security Act of 1985) if each of the plots have a unique combination of raster values, we had to group the NDVI and land cover rasters into broader classes.

The continuous NDVI raster was classified into 20 classes (using natural breaks in Arc Map 9.3) and the land-cover raster was reclassified into four broader classes namely broadleaved, conifers, mixed and non-forests. The predictor raster layers were then sent to the FIA unit at the Northern Research Station where the plot coordinates were intersected with the spatially referenced predictor rasters and the corresponding raster values were associated with each of the sample plots. In fact, we sent only the classified NDVI and land-cover rasters to the FIA unit in an anticipation to obtain as many sample plots data as possible after passing the security restrictions of FIA; the ground elevation and BAWHT data assigned to the plots were the direct values as measured by FIA. That means the elevation and BAWHT used in the training (reference) set for model building were from FIA while the values of target points for spatially extending the model for the study area were from the rasters. We simply assumed that the plot elevation and BAWHT values provided by FIA were the same as the corresponding values in the spatially referenced raster layers. We do acknowledge that since plot level elevation and BAWHT from FIA data may not

exactly match with the corresponding values in the geo-referenced rasters, there is likely bias associated at this stage. A simple flow diagram of the methods employed in the study is described in the Figure 5 below.

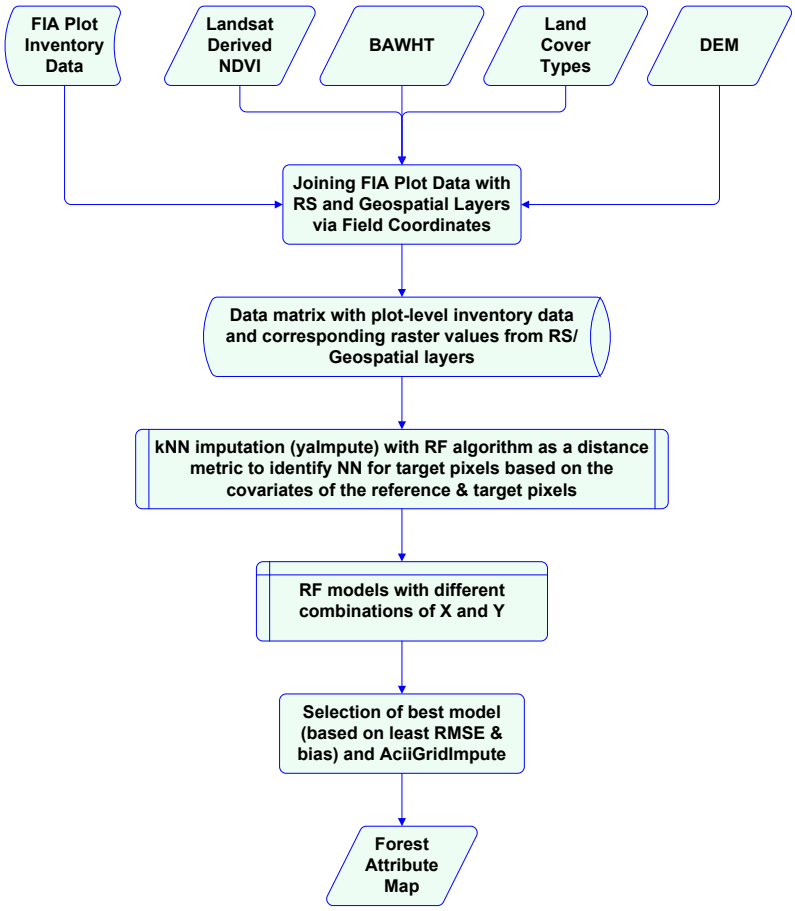

Figure 5: Flow diagram of the methods of study

Model Training and Prediction

The sample plots in the reference set are assumed to characterize the entire range of variability in the predictor layers, though the used predictor layers itself were not sufficient to fully describe the heterogeneous forest conditions of this regional study. Since imputation with a single nearest neighbor produces output with similar variance structure to that of the observations, the value of k (number of nearest neighbors) was set equal to one to maintain the variance structure of forest attributes in the imputation process.

To scale the plot level FIA data to maps, we developed RF imputation models relating field-measured response variables to plot level raster value as the predictor variables. As already mentioned NDVI, DEM, land-cover and BAWHT were the predictor variables, however the latter was not included while producing forest attributes distribution map because BAWHT data was available only for limited number

of plots. The yaImpute package in R statistical software (www.r-project.org) was used to implement the random forest algorithm (Crookston and Finley, 2008) to impute growing stock volume, growth, mortality and removal for whole of the study area. The yaImpute works with *.csv* file format of the reference data matrix and determines nearest neighbors for each of the target points based on spectral space distance using auxiliary variables of the reference and target points. The feature space distance considered in this study was based on random forest method. The random forest mode was set to regression, number of trees was fixed at 2000 and $k = 1$. After determining the best random forest model based on least statistical errors, the model was extended spatially using the Ascii Grids of the predictor layers for target area. Since the AsciiGridImpute function takes long time to generate spatial prediction, the imputation process was carried out by splitting the predictor rasters into a dozen tiles and running the imputation model separately for each tile. In this step, each of the target pixels are assigned an identifier from the reference data set and the response variable for the target pixel will be the same as of the assigned identifier.

