2014

APPLICATION OF AN IMPUTATION METHOD FOR GEOSPATIAL INVENTORY OF FOREST STRUCTURAL ATTRIBUTES ACROSS MULTIPLE SPATIAL SCALES IN THE LAKE STATES, U.S.A.

Ram K. Deo
Michigan Technological University

Recommended Citation

Deo, Ram K., "APPLICATION OF AN IMPUTATION METHOD FOR GEOSPATIAL INVENTORY OF FOREST STRUCTURAL ATTRIBUTES ACROSS MULTIPLE SPATIAL SCALES IN THE LAKE STATES, U.S.A.", Dissertation, Michigan Technological University, 2014.
https://digitalcommons.mtu.edu/etds/808

Follow this and additional works at: https://digitalcommons.mtu.edu/etds

Part of the Forest Sciences Commons, Natural Resources Management and Policy Commons, and the Remote Sensing Commons

APPLICATION OF AN IMPUTATION METHOD FOR GEOSPATIAL INVENTORY OF FOREST STRUCTURAL ATTRIBUTES ACROSS MULTIPLE SPATIAL SCALES IN THE LAKE STATES, U.S.A.

By

Ram K. Deo

A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In Forest Science

MICHIGAN TECHNOLOGICAL UNIVERSITY

2014

This dissertation has been approved in partial fulfillment of the requirements for the Degree of DOCTOR OF PHILOSOPHY in Forest Science.

School of Forest Resources and Environmental Science

Dissertation Co-Advisor:	*Robert E. Froese*
Dissertation Co-Advisor:	*Michael J. Falkowski*
Committee Member:	*David D. Reed*
Committee Member:	*David W. Watkins*
School Dean:	*Terry Sharik*

Table of Contents

List of Figures

List of Tables

List of Appendices

Preface

This dissertation consists of one overview and three major chapters. The major chapters are already in a standard paper format and will soon be published in peer reviewed journals. The dissertation author, Ram K. Deo, is the principal investigator who conducted all analysis, produced all figures and tables, and accomplished the writing. The co-advisors, Robert E. Froese and Michael J. Falkowski, contributed in refining research questions, methodology, and editing of the individual chapters. Ram K. Deo will be the first author of all the major papers to be published out of the dissertation, and the dissertation co-advisors, will stand as co-authors for each of the papers. Since no material has been submitted for publication, no chapter is under any copyright.

Acknowledgements

I am grateful to the School of Forest Resources and Environmental Science (SFRES) at Michigan Tech for letting me experience the top schooling environment with cooperative faculty and staff. I am humble to the Michigan Economic Development Corporation, U.S. Department of Energy, NASA, and US Forest Service, Forest Management Service Centre (FMSC) for supporting the three major research chapters of my dissertation. I am thankful to the Dean Dr. Terry Sharik and Asst. Dean Dr. Andrew Storer of SFRES for research assistantship support during the final semester of my stay at the SFRES.

I express my sincere gratitude to my supervisors, Dr. Robert E. Froese and Dr. Michael J. Falkowski, for their continuous support, encouragement and helpful comments from the designing of research proposal till the completion of this dissertation. I thank Dr. Froese for arranging funds and facilitating data request with the national inventory programs (FIA, and BIA) and FMSC for model calibration and validation. Dr. Falkowski arranged the funding and opened the door of LiDAR remote sensing to assist my genuine interest; I thank him for his effective guidance on technical writing and questions regarding GIS and remote sensing based modeling.

I owe sincere thankfulness to my research committee members, Dr. David D. Reed and Dr. David W. Watkins, for insightful comments and suggestions during my proposal defense, comprehensive exams, and other meetings. This helped me improve my research questions and methodological skills. I am also thankful to Dr. Erin Smith-Mateja of US Forest Service FMSC-FVS Group; Dr. Richard A. McCullough of US Forest Service, Northern Research Station; and Dr. Joseph Mortzheim of the Bureau of Indian Affairs, Midwest Regional Office for their assistance with the data for model calibration and validation.

I would like to thank Timothy Gebuhr for the hard work during my field data collection at the Ford Forestry Center of the Michigan Tech. I am highly indebted to Nan Pond and Michael Premer for the assistance with overstory validation datasets of the northern hardwood and aspen stands. I express my sincere appreciations to Aaron Poznanovic and Karl Meingast for the help during data analysis in the R statistical software, and also to Michael Hyslop, James Rivard, and Shawn O'Neil for GIS assistance. I would also like to thank all my lab-mates and colleagues who contributed in my study by providing moral and material support. Thanks to the Nepalese community at Michigan Tech for organizing frequent get-togethers in a cultural environment with native foods. I must also thank Erin Froese, Scott Hillard, Bridget Hillard, Gary Ewert, and Marianna Ewert for the intimacy and various supports.

Last but not least, I would like to thank the entire SFRES/MTU family who directly or indirectly contributed in my happy stay at Houghton. Extra thanks go to Linda Nagel, Debra Charlesworth, Jill Fisher, Ruth A Ojala, Andrea Longhini, Marjorie Lindley, and Nancy Byers Sprague for the academic and administrative supports.

I also thank my wife Shruti for continuous and unconditional support since we are married. I am proud of my papa, mom, brothers, sisters and other family members for their continuous support in my career building. I humbly dedicate this work to my parents and teachers who inspired me to study forestry for the benefits of nature and society.

Abstract

Credible spatial information characterizing the structure and site quality of forests is critical to sustainable forest management and planning, especially given the increasing demands and threats to forest products and services. Forest managers and planners are required to evaluate forest conditions over a broad range of scales, contingent on operational or reporting requirements. Traditionally, forest inventory estimates are generated via a design-based approach that involves generalizing sample plot measurements to characterize an unknown population across a larger area of interest. However, field plot measurements are costly and as a consequence spatial coverage is limited. Remote sensing technologies have shown remarkable success in augmenting limited sample plot data to generate stand- and landscape-level spatial predictions of forest inventory attributes. Further enhancement of forest inventory approaches that couple field measurements with cutting edge remotely sensed and geospatial datasets are essential to sustainable forest management. We evaluated a novel Random Forest based *k* Nearest Neighbors (RF-kNN) imputation approach to couple remote sensing and geospatial data with field inventory collected by different sampling methods to generate forest inventory information across large spatial extents. The forest inventory data collected by the FIA program of US Forest Service was integrated with optical remote sensing and other geospatial datasets to produce biomass distribution maps for a part of the Lake States and species-specific site index maps for the entire Lake State. Targeting small-area application of the state-of-art remote sensing, LiDAR (light detection and ranging) data was integrated with the field data collected by an inexpensive method, called variable plot sampling, in the Ford Forest of Michigan Tech to derive standing volume map in a cost-effective way. The outputs of the RF-kNN imputation were compared with independent validation datasets and extant map products based on different sampling and modeling strategies. The RF-kNN modeling approach was found to be very effective, especially for large-area estimation, and produced results statistically equivalent to the field observations or the estimates derived from secondary data sources. The models are useful to resource managers for operational and strategic purposes.

1. Introduction to the dissertation: Imputation for geospatial inventory of forest structural attributes at multiple spatial scales

1.1. Background

Spatially explicit inventory of forest structural attributes and site productivity is invaluable for informing strategic planning and proactive management of forests that face increasing demands for products and services such as bioenergy feedstock and carbon sequestration. However, exhaustive field measurement under a design-based framework (i.e., using only sample plots) for resource assessment across large spatial extents is implausibly expensive. An alternative is to couple remotely sensed data with sparse ground-sampled forest inventory data to extend sample plot measurements through both space and time. Spatial modeling algorithms using remotely sensed and other geospatial datasets have been recognized in intergovernmental initiatives towards climate change mitigation, such as the United Nations Framework Convention on Climate Change (UNFCCC, 1992). Such initiatives require estimation and verification of forest biomass which is an efficient apparatus (sink) to sequester rising level of atmospheric carbon dioxide. An example includes the global remote sensing survey of forest cover change by the United Nations Food and Agricultural Organization (FAO, 2010; D'Annunzio et al., 2014). Forest biomass mapping and inventory updates provide information on production, availability, and distribution which also support understanding of the role of forest ecosystems as carbon sinks (Powell et al., 2010; Zhang and Ni-meister, 2014). Accurate and practical spatial models are needed to assess status and trend of forest resources resulting from various management practices (Zheng et al., 2007; Song, 2012). Forest managers can apply an efficient spatial model to generate a baseline and biomass accrual information to gain economic incentives such as under the United Nations collaborative program on reducing emissions from deforestation and forest degradation (UN-REDD, 2010). High resolution wall-to-wall maps of inventory attributes facilitate managers to design and implement ecologically sound, economically viable, and socially acceptable forestry projects. Local-scale inventory information is essential for operational management such as harvest scheduling while regional-scale inventory supports strategic planning (McRoberts et al., 2007; Brosofske et al., 2014) such as an optimal site selection for a biofuel plant.

The operational inventory systems require methods of local and regional relevance (scope) with the qualities of cost-efficiency and reliability (Anaya et al., 2009). The reliability and accuracy of remote sensing methods for small-or large-areas inventory depends on quality and availability of spatially continuous auxiliary data along with representative ground registered sample plot data and a robust algorithm for modeling (Lu et al., 2012). High resolution remotely sensed data are required to accurately quantify biophysical attributes for operational management at local scale (Hudak et al., 2008; Falkowski et al., 2010) while coarse resolution data are mostly used for large-area estimation suitable for strategic planning (Brosofske et al., 2014). For example, MODIS data at spatial resolutions of 250 m or 500 m have been used for national or continental scale mapping (Baccini et al., 2004; Zheng et al., 2007; Blackard et al., 2008; Anaya et al., 2009) while Landsat data at spatial resolutions of 30 m have been used for local or regional scale mapping (Hall et al., 2006; Labrecque et al., 2006; Luther et al., 2006; Powell et al., 2010). By configuration, high spatial resolution of space born sensors is, however, associated with low temporal and radiometric resolutions. Conversely, coarse spatial resolution data generally have high temporal and radiometric resolutions but suffer from the impairing effect of mixed digital signatures in large pixels (Huete et al., 2002; Muukkonen and Heiskanen, 2007). The spatial mismatch between the plot size of ground data and pixel size of coarse optical data is also a serious issue in spatial modeling. However, this issue can be tackled by using high spatial resolution data (e.g. Landsat) as an intermediary feature to establish an empirical model with the field measured variable and then coarse resolution data (e.g. MODIS) can be used to spatially extend the model after radiometric calibration of image bands (Muukkonen and Heiskanen, 2007; Zheng et al., 2007; Wulder et al., 2012). The strategic *in situ* data collected by the national forest inventory and analysis (FIA) program of the US Forest Service have been historically used as a reference frame to produce estimates of forest structural attributes over large geographic area. The nationwide annual inventory system of FIA, operated since 1999, provides a comprehensive dataset to describe status and trend of forest resources over all forest types across the country. For any un-sampled small-area, a tactical inventory may employ algorithms that combine sample plot data from outside the area of interest with ancillary remotely sensed data that prevail explicitly over both sampled and unsampled area of interest (Yim et al., 2011).

The feature metrics derived from several air- or space-borne spectral sensors, either passive or active, are used in spatial inventory modeling. Optical, RADAR and LiDAR remote sensing data are commonly applied in biomass mapping (Zhang and Ni-meister, 2014). Passive optical data particularly from Landsat have long been applied in biomass mapping for several reasons including (i) free availability, (ii) historic data archive, (iii) large scene size, (iv) compatible spatial resolution with standard sample plots (e.g., FIA), and (v) sensitivity of spectral reflectance to canopy cover. However, the sensitivity of optical remote sensing data generally saturates in closed canopy forests which leads to underestimation at high biomass areas and overestimation in low biomass areas. Active sensors such as LiDAR and RADAR are credited to be most accurate in biomass mapping at local scales. LiDAR has gained popularity because of high accuracy recognized in the characterization of horizontal and vertical structure of the canopy even in complex forest types. The increasing availability of LiDAR data and processing platforms to derive numerous metrics representing vegetation height, sub-canopy topography, and ground elevation have facilitated operational use in forestry research and development. The optical data provides reliable information on horizontal dimension while LiDAR is the most suited means to trace the vertical profile of forest canopy (Walker et al., 2007). However, the issues of smaller spatial coverage and high cost of LiDAR acquisition favors optical satellite data application for large-area assessment. RADAR technology is promising for measurement of forest structural properties as it can acquire data independent of weather and time. However, like optical sensors RADAR signal also suffers from saturation at lower biomass density, ranging 20-100 $Mg.ha^{-1}$ (Ranson et al., 1997; Ahamed et al., 2011; Næsset et al., 2011).

Spectral reflectance, vegetation indices (e.g., normalized difference vegetation index, NDVI), land cover, and canopy density metrics are the fundamental predictors applied in optical remote sensing based inventory. Climatic (e.g., mean annual temperature and precipitation), soils, and topographic variables (elevation, slope, aspect) have also been applied in some studies as ancillary data because of their recognized association with biomass production and distribution (Ohmann and Gregory, 2002; Baccini et al., 2004; Anaya et al., 2009; Straub and Koch, 2011). Spatial models of canopy height and age derived from remotely sensed data have been found to have very good correlation with structural attributes. The efficiency of any model depends on the explanatory power, number and type of predictors. The privacy policy of FIA regarding confidentiality of plot coordinates (O'Connell et al., 2013) restricts the linking of plot measured response variables with desired number of geospatial features corresponding the plots. Although plot locations are not disclosed, the spatial data service of FIA helps attaching a limited number of spatial predictors to the plot data via actual coordinates after internal security screening. These privacy restrictions severely limit the number of options available for developing accurate and robust mapping models, particularly because it impedes data mining or model selection techniques to determine the best set of remote sensing and geospatial predictor variables. The FIA database, however, provides “fuzzed-swapped” coordinates for the plots. The fuzzing generally creates an offset by 0.8 km (0.5 mile) from the actual plot location, while swapping makes exchange of inventory attributes among 20% of similarly stocked plots under private ownership in each county. Despite these issues, FIA data still provide a reliable and attractive source of information for developing inventory models for large-area mapping. The FIA plot data are particularly useful for modeling because they are probability based samples and involve less bias compared to purposive samples (Jenkins et al., 2003).

Air-borne discrete return scanning LiDAR is the widely used state-of-art technology for accurate and detailed characterization of forest structural attributes for operational management (Lu et al., 2012; Wulder et al., 2012), especially at landscape (multi-stand) scale. Since LiDAR signals can penetrate canopy gaps and directly measure the horizontal and vertical profile of the canopy, several structural attributes can be modeled from LiDAR derived metrics, such as canopy height distribution, sub-canopy strata, cover, and crown dimensions (Popescu, 2007; Hudak et al., 2008; Falkowski et al., 2009; Falkowski et al., 2010; Popescu et al., 2011). The increasing resolution and coverage of new generation of sensors and publicly available processing platforms capable of generating numerous canopy, sup-canopy, density, cover, texture, and terrain metrics have made LiDAR data popular for operational use (Hudak et al., 2008; Næsset and Gobakken, 2008; Hudak et al., 2012). In the conventional area-based approach to forest attributes estimation using LiDAR data, field measured attributes from fixed dimension plots are generally related to LiDAR derived metrics for the same area, i.e. the spatial resolution of LiDAR metrics is restricted to the size of ground plot. The fixed area plot measurements are time consuming and costly, so implementation of this method generally involves a tradeoff between sample size and plot size. Further, smaller plots have higher variability (coefficient of variation) compared to larger plots which impacts accuracy of estimates. In the rapidly expanding era of LiDAR application, with promising accuracies revealed, improved approaches are continually being explored in remote sensing community to obtain cost-efficient results of acceptable accuracy. For example, LiDAR samples have also been used recently as substitutes for field plots (Wulder et al., 2012a).

An accurate metric of site productivity is important for forest growth modeling. Site index (SI), defined as the height of dominant and co-dominant trees in competition free environment at a given base age (e.g., 50 years in the Lake States), is a proxy for forest productivity (Rehfeldt et al., 2006; Crookston et al., 2010; Weiskittel et al., 2011). Accurately estimating SI depends on accurate estimates of total height and age of sample trees that are free of past competition and damage. Thus, the method is most suited to fully stocked even-aged stands of known or measurable age. Although tree height for SI calculation can be measured with higher accuracy, tree age estimation is often difficult or impossible to obtain, especially for diffuse-porous hardwood tree species that grow slowly. Further, the total height estimates may also be erroneous when tree tops are broken. In regions like the Lake States, many stands are characteristically composed of shade tolerant species in uneven-aged conditions, and it is not surprising that substantial error exists in SI estimation. Since finding sample trees of dominant quality in competition free niche is difficult at many sites, development of spatially explicit map of SI may be useful for many applications.

Site quality, and thus SI as an index of quality, depends on the interaction of several biogeoclimatic variables including local management regimes (Stage et al., 2001; Rehfeldt et al., 2006; Crookston et al., 2010; Weiskittel et al., 2011; Sharma et al., 2012). Spatial variability in topography, soil, climate, and complex biotic interactions leads to variations in site conditions. The moisture gradient across a landscape, soil depth, soil nutrient, and temperature characteristic can influence site productivity since physiological systems of vascular plants are rooted to these factors. The spatial and temporal variation in forest site productivity can be modeled dependent on measures of climate, soil moisture, soil nutrients, land cover type, canopy density, canopy height, topographic variables, and other satellite imagery derived digital metrics (Klinka and Carter, 1990; Monserud et al., 2006; Monserud et al., 2008). Since site productivity depends on climate and climate is changing, integration of climatic spatial data is essential to make SI prediction models sensitive to climate. The changing paradigm in forestry to holistic management justifies the search for alternatives to traditional SI (Pokharel and Froese, 2009). A number of geospatial layers of biogeoclimatic features are freely available through public web-portals. These spatial predictors can be coupled with FIA data in order to formulate SI models, and of course predicted SI can be incorporated into growth models to analyze the potential for broader application. The likelihood that FIA plots are evenly distributed over all age and site classes make the database more appropriate for regional SI modeling. The predicted SI may be a useful explanatory variable in other geospatial inventory models, especially for uneven-aged mixed species stands where site trees are difficult to identify and measure for total height and age. Spatial mapping of SI allows for estimation of site quality even for the areas that are presently devoid of forests but need afforestation.

1.2. Spatial inventory modeling and considerations

The conventional approach to estimating forest population parameters is to aggregate sample plot statistics, provided the sample adequately represents the population characteristics (Jenkins et al., 2003; Golinkoff et al., 2011; Brosofske et al., 2014). Such methods may be robust, but they are costly and time consuming (Wulder et al., 2012a). Consequently, spatial modeling has evolved as a strategy to extrapolate sample estimates of inventory attributes across a large unsampled area of interest via remotely sensed and other geospatial auxiliary layers that augments the estimation process (McRoberts et al., 2002). A remote sensing based inventory essentially requires a reference frame out of a sample of ground plots of known coordinates such that co-located auxiliary geospatial predictor values are attached to the measured response variables. Then, the process involves formulation of an empirical relationship from the reference frame and then spatial prediction across the entire target area via contiguous pixel units where only predictor variables are known as digital signatures.

Spatial inventory of forest structural attributes using a host of input spatial datasets (multi-resolution, and multi-coverage) and modeling frameworks have impact on the accuracy of estimates. Numerous modeling approaches prevail in myriad of published studies that have integrated various combination of optical, LiDAR, RADAR and other geospatial data with field sample data in the framework of parametric or non-parametric regression (Walker et al., 2007; Koch, 2010; Wulder et al., 2012; Brosofske et al., 2014). The parametric regression methods are the most common for biomass mapping (Fuchs et al., 2009; Powell et al., 2010); however, the inherent assumptions (e.g. independence, linearity, normality, and homoscedasticity) are often violated in multivariate remote sensing based assessment (Evans et al., 2011; Burkhart and Tomé, 2012; Brosofske et al., 2014). The regression assumptions that biomass is linearly related to spectral response, and individual predictors are unrelated leads to biased prediction since multicollinearity among remotely sensed data is rife (Rehfeldt et al., 2006).

A widely used non-parametric method that integrates sample inventory with remotely sensed and other geospatial data for large scale mapping is the k-nearest neighbors (kNN) imputation (Moeur et al., 1995; Katila and Tomppo, 2001; Haapanen et al., 2002; LeMay and Temesgen, 2005; Falkowski et al., 2008; Eskelson et al., 2009). The kNN has been extensively applied for local to regional scale estimation of forest attributes in many countries (Tomppo and Halme, 2004; McRoberts, 2012). The method has the ability to simultaneously predict multiple responses at unsampled locations based on the relationship of response and feature variables at the reference sample locations (Hudak et al., 2008; McRoberts, 2009). In the simplest form of kNN, the prediction at any target point is calculated as the weighted average of the nearest neighbors from the reference (training) set; the weight decreases with increasing distance (e.g., inverse distance squared). The imputation algorithm begins with the calculation of a similarity (nearness) between the target and reference points where the target points have known values of only the auxiliary features but the reference points have both auxiliary and response features. The nearness between a target and reference units can be determined in the feature space of covariates by using several methods (McRoberts et al., 2007; Crookston and Finley, 2008; Hudak et al., 2008; Falkowski et al., 2010). The Random Forest (Breiman, 2001) based proximity metric (Crookston and Finley, 2008) is a noble measure commonly applied in imputation mapping of forest resources using multivariate remotely sensed data (Rehfeldt et al., 2006; Falkowski et al., 2009; Ohmann et al., 2011). The RF algorithm can simultaneously handle categorical and continuous variables for multiple responses and predictors.

The RF algorithm works on the basis of aggregated result of an ensemble (forest or machine) of many classification and regression trees where each is generated independently out of a bootstrap sample (usually two-third) of reference data (Breiman, 2001a; Liaw and Wiener, 2002; Cutler et al., 2007). The individual trees in the forest are made independent (or uncorrelated) with the introduction of an additional random component in tree formation where each node split depends on the best predictor among a random subset of all predictors. The tree development also requires an optimization function to select a node, a predictor variable, and a cut-off value that result in the more homogeneous child nodes, as measured by the Gini index (Falkowski et al., 2009). The tree growth stops at a point when further splitting does not reduce the Gini index. Thus, each terminal node of a tree contains a cluster of most similar observations. The RF proximity measure is the proportion of trees where target observation is in the same terminal node as a reference observation (Breiman, 2001; Liaw and Wiener, 2002; Crookston and Finley, 2008). The RF algorithm is strictly non-parametric, flexible and robust with respect to non-linear and noisy relations among input variables (Cutler et al., 2007). Further, the algorithm does not require cross validation data since out-of-bag observations (about one-third) for each tree provide error of individual trees which are then summarized (averaging for continuous, or majority vote for categorical variables) across all trees to estimate the overall accuracy. The algorithm also gives relative importance ranking of predictors by randomly permuting the values of one predictor at a time and reporting the proportional increase in mean square error of the model (Liaw and Wiener, 2002; Falkowski et al., 2009).

The accuracy of imputation depends on the choice of predictors, explanatory power of auxiliary variables, size and distribution of reference sample, distance or nearness measure, number of neighbors (i.e. value of *k*), and weight function used for prediction (Ohmann et al., 2011). A large size of reference sample can be expected to improve the accuracy as closer matches of reference and target covariates would be available for imputation (LeMay and Temesgen, 2005). When a single nearest neighbor is applied (i.e., *k*=1), then the imputed value of a response at a target point will simply be the observation from one of the reference points. In the case of single neighbor imputation, the natural variation of inventory variables is retained in the prediction but accuracy is reduced at the plot level (Moeur et al., 1995; Haapanen et al., 2002; Holmstrom and Fransson, 2003). When more than one neighbor is selected, the accuracy of prediction may improve but at the cost of higher bias (McRoberts et al., 2002). The larger bias with higher number of neighbors can be reduced by weighted averaging (Katila and Tomppo, 2001).

The variation in forest composition, structure, and phenology (due to topographic, climatic, and site variables) over space and time combined with data scarcity restrict model calibration and expansion across a broader area and time scale (Foody et al., 2003; Lu et al., 2012). The optimization strategy and criteria of model selection depends on input data availability, accuracy requirements, simplicity, assumptions and limitations, and uncertainty (Zhang and Ni-meister, 2014). A large reference dataset, adequately capturing the compositional and structural diversity of the target area, is important for model training to achieve reliable predictions of inventory attributes (Labrecque et al., 2006; Song, 2012). It can be assumed that models including multiple variables may give better prediction but such models likely have limited application in lack of large-scale datasets. The RF based kNN (RF-kNN) has advantage of producing distribution-free models that can accommodate numerous variables (multivariate and multi-response). Use of variable selection algorithms have been found to be a good strategy to reduce multicollinearity among predictors (Falkowski et al., 2009; Hapfelmeier and Ulm, 2013). Validation of prediction estimates from spatial inventory models is critical for operational application. However, performance of a model may vary with location, spatial scale of prediction, and selected fit statistics (e.g., R^2, RMSE, and bias) with different validation dataset (Powell et al., 2010).

An efficient inventory model requires that predictions are consistent for both small and large spatial extents and that variance is known to the end users. Different modeling approaches and optimization criteria have been used in the past for large scale predictive mapping of biomass. For example, Blackard et al. (2008) used coarse resolution (250 m) MODIS data, a national land cover dataset, and geo-climatic variables along with the FIA plot data to prepare spatially explicit biomass map for the conterminous USA circa 2003. Similarly, the Woods Hole Research Centre has produced a finer resolution (30 m) biomass map as a part of the National Biomass and Carbon Dataset 2000 (NBCD 2000) for the conterminous USA by combining FIA data with high-resolution RADAR data acquired from the 2000 Shuttle RADAR Topography Mission (SRTM), and optical remote sensing data acquired from the Landsat ETM+ sensor (Kellndorfer et al., 2004; Kellndorfer et al., 2012). Further, researchers and managers can develop new spatial models based on FIA data in two ways: (i) using the fuzzed-swapped coordinates of plots, available publicly via online database, to attach any number of geospatial predictors to the plot data, and (ii) collaborating regional FIA units that can attach limited number of geospatial predictors to the plot data via actual coordinates. The map products obtained from different modeling approaches need accuracy analysis at multiple spatial scales prior to any application.

The cost-efficiency of LiDAR based inventory modeling can potentially be improved when integrated with sample data collected through a quick, unbiased, and easy technique called point or variable radius plot (VRP) sampling (Avery and Newton, 1965; Bropleh, 1967; Avery and Burkhart, 1994). The VRP sampling is especially useful for timber inventory as efforts are more focused on big trees that hold the most volume and value. However, integrating VRP data with LiDAR data requires special strategies as the exact (optimal) size of VRPs remains unknown even for a known basal area factor (BAF). A major challenge is to find the optimum plot size so that the inventory variable of a VRP best matches with the resolution at which plot-level LiDAR metrics are derived (Golinkoff et al., 2011; Hollaus et al. 2009).

An option to estimate SI for any target locations is to apply RF-kNN imputation procedures that can spatially extend the measured SI values from the FIA plots, based on referenced auxiliary biogeoclimatic spatial layers. FIA database provides a good source of species-specific SI measurements per plot and can be integrated with biogeoclimatic spatial layers for large scale mapping. The FIA computes species-specific SI for every tree in the sample plots, based on measurements of one or more dominant and co-dominant site-trees per plot (Woudenberg et al., 2010).

1.3. Dissertation focus

A common approach in all the chapters of this dissertation is to apply RF-kNN imputation algorithm for the spatial prediction of forest inventory attributes across multiple spatial scales by coupling field inventory data with remote sensing and geospatial data at various resolutions. The principal questions of the dissertation chapters are as below:

1. How does the choice of model type and input data affect biomass predictions from alternative model approaches at small to large spatial scales?
2. How accurately can structural attributes be mapped using LiDAR data when coupled with field data collected with two different sampling methods, one more focused on cost-efficiency and the other on accuracy?
3. How efficient is the regional scale species-specific digital map of site index developed from the combination of FIA and geospatial data?

The first chapter (Chapter 2) presents an approach to utilize FIA data for large area biomass mapping in two contrasting ways: (i) using a limited number of spatial predictors under the policy restrictions on actual plot coordinates to develop a high resolution map (30 m pixel), and (ii) leveraging a large number of spatial predictors related via fuzzed-swapped coordinates to develop a coarse resolution map (250 m pixel). The predictor layers included in the study were the product of optical remote sensing (Landsat derived vegetation index, and land cover and MODIS derived slope raster), and fusion of optical and RADAR remote sensing (basal area weighted height), along with other geo-climatic datasets. The two map products of this study were compared with two other existing maps for assessment of biomass estimation accuracy at plot, stand and county scales. The small-and large-areas biomass estimates of the individual models, developed using different data sources and optimization criteria, were compared to recommend a suitable model depending on the area of operation. A key focus was on generating a high resolution map of operational use based on publicly available datasets.

The second chapter (Chapter 3) intends to leverage the strength of LiDAR derived metrics with inexpensively collected field data (following an unbiased variable-radius sampling) from indeterminate ground coverage to perform cost-effective spatial inventory of standing volume at multi-stand scale. The accuracy of inventory estimates from the indeterminate (variable-radius) sampling based imputation model was evaluated on the basis of estimates obtained from fixed area sampling based imputation model. The comparison of the two model estimates was done only at the plot level (not at the stand level in absence of sufficient number of stand inventory data). The study was carried out in six conifer stands at the Ford Forest of Michigan Technological University.

The third chapter (Chapter 4) integrates FIA measured species-specific SI with a number of biogeoclimatic variables to produce spatially explicit map of SI for five major species of the Lake States (MI, WI, and MN) at a spatial resolution of 250 m. Accuracy of the SI imputation models was evaluated by comparing the predicted SI against the measured values at a set of FIA plots other than the ones used for model training. In addition, the performance of the imputed SI was analyzed in the Forest Vegetation Simulator's (Dixon, 2002) large tree diameter growth models which are characteristically dependent on measured SI. The diameter growth predictions based on the models separately using measured SI and imputed SI was validated against the field observations at the tribal lands in Minnesota and Wisconsin managed under the Bureau of Indian Affairs.

2. Evaluation of multivariate imputation methods for the spatial inventory of above-ground biomass in favor of operation planning in the Great Lakes region[1]

2.1. Introduction

Forest biomass is the largest terrestrial carbon sink and thus a crucial ecological variable for understanding and mitigating climate change (FAO, 2009; Hudak et al., 2012). Since live forest biomass sequesters atmospheric carbon and biomass removal or mortality causes greenhouse gas emissions, explicit assessment and mapping of biomass (dry weight of which can contain 45-50% of carbon) can improve our understanding of the carbon cycle. Consequently, international agreements and conventions are adding economic value to biomass. This has ultimately created a demand for baseline biomass maps as a means to quantify changes in carbon stocks in support of programs such as REDD+ (Reducing Emissions from Deforestation and forest Degradation) (Walker et al., 2007; Tomppo et al., 2008; UN-REDD, 2010). In addition, commercial conversion of woody biomass into a sustainable energy sources (as biofuel or electricity) has drawn more attention toward the assessment of the distribution and availability (quantity, and accessibility) of biomass as a resource for biofuel feedstock (White, 2010; Gleason and Im, 2011; Straub and Koch, 2011). Indeed, a detailed understanding of the spatial distribution of forest biomass is required for management operations and ecological sustainability in addition to estimating the carbon stock and bioenergy potential of a given area (Tuominen et al., 2010). When overlaid with land ownership, forest type, site index, and transportation layers, spatially explicit forest biomass information can assist in the identification of potential harvest areas and availability across multiple spatial extents. Small area inventory estimates at local scales guide operational management activities while large area regional assessments are necessary to inform national strategic plans and policies.

Forest biomass assessment can be done solely via *in situ* sampling or by integration of *in situ* data with remote sensing information in a modeling framework (FAO, 2009). *In situ* measurements from national forest inventories (NFI) are a reliable source to derive regional or national level biomass estimates (Jenkins et al., 2003). However, NFI sampling designs are insufficient for generating inventory information at resolutions appropriate to operational management and biofuels planning. This is because NFIs are

[1] This chapter is ready to submit in a remote sensing journal with Ram K. Deo as the first author and Robert E. Froese and Michael J. Falkowski as the second and third authors respectively.

specifically designed for large area assessment (e.g. regional or national level) by means of a systematic network of sparsely distributed permanent sample plots (Fazakas et al., 1999; Franco-Lopez et al., 2001; McRoberts, 2012). The large degree of spatial separation between NFI plots ultimately limits estimation of inventory attributes for small-areas due to insufficient sample representation. For example, the smallest area for which attributes were estimated at an acceptable accuracy was approximately 150,000 ha in Finland (Tomppo and Katila, 1991) and 500,000 ha in Sweden (Fazakas et al., 1999). Hence, only limited or inadequately precise forest statistics can be expected for small areas based on NFI measurements alone (Tomppo et al., 2008). This is a challenge for resource managers, who are often interested in generating forest inventory information on the amount and distribution of, say, biofuel feedstock at sub-regional levels such as forest stands. Two options are available for generating or improving inventory information for small-areas: (i) conduct additional field surveys within the small target area, or (ii) use sample data from outside the area of interest and leverage the combined strengths of the field and remotely sensed data through advanced modeling algorithms. Geographically localized small-area estimation methods generally adopt the latter strategy by augmenting NFI data with remotely sensed and geospatial predictors. The spatially explicit products from such integration could potentially benefit both small-area operations and regional forest planning, if sufficiently accurate.