Model Validation

For validation purpose of the imputation results, a set of data for growing stock volume, growth, mortality and removal was retrieved from the FIA database (FIADB) using EVALIDator program where these data are available for forestland at county-level. This reference set was used for validation of the imputation results for each of the 47 counties. In order to compare the estimates of the forest attributes (particularly volume) at plot and stand level, the growing stock volume data from the Capstone Forestry Project of Michigan Tech was utilized. The Capstone dataset has inventory records, only for the growing stock volume, for 358 permanent plots and 50 stands.

We have used the RMSE and biases (calculated from the difference between imputed and observed values) as a measure of predictive performance of yaImpute i.e. prediction accuracy of the RF imputation, in this study.

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

$$Bias = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)}{n}$$

Results

Three RF models were developed with the common predictors NDVI, DEM and land-cover. The root-mean square errors (RMSE) with and without BAWHT as an additional predictor in the three models, are given in the Tables 2, 3 and 4. The first model predicts volume and % hardwood; second model predicts volume, %hardwood and growth; and the third model predicts volume, %hardwood, growth and mortality. These model errors are comparable to the estimation error reported by Holmstrom and Fransson (2003), Tomppo *et.al.* (2002) and Nellson (1997) for wood volume using kNN technique (RMSEs of 50 m^3 ha^{-1}, 46 m^3 ha^{-1} and 56 m^3ha^{-1} are reported respectively).

Table 2. Error statistics of the random forest model predicting total volume and % hardwood

	RMSE Without BAWHT	RMSE With BAWHT	R^2 Without BAWHT	R^2 With BAWHT
Volume (m^3)	31.71	26.53	0.2444	0.4103
% Hardwood	33.32	30.16	0.4673	0.5470

Table 3. Error statistics of the random forest model predicting total volume, % hardwood and net growth of growing stock

	RMSE Without BAWHT	RMSE With BAWHT	R^2 Without BAWHT	R^2 With BAWHT
Volume (m^3)	32.36	27.38	0.2232	0.3855
% Hardwood	34.43	31.05	0.4394	0.5233
Growth (m^3yr^{-1})	1.19	1.27	0.0573	0.0268

Table 4. Error statistics of the random forest model predicting total volume, % hardwood, net growth and mortality of growing stock

	RMSE Without BAWHT	RMSE With BAWHT	R^2 Without BAWHT	R^2 With BAWHT

Volume (m^3)	32.74	28.02	0.2091	0.3663
% Hardwood	34.77	31.51	0.4289	0.5116
Growth (m^3yr^{-1})	1.17	1.26	0.0633	0.0279
Mortality (m^3yr^{-1})	0.68	0.75	0.0512	0.0159

The four response variables (viz. volume, growth, mortality and removals) were imputed for each of the forty-seven counties selected in Michigan and were compared with the independent reference data obtained from FIA database using the EVALIDator tool. The validation statistics are given in the Table 5. The comparisons of imputation estimates of the parameter with the reference data set at the county level are presented separately in the scatter plots in Figures 6A-6D.

Table 5. Statistics of county level forest parameters comparison

Parameter	**R^2**	**RMSE**	**Relative RMSE**	**Bias**	**Relative Bias**
Total Growing Stock Volume (m^3)	0.8940	2686069.42	17.99%	-365700.67	-2.45%
Growth (m^3yr^{-1})	0.7747	81248.23	30.40%	-49669.62	-18.59%
Mortality (m^3yr^{-1})	0.7270	50834.26	42.70%	-12394.02	-10.41%
Removals (m^3yr^{-1})	0.6843	1865.25	47.89%	-787.93	-20.23%