The success and synergies of remote sensing and geospatial data have been noteworthy in the past two decades for generating biomass maps at multiple spatial scales (Walker et al., 2007; Anaya et al., 2009). The assured quality of *in situ* data collected to a national standard by NFI, particularly in the United States, offers a good promise to improve small-area assessments when integrated with multisource high spatial resolution datasets. Published research has established relationships between biomass and a combination of remote sensing, topographic, and climatic data sources (Baccini et al., 2004; Saatchi et al., 2007; Wulder et al., 2008a). Remotely sensed reflectance and derived metrics such as normalized difference vegetation index (NDVI), percentage canopy density, canopy height, cover types, and texture are commonly used predictors in biomass modeling and mapping (Tomppo and Halme, 2004; Hall et al., 2006; Wulder et al., 2008a; Anaya et al., 2009). The selection of suitable metrics and algorithms form the basis of efficient modeling and mapping strategies (Lu et al., 2012). Due to the complexity of forest structure, composition, phenology, and sites, several ancillary variables representing soil productivity, topography, and climate are often included to account for the non-linear relationships between biomass and spatial predictors at the landscape scales (Baccini et al., 2004; Blackard et al., 2008; Powell et al., 2010; Ohmann et al., 2011; Brosofske et al., 2014). Further, time series spectral trend at pixel level is applied to leverage the temporal information (e.g. change in forest surface, growth, age) of satellite data (Powell et al., 2010; Le Maire et al., 2011). While multispectral data provide two-dimensional information on canopy coverage, integration of height information such as basal area weighted canopy height (BAWHT) from active sensors has potential to improve biomass prediction accuracy (Kellndorfer et al., 2004; Pond et al., 2014). A moderate resolution spatial dataset representing BAWHT (circa 2000) is freely available for the conterminous USA (Walker et al., 2007; Kellndorfer et al., 2012). Some studies combine spectrally calibrated moderate and coarse resolution optical data where moderate resolution data are used to establish a model with the field data and coarse resolution data are used to spatially extend the model (Zheng et al., 2007; Wulder et al., 2008a). However, solitary use of passive remotely sensed data is generally insufficient, particularly for small-area estimation. For example, Powell et al. (2010) used Landsat imagery and NFI data to map forest biomass in Minnesota, USA and attained a plot-level root mean square error (RMSE) between 61-69%. When estimating standing volume, from a combination of Landsat, geospatial, and NFI data, Tomppo et al. (2008) reported relative RMSE of 50-80% at the pixel (plot) level, 13-14% in small-areas of size 1 km^2, and 5% for the size of 100 km^2 in Finland and Sweden. Similarly, Franco-Lopez et al. (2001) observed a relative RMSE of 83.76% for plot level volume estimation in Minnesota from a combination of NFI data and Landsat imagery, while Yim et al. (2011) observed a relative RMSE of 55-75% for the plot level volume prediction when integrating Landsat and NFI data in central South Korea. Fazakas et al. (1999) modeled NFI and Landsat data and obtained 66-78% RMSE for plot-level biomass prediction in Sweden.

The accuracy of remote sensing based approaches to inventory modeling depends on several factors. These include the quality of the remote sensing and geospatial data (e.g., resolution, atmospheric attenuation), the sensitivity of sensor to the variation in forest structure, as well as characteristics of the NFI data such as quality, sampling intensity, and data availability (Hall et al., 2011). Spatial mismatch of field plots and corresponding pixels in remotely sensed imagery, incompatible size of the plots and pixels, poor sensitivity of image bands, radiometric variations within and among adjacent scenes, and mixed pixel effects in coarse resolution data are additional sources of uncertainty (Tuominen and Pekkarinen, 2005; Muukkonen and Heiskanen, 2007; Tomppo et al., 2008). Multispectral optical sensors such as Landsat or MODIS are inefficient to capture the spatial variability of forest structure as the sensors become insensitive in high biomass areas and possess limited power for discriminating species and cover types (Lefsky et al., 2002; Lu, 2006; Song, 2012). This is revealed in Huete et al.(2002) who pointed out the insensitivity of NDVI in high biomass regions and Steininger (2000) who observed saturation of canopy-reflectance and biomass relationship at around 150 Mg ha^{-1}. The insensitive spectral data ultimately leads to spatial models that generate estimates closer to the mean (called regression towards the mean effect) at every pixel, i.e. over-prediction in areas with low biomass and under-prediction in areas with high biomass (Baccini et al., 2004).

Active sensors such as LiDAR and RADAR are substantially more accurate in biomass mapping at local scales. LiDAR based state-of-art techniques for biomass assessment have shown remarkable success since the sensors accurately measure three dimensional canopy profile as well as terrain elevations (Koch, 2010; Lu et al., 2012; Wulder et al., 2012). But application of LiDAR is generally limited due to higher cost acquisition, and also narrow spatial coverage of the sensors. Synthetic aperture RADAR (SAR) operated from satellite platforms is less expensive and promising technology but suffers from saturation of backscatter intensity comparatively at lower biomass levels of around 20-100 Mg ha^{-1} (Ranson et al., 1997; Ahamed et al., 2011; Næsset et al., 2011). Interferometric SAR (InSAR), available from both spaceborne and airborne platforms, is more promising when used in concert with a digital terrain model (DTM) since the difference of the DTM and InSAR height is strongly related to the height of forest canopies and above ground biomass (Næsset et al., 2011). Although not as accurate as active remote sensing, optical remote sensing is still essential to monitor biomass and biomass change over a large spatial and temporal extents (Le Maire et al., 2011). Given the pressing need for regional-level spatially explicit biomass assessment, Landsat imagery is considered appropriate for biomass mapping for several reasons including (i) free availability, (ii) historic data archive, (iii) large scene size, and (iv) a spatial resolution comparable to the size of typical NFI plots (Labrecque et al., 2006; Main-Knorn et al., 2011). Publicly available multisource, multitemporal, and multisensor geospatial datasets including climate, and topographic layers are often combined to improve biomass estimation accuracy (Ahamed et al., 2011).

When developing remote sensing based biomass-mapping models, it is important to assure that sample data represent the entire range of variability in the forest conditions of the area of interest. The U.S. Forest Service’s national Forest Inventory and Analysis (FIA) program collects data annually over all ownerships via a network of design-based sample plots with an intensity of at least one plot per 2400 ha. The Great Lakes States (MI, WI and MN) have more than 44,000 permanent plots that represent the entire range of biomass variability in the region and hence is apt to develop a statistically robust prediction models (Franco-Lopez et al., 2001; Zheng et al., 2007). Privacy restrictions on data access are, however, a major barrier in the development of accurate models for biomass prediction. The legal security restrictions on FIA data do not allow outside entities access to true plot coordinates. Although outside users can request to have FIA program analysts intersect plot data with remotely sensed and geospatial data (i.e., to create a data frame for geospatial prediction), security restrictions limit the number of spatial data layers that can be used. This is because security restrictions are in place to ensure that data users outside of the FIA program cannot trace-back actual plot locations from the derived data frame. The FIA database, however, provides “fuzzed-swapped” coordinates for the plots; fuzzing generally creates an offset by 0.8 km from the actual plot location, and swapping makes exchange of inventory attributes among upto 20% of privately owned similarly stocked forested plots for each county (O'Connell et al., 2013). These privacy restrictions severely limit the number of options available for developing accurate and statistically robust mapping models, particularly because it impedes using data mining or model selection techniques to determine the best set of remote sensing and geospatial predictor variables. Despite these issues, FIA data still provide a reliable and attractive source of information for developing biomass models for large area mapping. Modeling of response variables from probability based samples (e.g. FIA plots) involves less bias compared to purposive samples (Blackard et al., 2008). This study explores the operational suitability of biomass models formulated from the FIA data obtained under the privacy constraints.

Application of several parametric and non-parametric regression approaches for biomass modeling prevails in myriad of published studies. However, parametric methods generally rely on statistical assumptions (such as independency, normality, linearity, and homoscedasticity) that do not hold true with remote sensing data (Evans et al., 2011; Robinson and Hamann, 2011; Burkhart and Tomé, 2012; Lu et al., 2012). Non-parametric models are gaining popularity for many reasons including the fact that they are robust and free from many statistical assumptions. One widely used non-parametric method is *k* nearest neighbor (kNN) imputation which has been applied in large area forest inventory since the early 1990s (Tomppo, 1991; Moeur et al., 1995; Ek et al., 1997; Van Deusen, 1997; Tomppo and Halme, 2004; McRoberts, 2012). In the kNN, predictions at a target location is the weighted average of response measurements at the *k* nearest neighbors in the domain of reference samples, where nearness is determined based on the similarity of spatial predictors known at every unit throughout the area of interest (McRoberts et al., 2002; McRoberts, 2012). An advantage of the kNN method is that multiple response variables can be predicted simultaneously at unsampled locations based on multiple spectral or auxiliary features. The application of more predictor variables in a model may improve its precision but the design also needs to be parsimonious and pragmatic since operational use requires inexpensive, accurate, update, and accessible auxiliary predictors valid for general conditions for spatially explicit mapping. An issue often noticed when using a large number of predictors from remotely sensed and other geospatial datasets is multicollinearity among predictors that may lead to unstable predictions. Indeed, research has demonstrated that including variable selection or model selection procedures can improve results (Falkowski et al., 2009; Latifi et al., 2010; Hapfelmeier and Ulm, 2013).

An efficient spatial inventory model requires that predictions are consistent for both small and large spatial extents and variance is known to the end users (for confidence in the predictions). Different modeling approaches and optimization criteria have been used in the past for large scale predictive mapping of biomass. For example, Blackard et al. (2008) used coarse resolution (250 m) MODIS data, a national land cover dataset, and geo-climatic variables along with the FIA plot data to prepare spatially explicit aboveground biomass map for the conterminous USA circa 2003. Similarly, the Woods Hole Research Centre has produced a finer resolution (30 m) spatially explicit above-ground dry biomass map as a part of the National Biomass and Carbon Dataset 2000 (NBCD 2000) for the conterminous USA by combining FIA 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 (Kellndorfer et al., 2012). However, the quality of such map products has not been compared with each other.

2.2. Objectives

Given that the FIA privacy protocols cause reduced power of explanatory variables while linking remotely sensed and geospatial values to the sample plot data, the general objective was to formulate multivariate spatial inventory models of biomass under the scenarios of true and fuzzed-swapped plot locations (both compromising with reduced power of predictors) and evaluate the accuracy at multiple spatial scales. The concept was to prepare biomass maps for a portion of the Lake States of the USA at 30 m and 250 m spatial resolutions, with the actual and fuzzed-swapped coordinate datasets respectively, and compare the correspondence of both small- and large-area estimates against the extant maps produced at 30 m resolution in the NBCD and at 250 m resolution by Blackard et al. (2008). In addition, performance of new models at small and large scales was intended to be verified with independent stand inventory datasets, and the county-level estimates from the FIA database. Realizing the need for cost-effective accurate biomass estimation for small-areas (e.g. stands) where only field observations external to the area are available, and large areas (e.g. county) at which FIA design may include a reasonable number of sample units, the following objectives were set:

I. To develop geospatial models and maps of biomass in favor of operational planning by employing restricted information from remote sensing and other geospatial datasets under the FIA privacy policy restrictions
II. To assess the accuracy of small- and large-area biomass estimates based on the new models and the extant models of NBCD and USFS

The ultimate focus was to ease the burden of model formulation and support prompt spatial inventory of biomass.

2.3. Methods

2.3.1. Study Area

A portion of the Lake States, specifically in the northern region of Michigan, USA, was selected as the study area (Figure 2.1). Forest biomass is the predominant target for biofuel production in the region, composed largely of mixed upland hardwood stands. The region supports diverse type of forests on glacial outwash plains within a matrix of lowland and upland mixed, conifer, and deciduous forests (Frelich, 2002). The major forest cover types in the region include wet deciduous (elm-ash-cottonwood), oak-hickory, mesic deciduous (maple-beech-northern hardwoods), pine, aspen, and wet coniferous-boreal (spruce-fir) (Dickmann and Leefers, 2003). The annual inventory system of FIA has been implemented in the region since 2000.

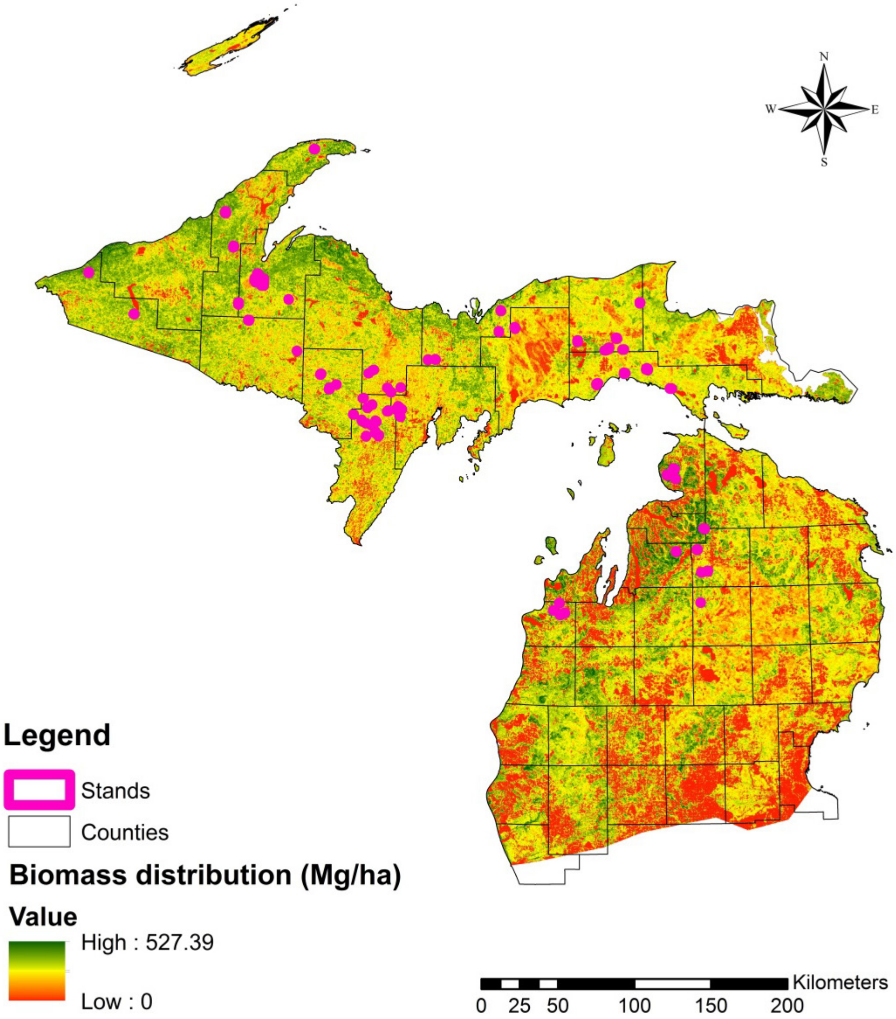


Figure 2.1. The study area and location of the validation stands in the upper Michigan, U.S.A.

2.3.2. FIA inventory system in the study area

The FIA plots in the study area are distributed across all public and private lands with sampling intensity of up to three plots per 2,400 ha of hexagonally gridded land area. There are more than 10,000 plots in the study area, approximately 20% of which are measured annually via panels with 5 years rotation. The plots are located such that each has at least 10% canopy cover within 0.4 ha neighborhood (Walker et al., 2007). Each plot consists of a cluster of 4 subplots (each with 7.32 m radius) in which all trees above 2.54 cm dbh are measured for numerous attributes. The plot-level data includes forest type, condition (forest/non-forest), site index, tree species, tree condition (live or dead), and tree size including dbh and height. The tree size measurements are used in species-specific allometric equations to derive individual tree volume or biomass which are then summarized to plot-level biomass on per unit area basis (O'Connell et al., 2013). The biomass (used interchangeably with above-ground biomass in this dissertation) in FIA parlance is sum of dry biomass in bole, stump, branches and twigs of all live trees above 2.54 cm dbh. The actual coordinates of the plots are kept confidential to maintain private owners' privacy, and also plot integrity.

2.3.3. Remote sensing and geospatial data

The initial set of spatial predictors for biomass mapping were Landsat 5 Thematic Mapper (TM) derived NDVI, land cover data from the state of Michigan's project called IFMAP (IFMAP, 2001), a digital elevation model (DEM) from the seamless data warehouse of the USGS, and basal area weighted canopy height (BAWHT) from the NBCD (Walker et al., 2007). All these raster layers are publicly available at a spatial resolution of 30 m. In addition, a coarse index of disturbance as an ancillary layer, named MODIS-slope, was derived using time series NDVI imagery from the MODIS sensor to account for the inter-annual vegetation phenology and structural variations.