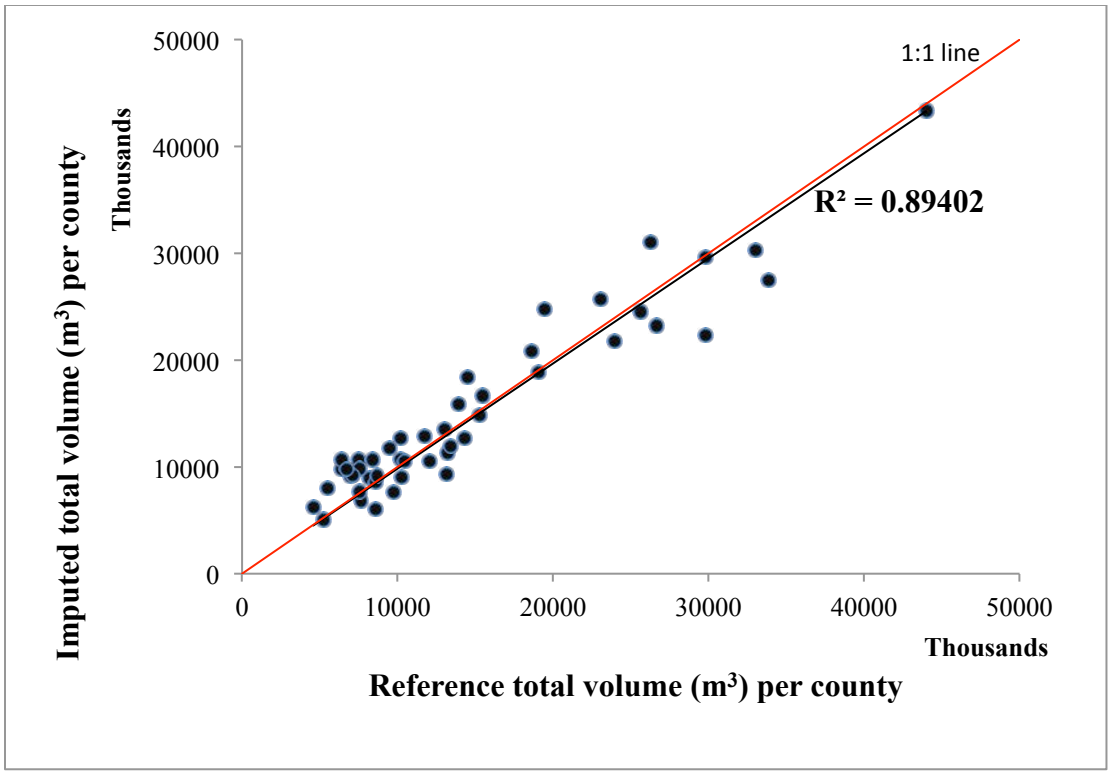

Figure 6A. Scatter plot of county level imputed volumes (m^3) against the reference volumes obtained from FIA database.

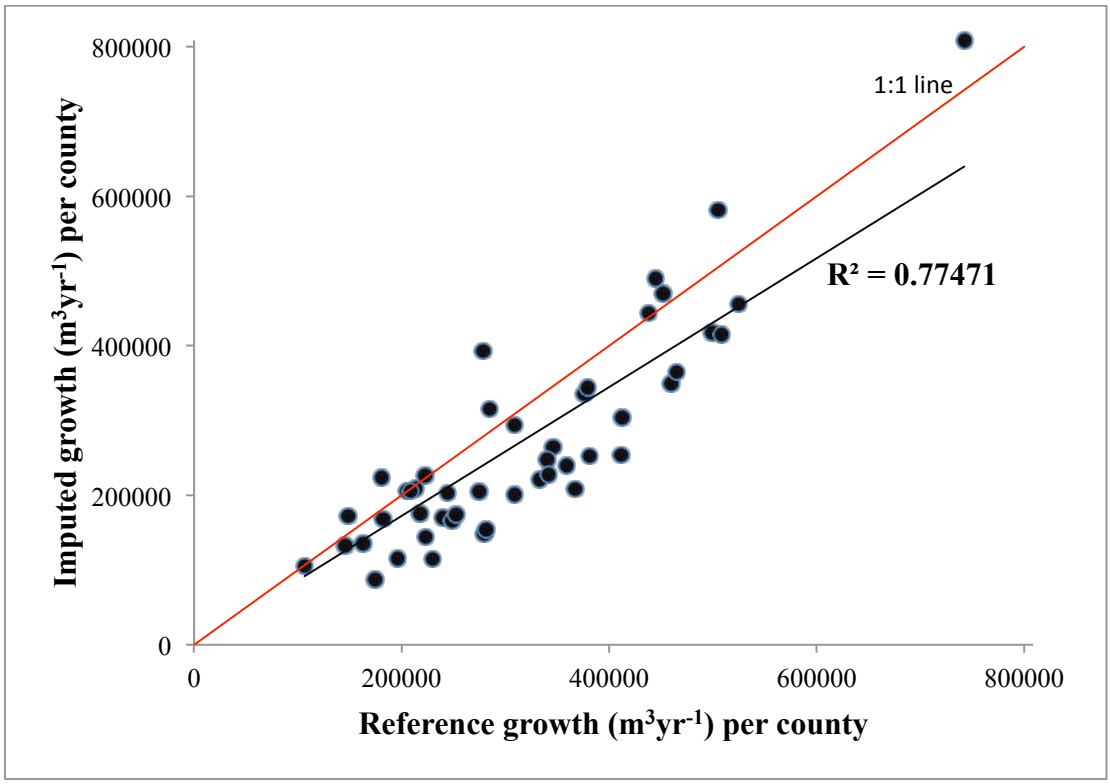

Figure 6B. Scatter plot of county level imputed growths (m^3yr^{-1}) against the reference growths obtained from FIA database.