The TM NDVI layer for the study area was prepared from 22 cloud-free Landsat scenes of the growing seasons (Jun-Aug) from the years 2006 to 2010 (Table 2.1). Only growing season imagery was considered to reduce the impact of seasonal phenological and solar zenith angle variations on spectral reflectance characteristics. A model builder in Erdas Imagine software was used for radiometric calibration of the images that involve conversion of the raw digital numbers to the absolute units of at-sensor spectral radiance (W m^{-2}) and finally to the top-of-atmosphere reflectance (TOA) (%) using algorithms and coefficients as in Chander et al., (2009). The TM NDVI raster which was produced from the TOA reflectance image as normalized ratio of the difference of near-infrared band (highly reflective to green leaves) and red band (highly absorptive to chlorophyll), was considered since the index is insensitive to many forms of multiplicative noise, and sensitive to the amount of green biomass (Huete et al., 2002; Jensen, 2005).

The original state-wide IFMAP land cover raster, derived through the classification of three-season Landsat TM imageries collected between 1997-2001, contained 32 thematic classes with 12 forest categories. For this study, the IFMAP cover types were reclassified into eight broader classes consisting of seven forest and one non-forest categories in an approach to assign a cover type to each of the FIA plot based on species composition (Table 2.2). The BAWHT raster was developed at the source using 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 MODIS-slope raster was produced from the MODIS/Terra derived 16-day composite NDVI images of anniversary dates from 2005 to 2010 (Table 2.3) in the peak growing season that offered similar solar zenith angle and phenological traits to the images. The images available at 250 m resolution were retrieved from the Land Processes Distributed Active Archive Center of the U.S. Geological Survey (LP DAAC, 2013). A simple linear regression, $y = a + b.x$, was fitted to the time series pixel values for the six years and a slope raster was calculated using the formula: $b = \left(N\sum xy - \sum x \sum y\right) / \left(N\sum x^2 - \left(\sum x\right)^2\right)$, where N represents number of years (i.e. six) and y is NDVI value of a pixel for the year x. The pixel values of the slope raster are assumed to characterize the growth, mortality and removal of growing stock. For consistency with other datasets related to the actual coordinate plots, the slope raster was resampled from 250 m to 30 m resolution using nearest neighbor approach in ArcMap 10 (ESRI, Redlands, CA, USA, 2011).

Some additional geospatial predictors were obtained and processed to model biomass directly from the fuzzed-swapped coordinates of FIA plots in the study area. These layers included the landcover dataset from the national gap analysis program (GAP, 2013) and geo-climatic variables. The climatic variables included frost-free degree-days above 5^0C (DD5), growing season precipitation (GSP), mean annual precipitation (MAP), mean annual temperature (MAT), and mean temperature in the warmest month (MTWM) that were obtained from a climate data server of the USFS Moscow Forest Sciences Laboratory (RMRS, 2013). The soil taxonomy dependent spatial layers, namely soil drainage index (DI) and productivity index (PI), were also used because these layers indicate long-term soil wetness, soil volume available for plant rooting, and potential tree stress areas (Schaetzl et al., 2012; Schaetzl et al., 2009). The DI and PI layers were downloaded from the forest health protection mapping and reporting portal (USDA Forest Service, 2013a). The GAP land cover dataset, originally produced from multi-season Landsat ETM+ imageries from1999-2001, is available at six different national vegetation hierarchies based on physiognomy (FGDC, 2008) but for ease of interpretation and analysis we considered only the macro-group with 59 classes and further aggregated into 20 broader classes by merging similar cover types. The climatic rasters were produced at the source by fitting Hutchinson's spline-surfaces to 30-year (1961-1990) normalized average monthly data from local meteorological stations throughout the North America (Rehfeldt et al., 2006; Crookston et al., 2010). These biogeoclimatic layers were resampled to a common spatial resolution of 250 m with exactly overlapping orientation of pixels in all the rasters which were used only with the fuzzed FIA plot data.

Table 2.1. Landsat imageries used in the study for the derivation of NDVI raster

WRS-2 path/ row	Lat/ long	Acquisition date	Scan time	UTM zone	Sun elevation	Earth-Sun distance*
20/ 29	44.6/ -82.7	2007-06-11	16:09:52	17	62.86	1.01536
20/ 30	43.2/ -83.2	2008-05-28	16:04:19	17	61.93	1.01355
20/ 31	41.8/ -83.7	2008-07-15	16:03:22	17	61.07	1.01646
21/ 28	46.0/ -83.8	2008-07-06	16:08:37	17	60.11	1.01670
21/ 29	44.6/ -84.3	2008-07-06	16:09:01	16	60.86	1.01670
21/ 30	43.2/ -84.8	2006-06-15	16:15:11	16	63.57	1.01577
21/ 31	41.8/ -85.3	2007-07-20	16:16:21	16	61.46	1.01616
22/ 28	46.0/ -85.3	2007-06-25	16:21:35	16	61.91	1.01652
22/ 29	44.6/ -85.8	2007-06-09	16:22:16	16	62.78	1.01513
22/ 30	43.2/ -86.3	2007-06-09	16:22:40	16	63.59	1.01513
22/ 31	41.8/ -86.8	2008-07-13	16:15:47	16	61.35	1.01655
23/ 28	46.0/ -86.9	2006-07-15	16:27:24	16	59.79	1.01646
23/ 29	44.6/ -87.4	2007-08-03	16:27:47	16	57.21	1.01471
23/ 30	43.2/ -87.9	2007-08-03	16:28:11	16	58.08	1.01471
24/ 27	47.4/ -87.9	2009-08-31	16:29:14	16	47.23	1.00946
24/ 28	46.0/ -88.4	2007-06-23	16:33:59	16	62.00	1.01642
24/ 29	44.6/ -88.9	2010-07-17	16:30:58	16	59.95	1.01635
24/ 30	43.2/ -89.4	2010-07-01	16:31:25	16	62.61	1.01667
25/ 27	47.4/ -89.4	2009-06-03	16:33:57	16	60.05	1.01433
25/ 28	46.0/ -89.9	2009-06-03	16:34:21	16	60.92	1.01433
25/ 29	44.6/ -90.5	2007-08-17	16:39:59	15	53.81	1.01244
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)

Table 2.2. Reclassification of the 12 IFMAP forest categories into 7 broader classes

IFMPAP Class	New Class	Description for categorization of FIA plots to the new class
14	1	Northern hardwood association (maples, American beech, American basswood, white ash, black cherry, and yellow birch exceeds 60% of total wood volume in plot)
15	2	Oak association (oak spp. exceeds 60% of total wood volume in
16	3	Aspen association (aspen exceeds 40% of total wood volume in
17, 18,	4	Deciduous dominant (deciduous trees exceeds 60% of total wood
19	5	Pines (pines exceeds 60% of total wood volume in plot)
20, 21, 25	6	Conifer dominant (conifers other than pines exceeds 60% of total wood volume in plot)
22, 26	7	Mixed forest (does not fall into any of the above categories; proportion of conifers and deciduous ranges from 40 to 60%)
all others	8	Non-forest

Table 2.3. MODIS/Terra NDVI imageries used to derive the MODIS-slope raster

H/V	Lat/Long	Scene ID	Date
11/4	45.0/-91.9	MOD13Q1.A2005209.h11v04.005	2005-08-13
11/4	45.0/-91.9	MOD13Q1.A2006209.h11v04.005	2006-07-28
11/4	45.0/-91.9	MOD13Q1.A2007209.h11v04.005	2007-07-28
11/4	45.0/-91.9	MOD13Q1.A2008209.h11v04.005	2008-07-27
11/4	45.0/-91.9	MOD13Q1.A2009209.h11v04.005	2009-07-28
11/4	45.0/-91.9	MOD13Q1.A2010209.h11v04.005	2010-07-28
12/4	45.0/-77.8	MOD13Q1.A2005209.h12v04.005	2005-07-28
12/4	45.0/-77.8	MOD13Q1.A2006209.h12v04.005	2006-07-28
12/4	45.0/-77.8	MOD13Q1.A2007209.h12v04.005	2007-07-28
12/4	45.0/-77.8	MOD13Q1.A2008209.h12v04.005	2008-07-27
12/4	45.0/-77.8	MOD13Q1.A2009209.h12v04.005	2009-07-28
12/4	45.0/-77.8	MOD13Q1.A2010209.h12v04.005	2010-07-28

2.3.4. Reference data frames

A reference set comprising of inventory data from 4,830 plots with actual coordinates intersected and attached to the key geospatial predictors were procured from FIA after an agreement abiding the privacy requirements. The TM NDVI and MODIS-slope rasters, both classified to 20 classes, were sent to the FIA unit at the Northern Research Station (Newtown Square, PA) to attach the auxiliary digital values to the inventory plot data. This strategy was pursued to maximize the number of reference plots under the FIA security screening that has set a minimum threshold on number of unique combinations of the raster values at the plot locations as well as non-sampled areas in each county. Pond et al. (2014) reported that at least 3 sample plots and 101.2 ha non-sampled areas need to have the same combination of raster values in each county for the data to be released. The plot-level inventory attributes obtained in the dataset were net timber volume, growth, mortality, and removal quantities on per unit area basis and also by species groups. In addition, FIA also attached ground elevation and BAWHT to each plots directly from its database. It was assumed that the FIA measured ground elevation and BAWHT are close to the respective values in the geo-referenced DEM and the BAWHT rasters used to spatially extrapolate the inventory data. Cover types of the plots similar to the 8 categories of the reclassified IFMAP dataset were deduced based on species dominance determined from plot-level net timber volume by species group. This strategy did not require us to request FIA for intersection and attachment of IFMAP values to the plots. Again, it was assumed that the IFMAP adequately represented cover types defined by FIA. The plot biomass was estimated from the net timber volume by applying average expansion factors of species group derived from FIA database.

An additional reference frame based on the data from the fuzzed-swapped coordinates of 7,322 FIA plots was prepared. The plot level biomass inventory and latitude and longitude data were obtained from the FIA database available online with the DataMart tool (FIA, 2013). The plot biomass was derived by summing up the above ground biomass of individual trees (≥ 12.5 cm dbh) per plot given in the tree table of the database. Only sampled plots (PLOT_STATUS_CD= 1&2) that were measured physically (SAMP_METHOD_CD=1) with standard quality assurance (QA_STATUS=1) were considered for preparation of the data frame for model training. The geospatial predictors attached to the plots via the fuzzed-swapped coordinates included BAWHT, TM NDVI, MODIS NDVI, GAP landcover, DD5, GSP, MAP, MAT, MTWM, DI and PI.

The large number of sample plots provided a good representation of actual forest conditions and diversity in the study area. Among the 4830 plots with actual coordinates intersected to the spatial predictors, almost 400 plots belonged to non-forest category with apparently no canopy height (i.e. zero BAWHT). The inventory data belonged to the seventh cycle of FIA in Michigan, measured in 2005-2009.

2.3.5. Validation data

Three independent sampling datasets termed FFC (Ford Forestry Center), Hardwood, and Aspen were used for the validation of spatial inventory models at the stand level; only FFC inventory data were used for the plot level validation because of the availability of highly accurate plot coordinates. The FFC dataset represents the ground truth from an intensive sampling of 51 stands with a network of 366 permanent plots (each 0.04 ha) measured in 2012 at the Michigan Tech's Ford Forestry Center, a research forest (about 1,400 ha) located in the Baraga County, Michigan. The coordinates of every plot center were determined using a Trimble GeoXH 6000 global positioning system and differential correction post-processing (via Trimble Pathfinder Office software) that resulted in an average horizontal precision of 1.50 m or less. The stands, roughly divided into jack pine (*Pinus banksiana*) and northern hardwood cover types have an average size of 22 ha. Most of these stands have been harvested more than once in the past 60 years.

The hardwood dataset comprised of sample measurements in 0.04 ha plots, established in 2010 and 2011, in 47 recently harvested (in 2006-2010) northern hardwood stands (area range 6.8 to 115.1 ha) throughout the study area. The sample measurements were made post-harvest and the pre-harvest inventory were derived using localized stem-to-breast height diameter prediction equation (Pond, 2012) . The Aspen dataset comprised of overstory sample data from 18 *Populus*-dominated stands (area range 1.9 to 16.1 ha) measured in 2012 at different locations of the central Upper Peninsula of Michigan. The aspen stands varied in age from 0 to 35 years.

The standing volume of the validation plots were calculated from the individual tree measurements following the algorithms adopted by FIA. The individual tree volumes were calculated using species-specific allometric equations for the Lake States as described in Miles and Hill (2010) and Woodall et al. (2010). This approach of volume calculation requires bole length estimation using models and coefficients from Ek et al. (1981) and Hahn (1984). The bole length is described as a function of stand basal area and site. Therefore, stand basal area was calculated from the plot inventory datasets and average site index values for individual species were estimated from tree lists available from the FIA DataMart (USFS, 2013). The individual tree volume estimates were summed to obtain plot level volume which were further summarized to the stand level estimates via up-scaling. The plot volume data were converted to above-ground biomass estimate by using species-specific expansion factors derived from the FIA database.

The county level estimates of biomass as a secondary dataset for the validation of landscape scale imputation estimates were obtained from the FIA database via the EVALIDator web-tool query (FIA, 2014) for the period 2005-2009.

2.3.6. Above-ground biomass modeling and accuracy assessment

The widely used novel extension of *k*-nearest neighbors (kNN) imputation (Tomppo et al., 2008; McRoberts, 2012) called random forest algorithm (Breiman, 2001; Crookston and Finley, 2008) was used to build empirical relationship between biomass and multisource spatial predictors for spatially extending the attribute across the study area. The kNN method is based on the premise that plots having similar spectral characteristics also have similar structural attributes. In the primitive form the kNN method estimates response variable at any unsampled target unit as the weighted average of the observed values from the nearest neighbors in the reference set. Several methods do exist to measure the nearness between the target and reference points on the basis of spatial covariates (Crookston and Finley, 2008; Latifi et al., 2010). The random forest (RF) based proximity is determined from the function of an ensemble of many classification and regression trees where each tree is built from a bootstrap sample of reference data and binary splitting of nodes in each tree is created with the best predictor selected out of a random subset of all predictors at each node (Liaw and Wiener, 2002). Conceptually, two observations are considered similar if they end up in the same terminal node of a tree, and the proportion of trees in the ensemble that place target and reference units in the same terminal node gives the distance measure. The RF based kNN (RF-kNN) imputation in regression mode predicts the response variable at any target point as the simple average of *k* nearest neighbors. Many recent studies have shown that the RF approach generally produces better results compared to other imputation methods (Hudak et al., 2008; Powell et al., 2010; Vauhkonen et al., 2010; Nelson et al., 2011; Ohmann et al., 2011; Coulston et al., 2012; Gleason and Im, 2012; Waske et al., 2012). The ability of internal cross-validation in RF also allows estimation of mean square error, and variable importance. In this study, the domain for nearest neighbor search for each target unit was the complete reference set since the auxiliary layers for the entire area were spectrally normalized using a standard procedure. The general advantage of imputation approach is that the prediction retains the natural variation as observed in the field measurements.

The RF-kNN imputation modeling was executed in the R statistical software (R Core Team, 2013) using the randomForest (Liaw and Wiener, 2002) and yaImpute packages (Crookston and Finley, 2008; Falkowski et al., 2010). Using different combinations of response and predictors, several RF imputation models were developed from the two reference data frames corresponding to the actual and fuzzed-swapped coordinates of FIA plots. Each model was based on 3,000 regression trees and the value of *k* parameter (i.e. number of nearest neighbors) set to 1 to maintain the natural variation in forest structure represented in the dataset. This mode of RF ensures that the imputed values at target points are exactly the same as one of the sample plots data in the reference set (rather than average of more than one plot). The details of the modeling approach with the actual coordinate data frame are described in Pond et al. (2014). The amount of (%) variation explained, mean square error (MSE), and bias for each of the models were compared to identify the best model to extend it spatially. The only model produced from the fuzzed-swapped coordinate data frame included eleven different biogeoclimatic variables to predict biomass. The RF variable importance ranking procedure was followed to include predictors in the models; the ranking depends on the criteria of proportional increase in the model's MSE when substituting random numbers for one predictor at a time while retaining others at actual values (i.e. permutation of predictors).

To scale up the FIA plot biomass spatially via the model from the actual coordinate data frame and also the model out of the fuzzed-swapped coordinate data frame, all the predictor rasters were converted to AsciiGrid format in ArcMap and were used as input to the two models in the yaImpute package of the R software. The model outputs in AsciiGrid format were finally converted to raster format, and the stand- and county-level summaries of biomass were obtained with the aid of the stand and county shapefiles in ArcMap.