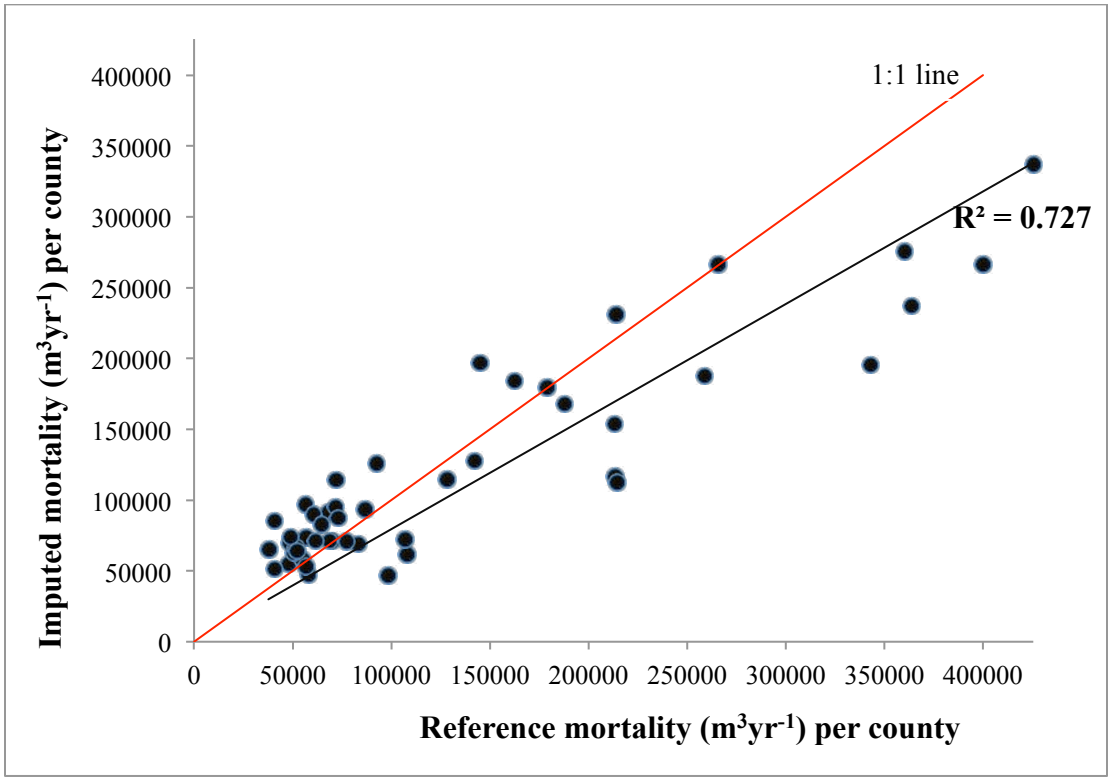

Figure 6C. Scatter plot of county level imputed mortalities (m^3yr^{-1}) against the reference mortalities obtained from FIA database.

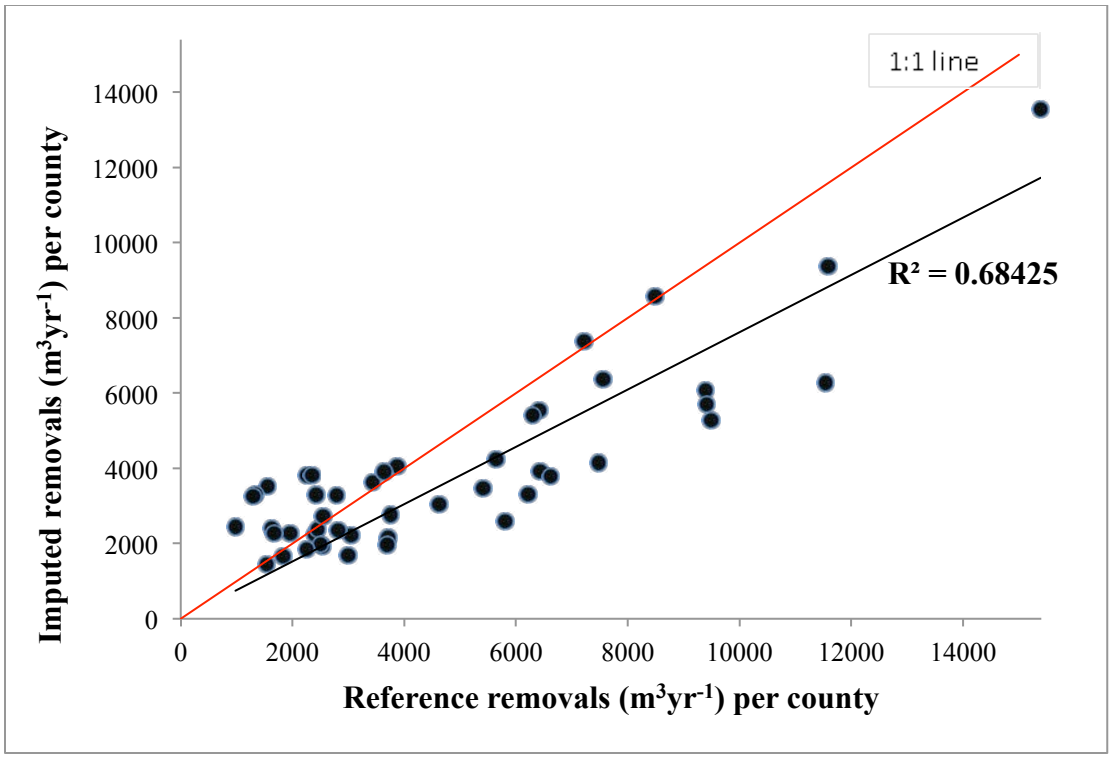

Figure 6D. Scatter plot of county level imputed removals (m^3yr^{-1}) against the reference removals obtained from FIA database.

The validation of stand level volumes (m^3) against the reference volumes obtained from the Capstone project has also shown encouraging results (see Figure 7 A), though substantial negative bias is obvious in the prediction (bias: -1083.43 m^3; relative bias: -46.27%). However, we could not compare the imputed growth, mortality and removals at the stand level because of unavailability of such reference data (in the Capstone project). As expected, the plot level comparisons of imputed and measured forest parameters were worst (see Figures 7B) and the possible reasons are described in discussion section below.

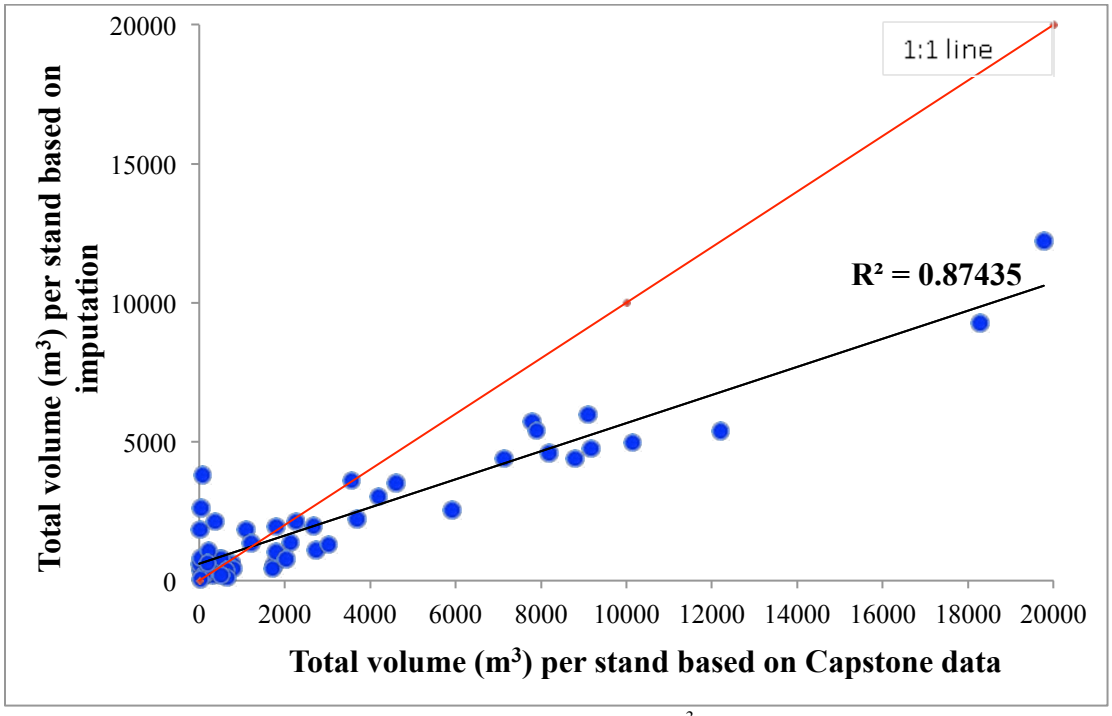

Figure 7A. Scatter plot of stand level imputed volumes (m^3) against reference volumes from the Capstone project.