The scatter plots of measured versus imputed total of plot-, stand-, and county-level biomass was produced and compared against a 1:1 line; the model predicting response mostly below the 1:1 line was inferred as negatively biased and vice-versa. The equivalence tests of inventory observation and imputation estimates as suggested by Robinson et al. (2005) and Robinson and Froese (2004) were carried out for the three spatial scales to evaluate the accuracy of the kNN technique at those scales. The accuracy was also measured in terms of root mean square error (RMSE), bias, and R^2. The experience from Finland and Sweden indicates that if the imputation estimates of inventory attributes are within 15 % of the measured values at stand level, then the results are reliable (Reese et al., 2002; Tomppo et al., 2008). Based on this premise, the smallest area for reliable estimate by imputation was explored.

The plot-, stand-, and county-level estimates of biomass obtained from generated imputation maps of this study were also compared with the estimates for the corresponding area derived from the extant biomass maps of the NBCD and Blackard et al. (2008). The county-level estimates from the four spatially explicit biomass layers were validated against the reference data derived via the FIA EVALIDator tool. The TukeyHSD (Tukey Honest Significant Differences) post-hoc test was also carried out to evaluate the significance of difference in the pairwise comparisons of mean estimates of the different methods at the three spatial extents. RMSEs and biases of the spatial inventory estimates based on the extant maps were calculated at stand-level based on the validation datasets.

Two comparisons were particularly emphasized to infer the utility of different approaches: (1) stand level comparison of imputation maps of this study with the map from NBCD, and (2) county level comparison of imputation maps of this study with the map by Blackard et al. (2008). If the NBCD maps provide better estimates than the imputation map of this study, then the new approach can be inferred as being constrained by the FIA privacy policy and the shortcomings of the optical remote sensing data; better results from NBCD can be justified because it used RADAR information. If the map of Blackard et al. (2008) and the new imputation maps provide similar results at the county scale, then the efforts at developing finer resolution maps for larger areas may not be necessary.

2.4. Results

The ranking of predictor variables used in the biomass imputation models based on the data from the actual and fuzzed-swapped coordinates of the reference plots are given in the Figures 2.2a and 2.2b that also represent the correlation of individual predictors with the biomass. It reveals that BAWHT is the most influential for both the models while elevation has very little explanatory power. The fuzzed-swapped data based model poorly explained the variations (% variance explained: 9.96) compared to actual coordinate data derived model (% variance explained: 32.23). It is also clear from the analysis that climatic variables, particularly mean annual precipitation, mean annual temperature, and growing season precipitation have important influence on biomass production and distribution (see Figure 2.2b).

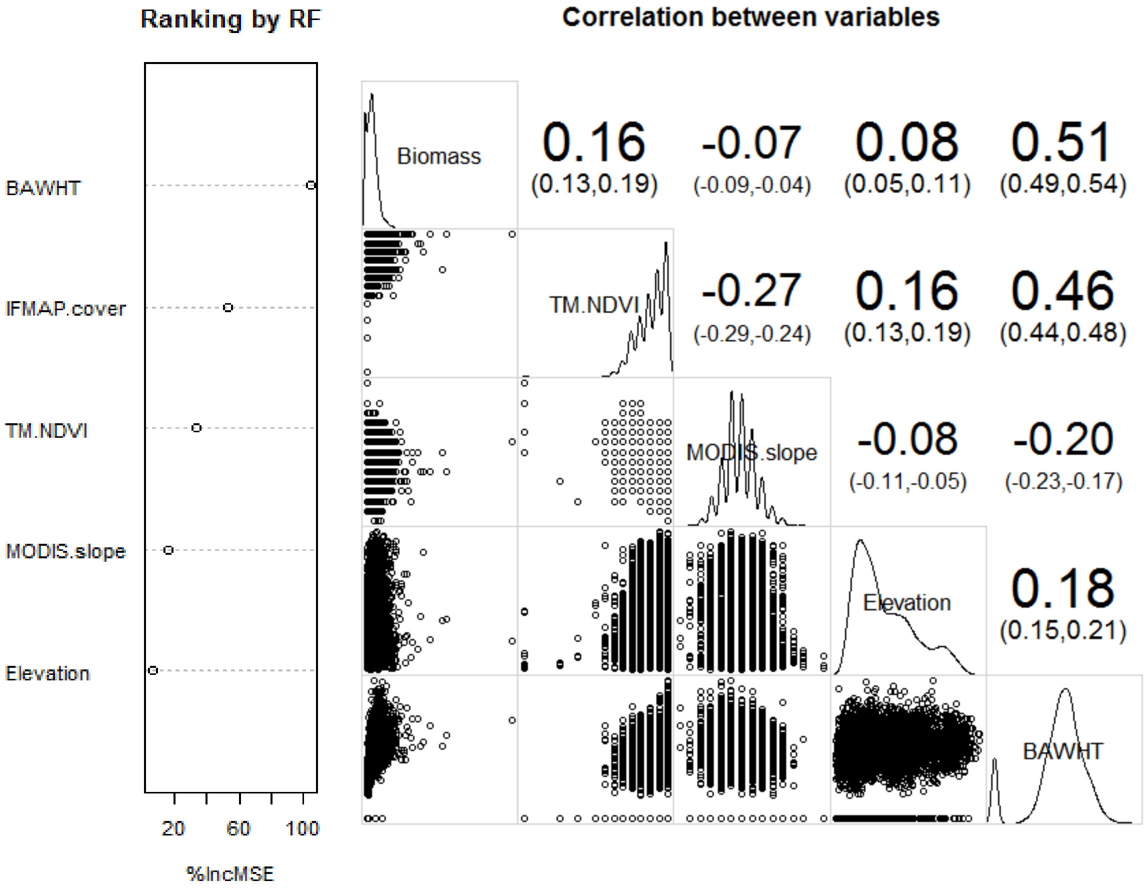


Figure 2.2a. Random forest based importance ranking (left) and correlation of predictors (right) used in biomass imputation modeling dependent on FIA database with actual plot-coordinate information.

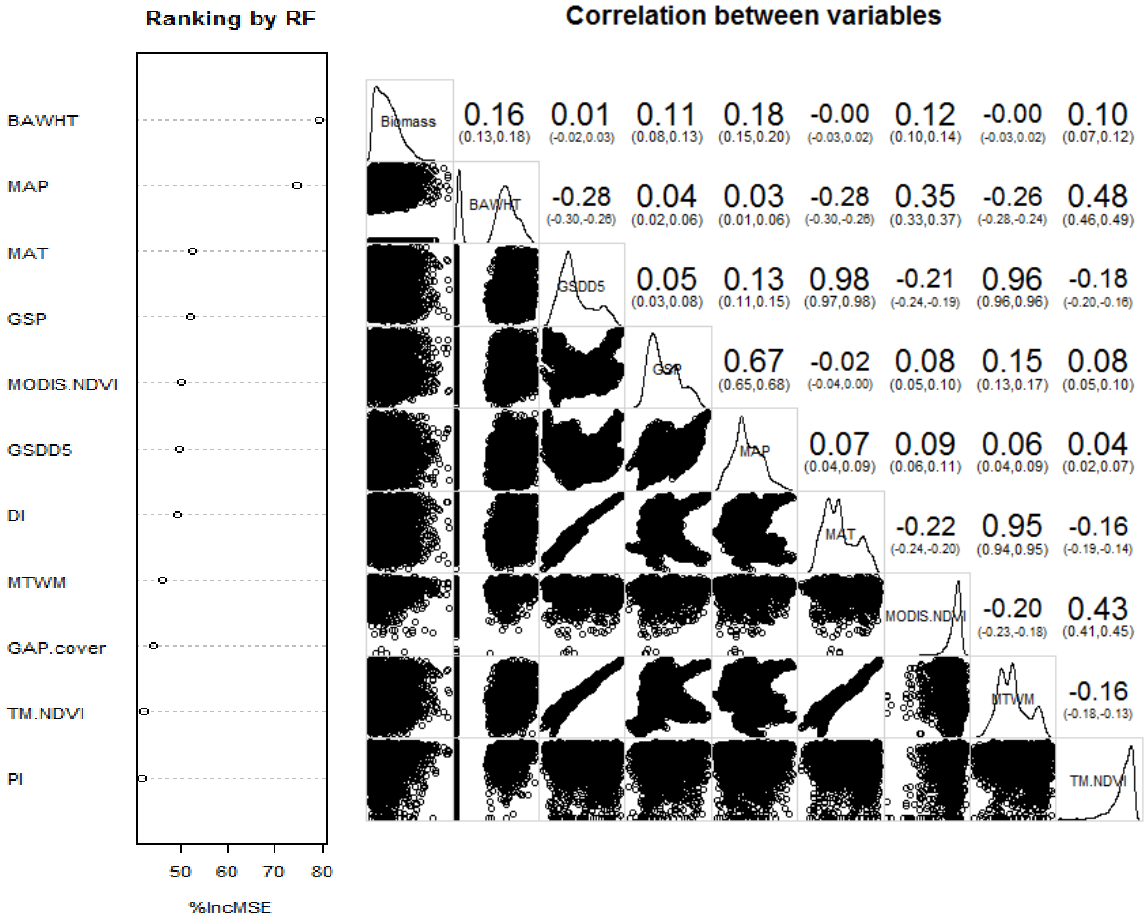


Figure 2.2b. Random forest based importance ranking and correlation of predictors used in biomass imputation modeling dependent on FIA database with fuzzed-swapped plot-coordinate information.

Plot-level imputation (average of 3×3 window of pixels) based on the newly formulated models and the extant models are compared in the Figure 2.3. From the box plot (Figure 2.3, right), it is clear that the existing models of NBCD and Blackard et al., 2008 (hereafter called USFS model) provide a narrow range of biomass predictions at the plot level compared to observed values which is more closely followed by the actual coordinate data derived imputation model, denoted by *Actu.imput* hereafter. All the models are negatively biased (i.e. do under prediction) in high biomass regions with density above 80 Mg ha^{-1} and positively biased in low biomass regions. As expected, the model derived from the fuzzed-swapped coordinate database (denoted by *Fuzz.imput* hereafter) performed poorly, while the model derived from the actual coordinate database is the best in terms of the linear correspondence between predictions and observations (Figure 2.3, left). However, none of the models are able to predict estimates statistically equivalent to the field measurements as evident from the result of equivalence test shown in the Figure 2.4. Further, the increasing spread of predictions on reference data towards higher biomass areas also signifies heteroscedasticity in the residuals which is a common nature of most spatial inventory systems.

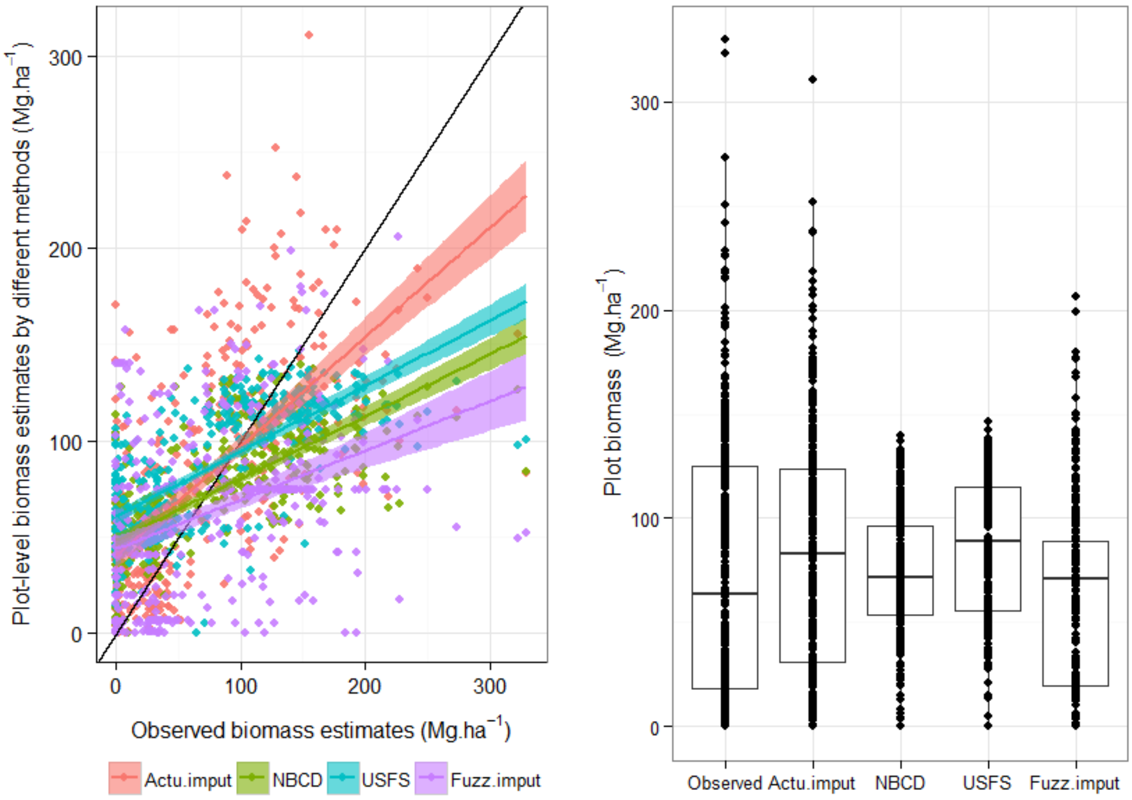


Figure 2.3. Comparison of plot-level biomass estimates by different imputation methods with the FFC inventory data**.**

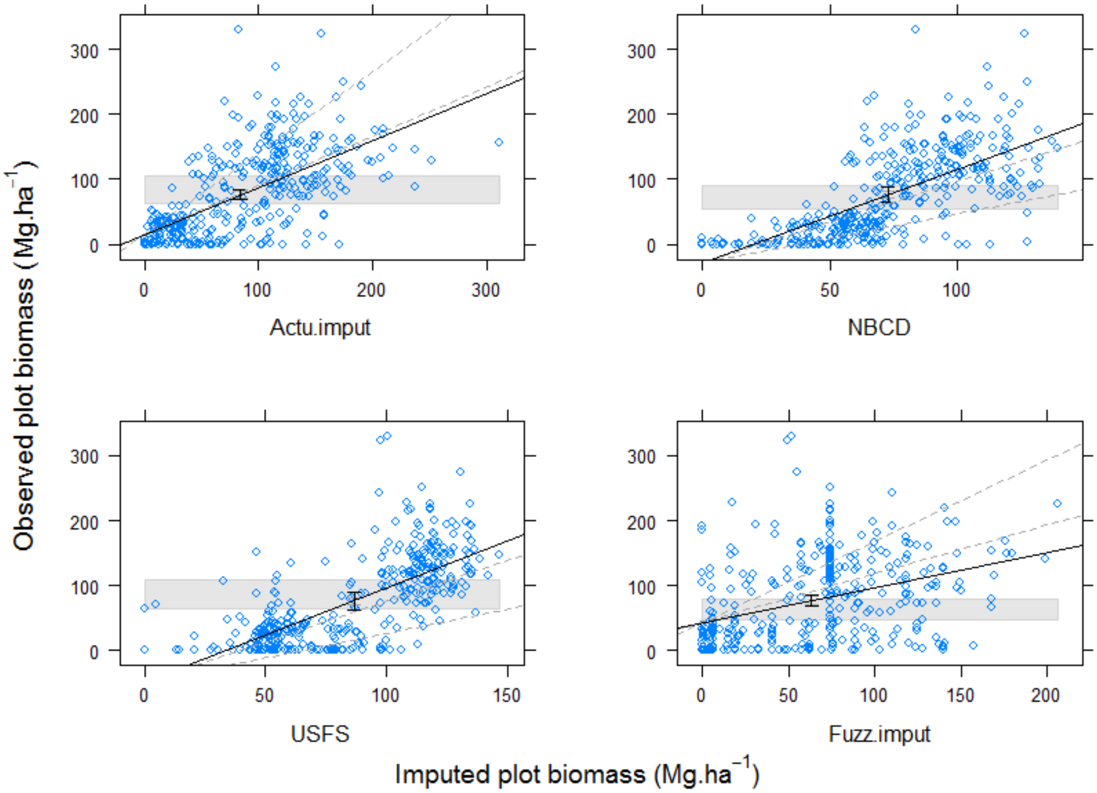


Figure 2.4. Equivalence plotting of the observed and imputed plot-level biomass estimates by the four methods. The black inclined line represents the line of best fit, the dashed gray lines represent 25% region of similarity for the slope, the shaded gray polygon represents 25% region of similarity for the intercept, and black vertical bar represents a confidence interval (at 5% alpha level) for the slope of the line of best fit.

The observed stand-level total biomass most closely matched the estimates based on the new imputation model developed out of the actual coordinate reference data frame (i.e. *Actu.imput* model); this is apparent in Figure 2.5 (left) as the 95% confidence interval of the fitted line includes the 1:1 line. The USFS and NBCD models were also found to be capable of producing estimates closer to the field observations; however, the model based on the fuzzed-swapped coordinate derived data frame (i.e. *Fuzz.imput* model) seemed unsatisfactory for stand-level biomass estimation. The insufficiency of the fuzzed-swapped model was verified from the equivalence test of the field observations against the imputation estimates as shown in the Figure 2.6. Except the *Fuzz.imput* model, the other models produced biomass estimates equivalent to the field observations.

A high degree of prediction variability was observed in younger stands due to low biomass but high canopy greenness (referred by NDVI) of the growing stock. It was particularly apparent in the young jack pine and aspen stands that have many trees smaller than the minimum threshold (10 cm) for dbh adopted in the field sampling.

Among the FFC stands, five were clear-cut harvested since 1995, and 15 were selectively harvested after 2006. Similarly, most of the hardwood stands were selectively harvested in between 2006-2011. So there is likely temporal mismatch between what was actually observed on the ground and what remote sensing device captured from space (e.g. Landsat derived NDVI were based on images from 2005-2010 while the field inventory data for validation were from 2010-2012). The stand level prediction error for the imputation methods were non-systematic with respect to stand size (e.g., larger stands did not necessarily had lower RMSE), hence it can be inferred that the imputation accuracy at stand level depends not only on the size but also the cover type and disturbance history. So, the initial goal of determining the minimum area for which imputation may provide a reliable estimate (say, within 15% of the observed biomass) is subject to additional information on stand age, management harvesting, and disturbance history besides a large range of stand sizes.

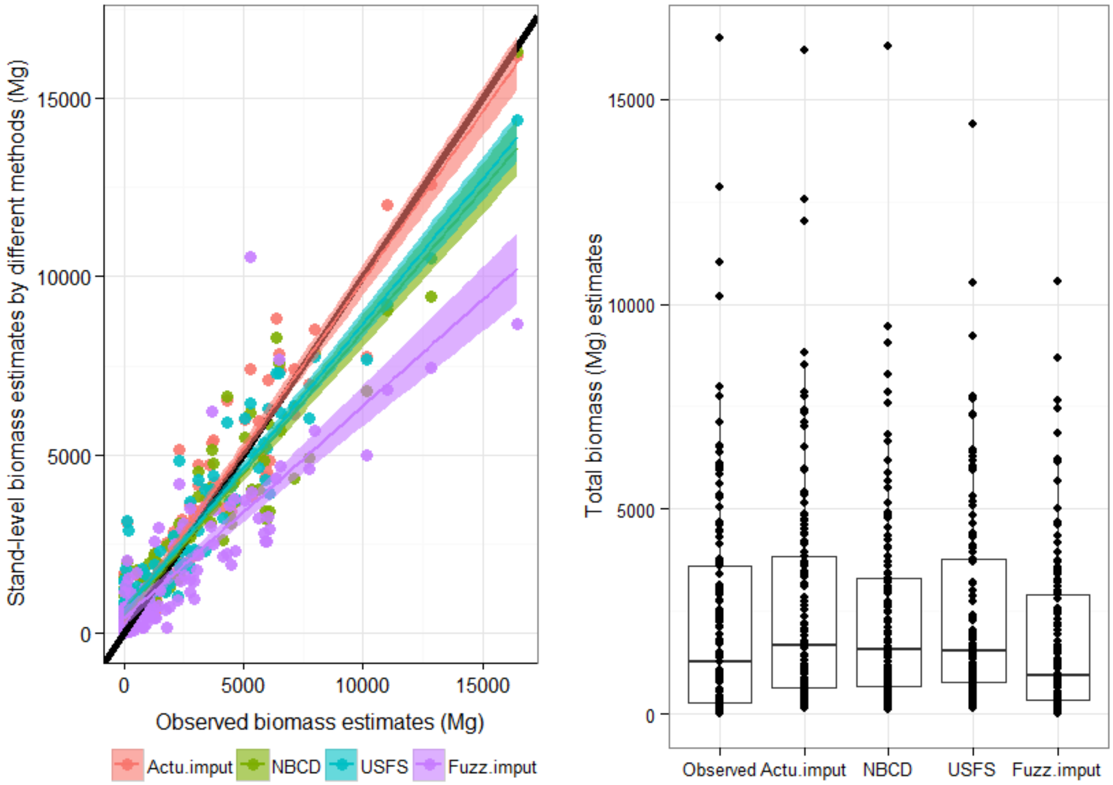


Figure 2.5. Comparison of stand-level total biomass estimates by the different imputation methods with the total estimates obtained from stand inventories from different locations of the upper Michigan.

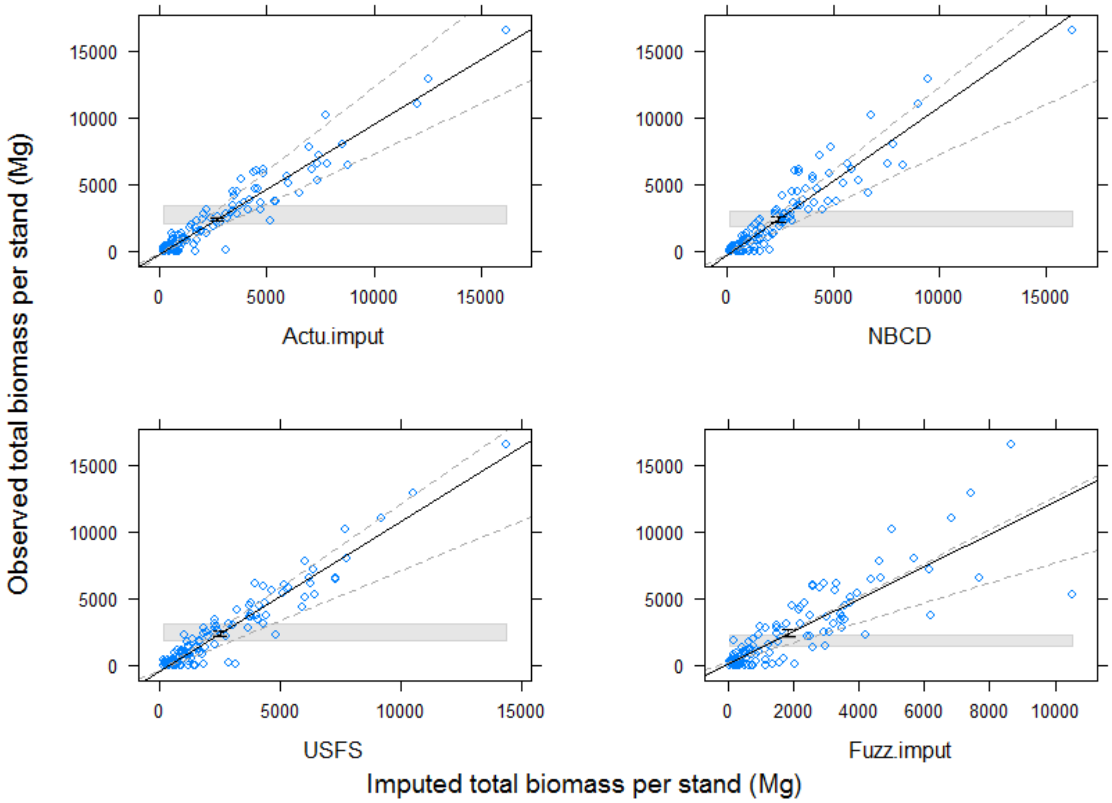


Figure 2.6. Equivalence plotting of the observed and imputed stand-level total biomass estimates by the four methods. The black inclined line represents the line of best fit, the dashed gray lines represent 25% region of similarity for the slope, the shaded gray polygon represents 25% region of similarity for the intercept, and black vertical bar represents a confidence interval (at 5% alpha level) for the slope of the line of best fit.

The county-level estimates are achievable to similar precision with any of the four models (see Figure 2.7); however, each of the models are generally producing over-estimation if we consider the FIA database with the county-level biomass estimate (via EVALIDator tool) as the accurate reference. The equivalence test (Figure 2.8) indicates that the NBCD model is providing the best correspondence of prediction estimates with the reference data, however, the *Fuzz.imput* and *Actu.imput* models are close competitors. In fact, application of large number of local samples in the training set of the fuzzed-swapped model may be attributed to less bias compared to the other models. This result implies that high degree of sophistication or adjustments to offset the spatial mismatch concern of sample plot data and remote sensing or geospatial predictors is not necessary during the reference data acquisition, processing, and analysis while undertaking large area biomass estimation.

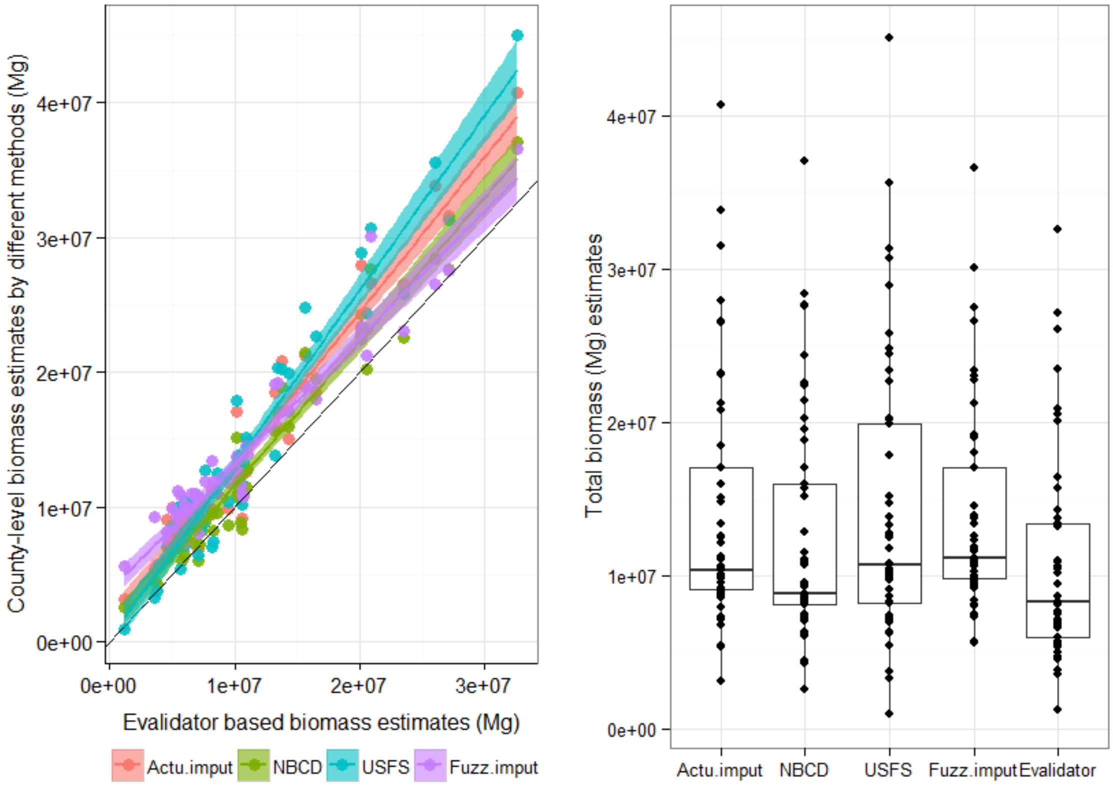


Figure 2.7. Comparison of county-level total biomass estimation by different imputation methods against the reference data obtained from the FIA database via EVALIDator tool**.**

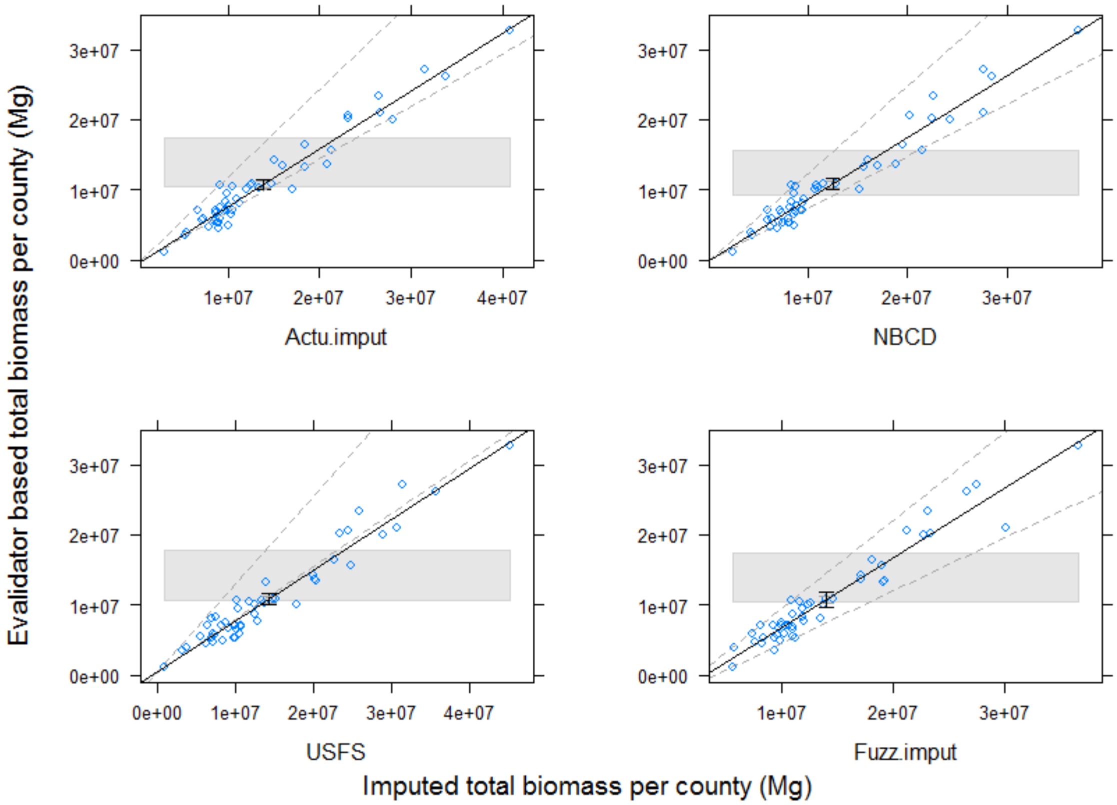


Figure 2.8. Equivalence plotting of the FIA estimated (via EVALIDator tool) and imputed county-level total biomass estimates by the four methods. The black inclined line represents the line of best fit, the dashed gray lines represent 25% region of similarity for the slope, the shaded gray polygon represents 25% region of similarity for the intercept, and black vertical bar represents a confidence interval (at 5% alpha level) for the slope of the line of best fit.

The highest coefficient of determination (R^2) was obtained for the county-level estimation by all the imputation methods compared to the validation data from FIA (Table 2.4). The R^2 values slightly decreased from county to stand-scale estimation but a large decline was found for plot-level estimates as compared to the field measurements. The plot-level RMSE was surprisingly lowest (58.19%) with the USFS model; however, the RMSE of *Actu.imput* model was lowest at the stand level and competitively similar at the county-level. Since the spatial distribution of the validation stands represent high diversity of conditions compared to the narrow capture of diversity in the validation plots (only at FFC), it can be concluded that the *Actu.imput* model is better. The smaller bias of NBCD model at all the scales can be attributed to the large sample size applied for model training and also inclusion of canopy height information from the InSAR data.

Table 2.4. Validation of the four models in terms of fit statistics at plot, stand and county scales

Spatial extent	Model	R^2	RMSE (Mg)	Relative RMSE (%)	Bias (Mg)	Relative bias (%)
Plot	*Actu.imput*	0.4103	53.97	64.35	7.33	8.73
Plot	NBCD	0.4628	50.61	69.59	**-3.82**	**-5.24**
Plot	USFS	**0.4942**	**50.36**	**58.19**	9.99	11.55
Plot	*Fuzz.imput*	0.1446	66.49	102.62	-13.33	-20.57
Stand	*Actu.imput*	**0.9164**	**896.07**	**33.38**	313.94	11.69
Stand	NBCD	0.8795	1035.15	43.25	**22.75**	**0.95**
Stand	USFS	0.9112	923.01	36.88	131.72	5.26
Stand	*Fuzz.imput*	0.717	1679.05	91.47	-529.16	-28.98
County	*Actu.imput*	**0.9497**	3732479.88	26.82	3111837.41	22.36
County	NBCD	**0.9431**	**2494049.01**	**19.97**	**1684509.06**	**13.49**
County	USFS	0.9347	4586442.74	32.25	-3418411.15	-24.04
County	*Fuzz.imput*	0.9298	3623739.84	25.98	3147890.45	22.56

The analysis of variance (ANOVA) test revealed that overall means of the estimation methods differ significantly only at the plot-level but not at the stand- and county-level at 95% confidence level. The TukeyHSD test showed that only four pairs of methods (namely, *Actu.imput*-Observed, NBCD-Observed, USFS-*Actu.imput*, and *Fuzz.imput*-NBCD) generate overall plot-level means that do not differ significantly with the pair of methods at 95% confidence level (Table 2.5a). All the pairs of methods produced stand- and county-level average estimates that are not significantly different at those scales (Tables 2.5a, 2.5b, 2.5c).