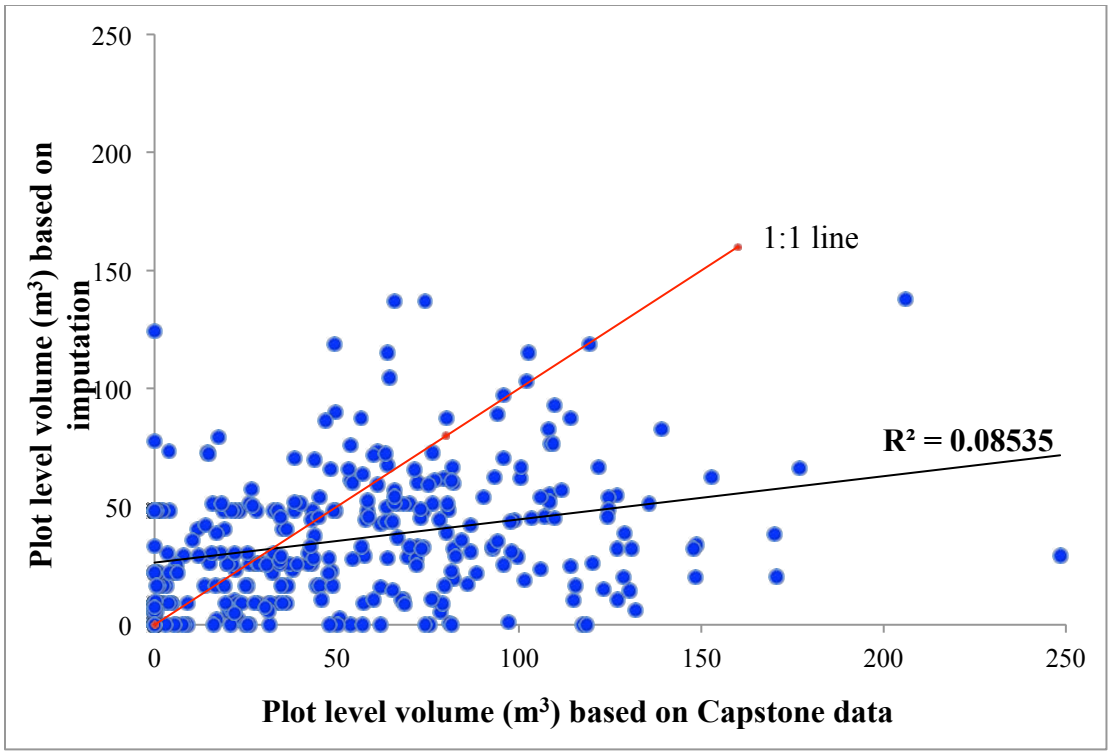

Figure 7 B. Comparison of plot level volumes (m^3) estimated from the imputation technique and Capstone project inventory.

Discussion

The results observed in the study are consistent with previous studies. Although we selected k=1 to retain the variability of field attributes in the imputation mapping, the pixel-plot level accuracy of estimates was least and county level estimates were the best. Other studies have also found that pixel level accuracy of forest attribute estimations using kNN is low, but for larger areas more acceptable accuracy is reached (Holmstrom, 2003; McRoberts *et.al.,* 2007; Nilsson, 2002; Tomppo *et.al.,* 2002). For example, Reese *et.al.*, (2002) found that accuracy of the kNN estimates for all forest parameters was low at the pixel level (RMSE for total wood volume ranged from 58–80%), however, better accuracy was achieved over larger areas, with best results being 10% RMSE over a 100 ha aggregation.

The weak association between pixel level measured and imputed values can also be justified on the basis of error in spatial referencing of pixels, error in GPS coordinates of inventory plots and also the design of inventory plots. For example, the layout of the FIA plots is such that the four 24-foot-radius subplots are spread over a minimum of 4 pixels (see Figure 3). The plot level per acre values for the attributes supplied by FIA (used in this study) is based on avenging and up-scaling of the values from the four sub-plots. Therefore, there is not a clear one-to-one relationship between spectral values of Landsat pixel and corresponding FIA plot data. Additionally, the year of FIA plot measurements did not exactly match the year of Landsat image acquisition. In fact, the best available Landsat scenes (used for developing NDVI mosaic), over the years 2006-2010, from the peak of growing season (June to September) were considered to avoid the nuisance of cloud cover.

The county-level data for the four attributes (viz. volume, net growth, removal and mortality) retrieved (for validation purpose) directly from the FIA database using EVALIDator tool, were the estimates of growing-stock on forestland only that includes timberland, reserved forest land, and other forest land. The forestland according FIA definition is any land having at least 10 % crown cover. The FIA plots, according to current design, can also include areas having <10% crown cover (i.e. non-forest) and the imputation technique also considerers such plots and gives prediction for both forested and non-forested pixels. However, in a surprising way our results show that the total of the imputed values for most of the counties are below the reference values (obtained from EVALIDator) as shown by the distribution of points below the 1:1 line in the Figures 6 and 7. This means that the imputation models are under predicting the forest parameters. This is something that we are still working on by excluding non-forest lands and including more predictor layers in the ongoing imputation process. Use of mask to exclude non-forest area and running imputation only for the forested region could possibly improve prediction and also reduce imputation time. This will be carried out in the upcoming version of Forest Biomass Information System (FBIS).