Table 2.5a. TukeyHSD test for difference of plot-level overall means by all pairs of methods

Pair of Methods	Mean diff (Mg)	Lower limit	Upper limit	p-adjusted
Actu.imput-Observed	7.3279	-2.5632	17.2189	**0.2553**
NBCD-Observed	-3.8175	-13.7085	6.0735	**0.8299**
USFS-Observed	9.9993	0.1083	19.8903	0.0460
Fuzz.imput-Observed	-13.3312	-23.2222	-3.4402	0.0022
NBCD-*Actu.imput*	-11.1453	-21.0363	-1.2543	0.0180
USFS-*Actu.imput*	2.6714	-7.2196	12.5624	**0.9477**
Fuzz.imput-Actu.imput	-20.6591	-30.5501	-10.7680	<0.0001
USFS-NBCD	13.8168	3.9257	23.7078	0.0013
Fuzz.imput-NBCD	-9.5137	-19.4047	0.3773	**0.0660**
Fuzz.imput-USFS	-23.3305	-33.2215	-13.4395	<0.0001

Table 2.5b. TukeyHSD test for difference of stand-level overall means by all pairs of methods

Pair of Methods	Mean diff (Mg)	Lower limit	Upper limit	p-adjusted
Actu.imput-Observed	313.9410	-607.5749	1235.4568	**0.8843**
NBCD-Observed	22.7516	-898.7643	944.2675	**0.9999**
USFS-Observed	131.7232	-789.7927	1053.2390	**0.9950**
Fuzz.imput-Observed	-534.8651	-1456.3810	386.6508	**0.5055**
NBCD-*Actu.imput*	-291.1894	-1212.7052	630.3265	**0.9096**
USFS-*Actu.imput*	-182.2178	-1103.7337	739.2981	**0.9829**
Fuzz.imput-Actu.imput	-848.8061	-1770.3220	72.7098	**0.0874**
USFS-NBCD	108.9716	-812.5443	1030.4874	**0.9976**
Fuzz.imput-NBCD	-557.6167	-1479.1326	363.8992	**0.4623**
Fuzz.imput-USFS	-666.5883	-1588.1042	254.9276	**0.2773**

Table 2.5c. TukeyHSD test for difference of county-level overall means by all pairs of methods

Pair of Methods	Mean diff (Mg)	Lower limit	Upper limit	p-adjusted
NBCD-*Actu.imput*	-1427328.34	-5706594.0	2851937.5	**0.8902**
USFS-*Actu.imput*	306573.74	-3972692.0	4585839.6	**0.9996**
Fuzz.imput-Actu.imput	36053.04	-4243213.0	4315318.9	**0.9999**
Evalidator-*Actu.imput*	-3111837.41	-7391103.0	1167428.4	**0.2697**
USFS-NBCD	1733902.08	-2545364.0	6013167.9	**0.7992**
Fuzz.imput-NBCD	1463381.38	-2815884.0	5742647.2	**0.8810**
Evalidator-NBCD	-1684509.07	-5963775.0	2594756.8	**0.8157**
Fuzz.imput-USFS	-270520.70	-4549787.0	4008745.1	**0.9997**
Evalidator-USFS	-3418411.15	-7697677.0	860854.7	**0.1848**
Evalidator-*Fuzz.imput*	-3147890.45	-7427156.0	1131375.4	**0.2586**

2.5. Discussion

Model formulation and validation are crucial for remote sensing based forest biomass assessment and its application. Contemporary spatial inventory systems for biomass estimation apply different modeling strategies contingent on scale of operation, required accuracy, data availability, funding and logistic support. High resolution data are more often used in small-area estimation and coarse resolution data for regional or national scale studies. The NBCD dataset is focused for the year 2000 and uses multi-source high resolution data with least temporal difference among the input variables. The InSAR data applied in the NBCD model accounts for canopy height, but the USFS model lacks height information. The inputs of USFS model also have a smaller temporal gap between the collection of field data and geospatial variables. The NBCD leveraged the SRTM 2000 data and hence is not replicable unless a similar data acquisition mission is conducted. The USFS method using MODIS data is reproducible though resulting in a coarser resolution output. Biomass estimation using MODIS data has limited success for small-area estimation because of the occurrence of mixed pixels which hinders integration of sample data from small sized plots with the remote sensing signatures from bigger pixels. Considering the potential of canopy height information and the linkage of biomass abundance with cover types, BAWHT and IFMAP layers were included as explanatory variables in the new models formulated in this study. But these two spatial layers were prepared at least 5 years before the FIA field data (collected in 2005-2009) and thus there is temporal mismatch.

The constraint of inadequate data for model training and validation hampered efficient modeling and data mining efforts. The fuzzed-swapped coordinates of the inventory plots in the FIA database are useful because as many predictor variables as available can be attached to the plot locations; the tradeoff is that the resulting reference frame suffers from spatial mismatch of the ground response and the feature explanatory variables. On the other hand, the true coordinates accessible only via FIA regional units allow joining of limited number of feature variables, often after reclassification, to the plot data. The FIA privacy policy led to compromise with predictive power of the auxiliary layers used in this study. Since it required grouping of NDVI and MODIS-slope rasters into 20 broader classes to pass the FIA security clearance with maximum number of reference data, the predictive power of the spatial layers was obviously reduced from what it would have been without the grouping. Further, the BAWHT values attached to the plots in the reference data matrix were not extracted from the raster used for spatially extending the model. These values were instead derived from the field measurements made by FIA at the inventory plots. Therefore, there is potential inconsistency between the raster values and the plot-level values of the two sources which may also introduce bias in the estimation. These constraints motivate evaluation of alternative modeling approaches for application to multiple spatial scales.

The way spatial inventory modeling was designed in this study is particularly important for small-area estimation where the areas are devoid of field inventory data. The spatial distribution of biomass obtained using the *Actu.imput* model suggest that the FIA database holds good promise for stand-level estimation provided that at least the restricted remote sensing and geospatial variables can be attached to the plot data. An advantage of using the FIA data is that it provides unbiased estimates as the sample units have random layout. The NBCD model did not produce better stand-level estimates than the constrained *Actu.imput* model. This implies that formulation and application of a model similar to the *Actu.imput* model is appropriate to update inventory information for operational planning.

The *Actu.imput* model validation results are consistent with previous studies. The pixel-plot level accuracy of estimates was least and county-level estimates were the best. Published works using kNN have shown that pixel level accuracy of forest attribute estimations is low, but for larger areas more acceptable accuracy is reached (Nilsson, 2002; Tomppo et al., 2002; Holmstrom and Fransson, 2003; McRoberts et al., 2007). For example, Reese et al. (2002) found low accuracy at the pixel level (58–80% relative RMSE for standing volume), and better accuracy over larger areas, with the best result of 10% relative RMSE over a 100 ha aggregation. Similarly, Chirici et al. (2008) reported 44-63% and Fazakas et al. (1999) reported about 74% relative RMSE against the measured mean standing volume. The under prediction in high biomass areas and over-prediction in low biomass areas is pursuant to Baccini et al. (2004) who observed under-prediction above 250 Mg ha^{-1} and over-prediction below 45 Mg ha^{-1}.

The relative importance of climatic variables in the *Fuzz.imput* model (Figure 2.2b) can be explained on the basis of Zheng et al. (2007) who reported increasing density of biomass from west to east and north to south that respectively follows increasing precipitation and temperature trend in the Lake States. Baccini et al.(2004) also observed positive association of total annual precipitation and biomass.

The weak association between plot-level measured and imputed values can also be justified on the basis of FIA plot design. The layout of the FIA plots is such that the four subplots (each 7.3 m in radius) are spread over a minimum of 4 pixels within a 3×3 window. The reference plot estimate of biomass used in this study was based on averaging and up-scaling of the values from the four sub-plots. As there is not direct correspondence between single pixels and FIA plots, an average of 3×3 window from the biomass output raster was used for cross validation with the field measurements in the FFC. This resulted in a higher R^2 value than when extracting values from single pixels. Another source of uncertainty in the estimation was the difference in the years of FIA measurements and the years of Landsat image acquisition.

The county-level validation data retrieved directly from the FIA database were the estimates of growing-stock on forest land only that includes timberland, reserved forest land, and other forest land. The definition of forest land in FIA protocol is set to the criteria of "at least 0.405 ha in size, 36.58 m continuous canopy width and 10% stocking where understory is not disturbed by non-forest land use such as agriculture or residence" (Blackard et al., 2008). The FIA plots, according to current design, can also include areas having <10% crown cover (e.g., clear-cut) and the imputation models of this study also considered such plots and gave predictions for both forested and non-forested pixels. As expected our results show that the total of the imputed values for most of the counties are above the reference values as shown by the distribution of points above 1:1 line in the Figure 2.7. This means that the imputation models are also predicting biomass at some cover types that are actually not forest. Use of a mask to exclude non-forest area and running imputation only for the forested region was not deemed necessary as the reference frame for model training included a large number of samples representing the full range of cover types of the study area. Although *Actu.imput* model was superior to the USFS model in terms of R^2, RMSE and bias at the county-scale, high efforts in developing fine resolution maps for large-area estimation is not necessary as a model generated out of fuzzed-swapped coordinate database also provided acceptable estimates at the county-level. The superiority of *Actu.imput* model may be because the FIA plot size is comparable to the pixel size of Landsat but not to the size of MODIS.

In general, spatial models have an inherent characteristic of being area specific, i.e. a model extended beyond the region of reference data may provide biased estimates. The NBCD and USFS models are for the entire conterminous USA but the models produced in this study are specific to a region of the Lake States so its application outside the study area even within the Lake Sates requires further scrutiny. The factors causing uncertainty in biomass prediction are described in Lu et al. (2012). The complex forest structure and composition across the landscape, use of improper allometric equations, and non-linear relationship between biomass and canopy cover are some factors hindering accurate mapping. For example, biomass continues to accumulate in trees even after canopy closure of forest, limiting the extent to which optical reflectance from canopy can be used to estimate biomass. The performance (or superiority) of models varies with scale of validation data, and choice of statistical measures derived from the prediction and validation datasets.

2.6. Conclusions

i. Although the restricted FIA plot data based imputation model (*Actu.imput*) provided better plot level estimates, none of the evaluated models can be applied for the plot level biomass prediction because none of the prediction estimates were statistically equivalent to the field based observations.

ii. Stand-level biomass estimate is most accurately provided by the *Actu.imput* model in terms of RMSE. The NBCD and USFS models are also capable of producing estimates closer to the field observations, but *Fuzz.imput* model derived from the fuzzed-swapped coordinate FIA database is not appropriate.

iii. A high degree of prediction variability was observed in younger stands with all the models, mostly because of temporal mismatch between remote sensing and field data.

iv. The county level estimates can be satisfactorily obtained with any of the tested models. A high degree of sophistication and adjustments to offset the spatial mismatch concerns of field plot data and remote sensing data is not necessary in modeling and mapping. This is because even the *Fuzz.imput* model generated results statistically equivalent to reference data.

v. The performance of models varied with the size of target area, choice of statistical measure to test goodness-of-fit, and the quality of calibration and validation data.

2.7. Suggested further study

Addition of more stands of known age, known disturbance history, and larger size that hold large trees above the minimum dbh threshold (10 cm in this study) could facilitate identifying the minimum area for reliable prediction of biomass by the imputation methods of operational value. The integration of un-binned NDVI and MODIS-slope and single panel (one year measurement) FIA data closest to the year of acquisition of the remote sensing images can be expected to improve the prediction accuracy, and this needs to be tested. The hypothesis that reducing temporal mismatch between remote sensing, geospatial data and field data used in model formulation could reduce prediction error could be verified through a similar study. Unsupervised or supervised classification of satellite image can be substituted for the outdated IFMAP layer.

3. Integration of variable radius plot and LiDAR data for multi-stand imputation and mapping of forest attributes[2]

3.1. Introduction

Assessment of forest structural attributes such as growing stock volume and biomass is essential for understanding ecosystem productivity and carbon cycling (Gleason and Im, 2011). Spatially explicit mapping of biomass has especially gained more attention in the frameworks of international conventions on climate change mitigation and sustainable forest management (Luther et al., 2006; FAO, 2009, 2010). In addition, accurate and cost effective spatial inventory information is demanding in operational forest management planning.

The traditional field-based techniques of generating inventory information on structural parameters via the use of sample data to extrapolate attributes over broader areas are often constrained by time and resources. Therefore, improved tools and techniques are continually being searched and developed to achieve better inventory accuracy in a cost-efficient manner. As intensive management practices demand low-cost and high resolution structural information especially for stand level monitoring, a customary practice is to integrate remote sensing (RS) data with a sparse network of field plot measurements (Wulder et al., 2008). Fundamentally, any geospatial inventory technique requires predictive models developed from a reference data frame based on samples of field measured response and co-located remotely sensed predictor variables across the entire area of interest. The formulated empirical relationship between field and remote sensing measures is then applied at spatially contiguous pixel units where only predictor variables are known across the entire target area.

A wide range of RS data and analysis techniques have been applied to augment *in situ* sample inventory data with the intent of providing timely, unbiased and cost-efficient assessments of forest structural attributes over progressively large spatial extents (Falkowski et al., 2006; Zhao et al., 2009; Gleason and Im, 2011). Indeed, RS data from both passive and active sensors have the capability to supplement the traditional approach of forest inventory and assessment (Lu, 2006; Song, 2012). Some satellite data (e.g. Landsat) are readily available globally and can be useful for estimating forest attributes. However, when considering diverse forest structures, Landsat and similar optical RS data are constrained by the fact that reflectance signals saturate in high biomass areas (Lefsky

[2] This chapter is ready to submit in a remote sensing journal with Ram K. Deo as the first author and Michael J. Falkowski and Robert E. Froese as the second and third authors respectively.

et al., 2002; Lu et al., 2012). Light detection and ranging (LiDAR) technology has demonstrated potential for addressing weaknesses inherent in the optical RS because LiDAR signals can penetrate the canopy gaps and directly measure horizontal and vertical profiles of canopy and terrain (Lefsky et al., 2002). Consequently, contemporary forestry research on structural assessments tends to leverage LiDAR for improved detection of 3-dimensional forest characteristics. Several studies have demonstrated high correlation between LiDAR data and forest structural attributes (Van Aardt et al., 2006; Wulder et al., 2008; Gleason and Im, 2012). LiDAR data is often leveraged in geospatial forest inventories because the sensor directly measures vegetation height, sub-canopy topography (i.e., elevations) with high degrees of accuracy and precision even in closed canopy and inaccessible forests. Indeed, the application of LiDAR data in forest resource inventory is rapidly expanding in the last 20 years (Hudak et al., 2009). In the future, regional or national level assessments may be realistic given the increasing availability, resolution, and coverage of sensors and acquisitions (Hudak et al., 2009; 2012). Further, incorporating LiDAR data into an operational forest inventory has been found to improve cost-efficiency as compared to the traditional field based approach (Hummel et al., 2011) and the advantages of LiDAR data may supersede the difficulties posed by cost-intensive field campaigns (e.g., sampling in remote areas).

LiDAR data are well suited to characterize horizontal and vertical attributes of forest canopies and underlying terrain (Wulder et al., 2012). The point cloud data obtained from LiDAR instruments accurately represents the elevations of vegetation and the ground surface (Mitchell et al., 2011; Sun et al., 2011). The vertical profile of the point cloud can be separated into ground and non-ground returns and a digital elevation model (DEM) at an appropriate resolution can be created (e.g. via nearest neighbor, kriging, or spline interpolation methods) based upon the ground returns only. The normalized point cloud (difference between individual point elevation and the DEM) then characterizes the vertical and horizontal distribution of vegetation in a forest, and various metrics representing statistical distribution of canopy height, canopy cover, strata density, and strength of near infrared return signals (i.e. intensity) can be derived (Gobakken and Naesset, 2008; Hudak et al., 2008). Such metrics have been used extensively as explanatory variables in many studies employing empirical or semi-empirical models under parametric or non-parametric frameworks for the prediction of various forest biophysical attributes (Goerndt et al., 2010; Pesonen et al., 2010; Gleason and Im, 2012). However, performance of the predictors varies with data characteristics, forest type, sampling design, and the modeling approach employed (Chen et al., 2012; Vincent et al., 2012). LiDAR derived metrics have been used successfully in previous studies for the prediction of biophysical parameters such as biomass (Lefsky et al., 2002; Popescu et al., 2011; Straub and Koch, 2011; Chen et al., 2012; Gleason and Im, 2012; Nelson et al., 2012), standing volume (Nilsson, 1996; Van Aardt et al., 2006; Straub et al., 2009; Latifi et al., 2010; Tesfamichael et al., 2010), stand basal area (Hudak et al., 2006; 2008; Vincent et al., 2012), tree density (Hudak et al., 2006; 2008), crown diameter (Popescu et al., 2003), DBH (Salas et al., 2010), LAI (Solberg et al., 2009), and vegetation structural development stages (Falkowski et al., 2009).