Since the study required grouping of NDVI and land-cover rasters in broader classes because of the security issues of FIA, the predictive power of the auxiliary layers was obviously less than it would have been without the grouping. The inclusion of BAWHT though improved the prediction accuracy significantly, our team decided not to use this variable while mapping the forest parameters because the height data was available for fewer numbers of plots in the study area. Further, the BAWHT and elevation values attached to the available plots in the reference data matrix were not extracted from the raster layers (we supplied to FAI), rather these values were the results of field measurements made by FIA at the inventory plots. So there is likely inconsistency between the raster values and the plot-level values for

these two variables and hence also bias in the estimation. Our team has further contemplated to include a disturbance element as an additional predictor to better estimate the forest attributes, particularly net growth, in the final version of imputation mapping. We are going to use time-series NDVI imageries from MODIS sensor to develop a raster layer as a predictor of disturbance in the upcoming edition of FBIS.

This research did not consider the differences in vegetation zones over the large study area. Operational application of the kNN method has also shown that the predictions may be biased if the area of interest is large and covers several vegetation zones with different tree species compositions (Tomppo, 2006). The biases can be reduced if the set of potential nearest neighbors can be restricted to strata by classifying the area into different vegetation zones. This is something to look at in future research.

Literature Cited

Bernier, P.Y.; Daigle, G.; Rivest, L.P.; Ung, C.H.; Labbe, F.; Bergeron, C; Patry, A. From plots to landscape: A k-NN based method for estimating stand-level merchantable volume in the Province of Quebec, Canada. The Forestry Chronicle, Vol. 86, No. 4, pp. 461-468.

Breidenbach, J.; Nothdurft, A.; Kandler, G. 2010. Comparision of nearest neighbor approaches for small area estimation of tree species-specific forest inventory attributes in central Europe using airborne laser scanner data. Eur. J. Forest Res 129: 833-846.

Breiman, L.; Cutler. A. 2004. Random Forests. [Online] http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm (Accessed on March 5, 2011).

Breiman L. 2001. Random Forests. *Machine Learning*, 45(1), 5-32.

Burkman, B. 2005. Forest Inventory and Analysis, Sampling and Plot Design. FIA Fact Sheet Series, 2/3/05. Available at: http://fia.fs.fed.us/library/fact-sheets/data-collections/Sampling%20and%20Plot%20Design.pdf

Chander, G.; Markham, B.L.; Helder, D.L. 2009. Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+ and EO-1 ALI sensors. Remote Sensing of Environment 113: pp. 893-903.

Crookston, N.L. and Finley, A.O. 2008. yaImpute: an R package for kNN imputation. *Journal of Statistical Software*, 23(10), 1-16.

Cutler, D.R., Edwards, T.C. Jr., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J. and Lawler, J.J. 2007. Random Forests for Classification in Ecology. *Ecology*, 88(11): pp. 2783-2792.

Eskelson, B.N.I.; Temesgen, H.; LeMay, V.; Barrett, T.M.; Crookston,N.L.; Hudak, A.T. 2009. The role of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. Scandinavian Journal of Forest Research 24: pp. 235-246.

Falkowski, M.J., Evans, J.S., Martinuzzi, S., Gessler, and P.E., Hudak, A.T. 2009. Characterizing forest succession with lidar data: an evaluation for the inland northwest, USA. *Remote Sensing of Environment*, 113: pp. 946-956.

Falkowski, M.J. 2008. Improving conifer forest inventory and assessment with discrete return LiDAR data. Ph.D. dissertation submitted to University of Idaho.

Falkowski, M.J.; Hudak, A.T.; Crookston, N.L.; Gessler, P.E.; Uebler, E.H.; Smith, M.S. 2010. Landscape-scale parameterization of a tree-level forest growth model: a *k*-nearest neighbor imputation approach incorporating LiDAR data. Can. J. For. Res. 40: 184-199.

Franco-Lopez, H., Ek, A.R. and 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: pp. 251-274.

Grossmann, E.B.; Ohmann, J.L.; Gregory, M.J.; and May, H.K. 2009. Nationwide Forest Imputation Study (NaFIS)- Western Team Final Report. Final Report of the Nationwide Forest imputation Study (NaFIS) available at http://www.fsl.orst.edu/lemma/main.php?project=common&id=publications&ref=nafis

Haapanen, R. and Ek, A.R. 2001. Software and instructions for kNN application in forest resources description and estimation. Staff Paper Series Number 152. Department of Forest Resources, College of Natural Resources and Minnesota Agricultural Experiment Station, University of Minnesota, St. Paul, Minnesota.

Haapanen, R.; Lehtinen,K.; Miettinen, J.; Bauer, M.E.; and Ek, A.R. 2002. Progress in adapting k-NN methods for forest mapping and estimation using the new annual Forest Inventory and Analysis data. In: McRoberts, R.E.; Reams, G.A.; Van Deusen, P.C.; Moser, J.W., eds. Proceedings of the Third Annual Forest Inventory and Analysis Symposium; Gen. Tech. Rep. NC-230. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station: 87-95. Available online at http://nrs.fs.fed.us/pubs/4432

Holmstrom, H; Fransson, J.E.S. 2003. Combining remotely sensed optical and radar data in *k*NN-estimation of forest variables. Forest Science 49(3): pp. 409-418.

Hoppus, M.L.; Lister, A.J. 2002. A statistically valid method for using FIA plots to guide spectral class rejection in producing stratification maps. In: McRoberts, R. E.; Reams, G. A.; Van Deusen, P. C.; Moser, J. W., eds. Proceedings of the Third Annual Forest Inventory and Analysis Symposium; Gen. Tech. Rep. NC-230. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station: 44-49.

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 Sensing of Environment: pp. 2232-2245.

Katila, M. and Tomppo, E. 2001. Selecting estimation parameters for the Finnish multisource National Forest Inventory. Remote Sensing of Environment 76: pp. 16-32.

LeMay, V.; Temesgen, H. 2005. Comparision of nearest-neighbor methods for estimating basal area and stems per hectare using Aerial auxiliary variables. Forest Science 51(2): pp. 109-119.

Liaw, L.A. and Wiener, M. 2002. Classification and regression by randomForest. *R News*, 2(3), 18-22. URL http://CRAN.R-project.org/doc/Rnews/.

McRoberts, R.E. 2009. A two-step nearest neighbors algorithm using satellite imagery for predicting forest structure within species composition classes. Remote Sensing of Environment. 113: pp. 532-545.

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: pp. 457-468.

McRoberts, R.E.; Tomppo, E.O.; Finley, A.O.; Heikkinen, J. 2007.Estimating areal means and variances of forest attributes using the *k*-nearest neighbors technique and satellite imagery. Remote Sensing of Environment 111: 466-480.

Moeur, K.S.; Coble, D.W.; McMahan, A.L.; Smith, E.L., 1995. Most Similar Neighbor- an improved sampling inference procedure for natural resource planning. Forest Science 41: pp. 337-359.

Moeur and Stage, 1995. Most Similar Neighbor- an Improved Sampling Inference Procedure for Natural Resource Planning. Forest Science 41: pp. 337-359.

Nilsson, M. 2002. Deriving nationwide estimates of forest variables for Sweden using Landsat ETM+ and field data. Department of Forest Resources Management and Geomatics. Swedish University of Agricultural Sciences. Paper presented at ForestSAT Symposium, Heriot Watt University, Edinburgh.

Ohmann, J.L. and 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. Can. J. For. Res. 32: pp. 725-741.

Ohmann, J.L.; Gregory, M.J.; Henderson, E.B.; Roberts, H.M. 2011. Mapping gradients of community composition with nearest-neighbour imputation: extending plot data for landscape analysis. Journal of vegetation Science, 22(4): pp. 660-676.

Pang, H., Lin, A., Holford, M., Enerson, B.E., Lu, B., Lawton, M.P., Floyd, E., and Zhao, H. 2006. Pathway analysis using random forests classification and regression. *Bioinformatics* 22(16): pp. 2028-2036.

Powell, S.L.; Cohen, W.B.; Healey, S.P.; Kennedy, R.E.; Moisen, G.G.; Pierce, K.B.; Ohmann, J.L. 2010. Quantification of live aboveground forest biomass dynamics with Landsat time-series and field inventory data: A comaparision of empirical modeling approaches. Remote Sensing of Environment 114: pp. 1053-1068.

Reese, H.; Nilsson, M.; Sandstrom, P.; and Olsson, H. 2002. Applications using estimates of forest parameters derived from satellite and forest inventory data. Computers and Electronics in Agriculture 37(1): 37-55.

Stage, A.R. and Crookston, N.L. 2007. Partitioning error components for accuracy assessment of near-neighbor methods of imputation. *Forest Science*, 53(1), 62-72.

Tomppo, E. 2006. The Finnish Multisource National Forest Inventory: Small-Area Estimation and Map Production. Proceedings of the Eighth Annual Forest Inventory and Analysis Symposium.

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: pp. 156-171.

Walker, W.S.; Kellndorfer, J.M.; LaPoint, E.; Hoppus, M.; Westfall, J. 2007. An empirical InSar-optical fusion approach to mapping vegetation canopy height. Remote Sensing of Environment 109: pp. 482-499.

Woudenberg, S.W.; Conkling, B.L.; O'Connell, B.M.; LaPoint, E.B.; Turner, J.A.; Waddell, K.L. 2010. The Forest Inventory and Analysis Database: Database description and users' manual version 4.0 for Phase 2. Gen. Tech. Rep. RMRS-GTR-245. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 339 p.

Yang, B.S.; Di, X.; Han, T. 2008. Random forests classifier for machine fault diagnosis. Journal of Mechanical Science and Technology, 22: pp. 1716-1725.