Any LiDAR assisted forest inventory, similar to other RS based methods, involves empirical model building by relating dependent variables measured in sufficient number of sample plots with coinciding LiDAR derived predictor metrics. Hence, in addition to LiDAR data acquisition, a significant portion of the total inventory cost is associated with field sampling measurements. The conventional approach for integrating LiDAR and field sample plot data is to use inventory parameters from fixed dimension plots, with the size of the plots approximately equal to the spatial resolution (i.e., grid size) at which LiDAR predictor variables are calculated (Hudak et al., 2008). However, forest inventory based on fixed dimension plots can be costly since that entails detailed measurements of every tree within the plot boundary. In addition, statistical validity of the models requires that the sample plots should characteristically represent the full coverage of forest structural variability across the area of interest. Since the size, number (intensity), and distribution of sample plots is directly related to the cost of inventory, resource managers often have vested interest in inventory protocols that maintain sampling efficiencies and produce accurate estimates of attributes. Nonetheless, increasing the sampling intensity is one solution to improving the representation of forest variability in the sample dataset. A commonly applied strategy to increase the sampling intensity is to employ variable radius plots (VRP), or point sampling. VRP sampling is quick, unbiased, and easy to implement as the sample trees are selected with probabilities proportional to their basal area and inverse-distance from the plot center (Avery and Burkhart, 1994). The technique basically requires the cruiser to stand at a point, view every tree at the breast height level (1.37 m) through an angle gauge (prism or Relaskop) in a 360^0 sweep, and count only those trees the bole of which completely covers the projection angle of the device. The technique depends on a pre-selected angle gauge that corresponds to a constant basal area on per unit forest-area basis, called basal area factor (BAF), for each tally tree regardless of the DBH. The conventional practice is to apply a single BAF to cruise one stand. The choice of BAF in an operational inventory depends on tree size distribution and density. A smaller BAF generally corresponds to a larger coverage of unknown and inconsistent area and results in more tally trees per point compared to a larger BAF (Reed and Mroz, 1997); smaller BAFs have risk of missing or double counting sample trees in dense stands but reduce the likelihood of edge effect (White et al., 2013). Some prior publications recommend selection of BAFs that provide an average of 4-8 tally trees per point (Reed and Mroz, 1997; USFS, 2000). The tallied trees need also to be measured for

DBH (also total height for better results) if the objective is computation of tree density or standing volume. This type of sampling strategy is especially useful for timber inventory as efforts are focused on larger trees that hold the most volume and value. The canopy height and density information from LiDAR data can be useful to determine an optimum BAF for multi-stand sampling and modeling. Currently, studies that integrate VRP data with LiDAR data are limited.

Integrating VRP data into a LiDAR based forest inventory is problematic because size of the sample plots is indeterminate and inconsistent even for a specified BAF. In other words, spatial mismatch issues arise when formulating prediction models dependent on LiDAR derived metrics with a fixed spatial resolution (i.e. raster grids). The variable size of input field plots is the major source inducing uncertainties in the models. Thus a key challenge for improving the integration of VRP and LiDAR data is finding the optimum grid size to which LiDAR data be binned so that inventory parameters of a VRP best matches with the coinciding LiDAR derived metrics (Golinkoff et al., 2011). The purpose of finding an optimal size is to reduce the variability and spatial mismatch between the LiDAR data and plot measured attributes (Hollaus et al., 2007; Jochem et al., 2011). Hollaus et al. (2009) selected an approximate grid size of LiDAR metrics by analyzing only four different arbitrarily selected plot-diameters (16, 20, 24, and 28 m) while Kronseder et al. (2012) used a one hectare circular area for each plot to extract LiDAR metrics, citing that the VRP method provides attribute estimates on per hectare basis for each plot. Hollaus et al. (2007) also used five different circular areas to extract the LiDAR data, and evaluated the impact of five resolutions of predictor metrics on response variables to decide an approximate (or average) VRP size. Van Aardt et al. (2006) coupled LiDAR distributional parameters on per segment basis (segment derived from canopy height model) with multiple VRPs per segment for modeling and mapping of volume and biomass. Gobakken and Naesset (2008) evaluated the effect of different sized fixed dimension plots on the accuracy of a LiDAR based inventory and noticed that the effect varies with canopy structure and stem density. Some authors caution against integrating LiDAR with VRP data simply because of the concerns for mismatch between field-measured attributes and corresponding fixed resolution LiDAR metrics (Laes et al., 2011).

A large area inventory across multiple stands using VRP sampling may involve several BAFs that in principle vary with stand structure. However, Reed and Mroz (1997) have indicated that foresters often prefer to work with a compromise BAF for multi-stand inventory to avoid practical difficulties that arise due to change in limiting distance for a given tree size with a change in BAF. Indeed, an efficient strategy for multi-stand inventory and assessment dependent on point sampling and LiDAR data would be to apply a common BAF suitable for all target stands. This strategy would support prompt resource assessment goals as a single BAF may allow the use of a single grid size of LiDAR metrics for the entire target area, and hence less time for LiDAR data processing. An intuitive approach for leveraging LiDAR data with multi-stand VRP data would be trial-based, where several LiDAR samples of varying size can be extracted successively at each VRP location, and then models can be developed at various resolutions by associating the LiDAR samples with the inventory attributes obtained from point sampling. Further strategy to facilitate modeling could be grouping of the plots cruised with the same BAF and formulating model for each size (resolution) of LiDAR samples.

3.2. Objectives

The goal of this study was to assess the efficacy of supporting LiDAR based forest inventories with VRP data, and subsequently to develop an effective methodology for integrating LiDAR and VRP data to perform a large area assessment of standing volume. The general research questions include:

- Can VRP data be effectively integrated with LiDAR data to improve the efficiency of geospatial forest inventories?
- Can VRP data be substituted for fixed radius plot (FRP) data in the case where insufficient FRPs exist?
- How does a LiDAR-based stand level inventory modeled using VRP data compare with the field measurements?

3.3. Methods

3.3.1. Overall approach

The VRP data based modeling and mapping approach involved four principal steps. First, the VRP inventory data was collected by using a small angle gauge of BAF 1.15 m^2 ha^{-1}, denoted hereafter as BAF 5 (corresponding to the imperial units) for easy spelling. Second, the BAF 5 plot data was re-processed iteratively to derive data for larger BAFs since a smaller BAF corresponds to a larger plot size (but indeterminate area) with more tally trees. Third, LiDAR point cloud data were extracted at each plot locations for a range of radii between average and maximum limiting distances (see sections 3.3.3 and 3.3.5) that depend on the BAF and DBH distribution of each plot. Then ninety potential LiDAR metrics were generated and models were fitted separately for each sample size; the size yielding the best fit statistics were inferred the optimal plot size. Fourth, all tiles of LiDAR data for the area of interest were processed and predictor metrics were generated at spatial resolutions corresponding to the optimal plot size of the two best VRP models, and finally the models were extended spatially to develop standing volume distribution maps.

The FRP data based modeling and mapping followed the conventional procedures where LiDAR data was processed and ninety metrics were prepared to the resolution of plot diameter (i.e., 22.6 m). Then, the metrics were attached to each plot data and a model was fitted which was finally extended spatially across the study area.

Thus three volume maps, two corresponding to the best two VRP models and one corresponding to the best FRP model, were produced. The volume estimates at the plot-level by the FRP and VRP models were compared at the last.

3.3.2. Study area

The study was carried out in six conifer stands (Figure 3.1) of Michigan Technological University's Ford Forestry Center (FFC), located in the western Upper Peninsula of Michigan, U.S.A (Latitude 46°37'N, Longitude 88°29'W). The total FFC area is approximately 1400 ha and has been divided into 54 stands with an average size of 22 ha. The stands have been subject to various management activities since 1954. The Ford Forest is predominantly occupied by jack pine and hemlock-northern hardwood cover types, but also contains smaller areas of quaking aspen and natural (fire-origin) red pine. The dominant overstory tree species in the stands are jack pine (*Pinus banksiana*), sugar maple (*Acer saccharum*), red maple (*Acer rubrum*), eastern hemlock (*Tsuga canadensis*), and yellow birch (*Betula alleghaniensis*). The minor overstory species in the stands include red pine (*Pinus resinosa*), white pine (*Pinus strobus*), quaking aspen (*Populus tremuloides*), black cherry (*Prunus serotina*), American basswood (*Tilia Americana*), American elm (*Ulmus americana*), ironwood (*Ostrya virginiana*), eastern white pine (*Pinus strobus*), red pine (*Pinus resinosa*), balsam fir (*Abies balsamea*), white spruce (*Picea glauca*), northern white cedar (*Thuja occidentalis*), black spruce (*Picea mariana*), and tamarack (*Larix laricina*). Elevation of the target stands range from 359- 425 m above sea level and the soils primarily belong to the orders Spodosols and Entisols, originated out of glacial outwash, with varying nutrient status where mesic to dry silt loams support plant communities typical of northern hardwood complex, and xeric sandy areas harbor jack pine dominant overstory (Berndt, 1988).

The area of interest differs by age, stoking, species composition, and complexity. The six target stands (Figure 3.1) include one jack pine dominated old protected reserve, four pure even-aged young jack pine stands, and one mixed uneven-aged red pine dominant stand. The stands have not been harvested since 2007; however, selective harvesting did occur in two of the six stands in 1991 and 2007.

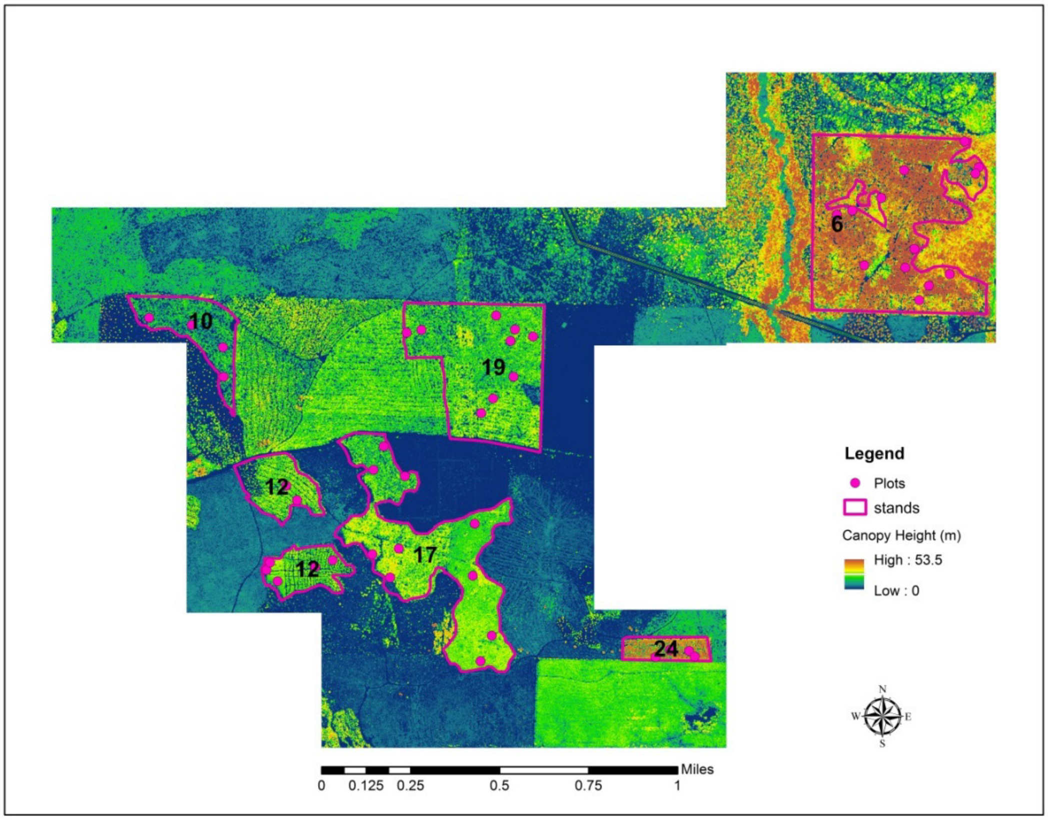


Figure 3.1. The six target conifer stands and sample plot locations overlaid on the canopy height model derived from the LiDAR data. The stands belong to different age, size (height, and diameter) and stocking classes.

Table 3.1. Characteristics of the sample stands based on overstory measurements

Stand ID	No. of plots	Area (ha)	Max DBH (cm)	QMD* (cm)	Trees per ha	SBA$^{\phi}$ $m^{2}\cdot ha^{-1}$	BAWHT † (m)	Remarks
6	13	50.29	82.0	26.9	515	29.43	18.7	red pine dominant
10	4	10.14	23.8	17.8	412	10.29	12.36	pure jack pine
12	7	14.22	23.6	5.8	803	13.82	11.12	pure jack pine
17	10	33.09	32.2	15.9	840	16.77	11.76	pure jack pine
19	9	32.12	27.6	15.3	961	17.74	11.35	pure jack pine
24	4	4.08	35.5	22.6	729	29.50	14.92	jack pine dominant

* QMD: quadratic mean diameter; ϕ SBA: stand basal area; † BAWHT: basal area weighted canopy height

3.3.3. Field inventory data

A field inventory was carried out in the summer 2012 over a network of fixed radius plots (FRP, 0.04 ha in size) in the stands. The permanent plots were distributed randomly across the study area in a stratified sampling design. The number of plots per stand ranged from 4 to 13 depending on stand size, density, and heterogeneity so that the minimum intensity was 1 plot per 3.8 ha (Table 3.1). The plots distribution intensity was based on a maximum sampling error objective of 20%. The FRP inventory dataset was supplemented with VRP inventory in September 2013 at each of the existing plot locations by using a prism of BAF 5. Data from both fixed radius and variable radius sampling techniques were obtained for a total of 47 plots in the target stands. The VRP inventory was made with a smaller BAF of 5 in order to obtain a large number of sample trees per plot so that a simulated inventory with larger BAFs result in at least four tally trees per plot. In the FRP sampling, species and DBH of each tally tree 10 cm or greater were recorded in addition to the ground slope, aspect, elevation, and coordinates of each plot. The coordinates of every plot center were determined using a Trimble GeoXH 6000 global positioning system and differential correction post-processing (via Trimble Pathfinder Office software) that resulted in an average horizontal precision of 0.80 m. In addition, total height of the smallest and largest trees (by DBH) of each species occurring within the fixed-size plots was measured using a Haglof Vertex Laser VL400 Hypsometer. In the case of the VRP sampling with BAF 5 prism, species and DBH of tallied trees 10 cm or greater, as well as the distance of each tree from the plot center, were measured. A laser hypsometer, fixed on a tripod over the plot center, was used to measure the distance (at breast height level) of the tallied trees. As a protocol, the sample tree measurement at each plot was started from due north to avoid measurement bias.

From the BAF 5 measurements, VRP data for additional six different BAFs of 1.60, 2.06, 2.29, 2.75, 3.21, and 3.44 m^2 ha^{-1} tree^{-1} were simulated; these six factors are hereafter denoted by BAF 7, BAF 9, BAF 10, BAF 12, BAF 14, and BAF 15 (corresponding the imperial units) respectively, for ease of spelling. The simulation was based on the comparison of measured horizontal distance of a tree from the plot center to the calculated limiting distance (R) (see Equation I). The limiting distance is the maximum horizontal distance from plot center to the face of a tree of given DBH such that the tree would still be considered “in”. If the measured distance was less than or equal to the limiting distance then the subject tree was considered to be “in” for the selected BAF. The limiting distance (R) was calculated as:

$R = PRF \times DBH$; Where $PRF = Plot\ Radius\ Factor = \frac{8.6962}{\sqrt{BAF}}$ ……….. Equation I

Tree basal area, volume, and other metrics corresponding to both FRP and VRP data were separately computed using standard procedures. The individual tree volumes were calculated using the species-specific equations for the Lake States as adopted by the FIA program (O'Connell et al., 2013) and detailed in Miles and Hill (2010) and Woodall et al. (2010). This approach of volume calculation requires bole length estimation using models and coefficients from Ek et al. (1981) and Hahn (1984). The bole length is described as a function of stand basal area and site. Therefore, stand basal area was calculated from the FRP inventory data and average site index for individual species were estimated from the tree lists available online via the FIA DataMart tool (USFS, 2013). The individual tree measurements from the FRPs were summed to obtain plot level estimates which were further summarized to the stand level estimates through up-scaling. Volume on a per unit area basis at each plot was calculated by summing the volume of individual trees, multiplied by appropriate sampling weight. In the case of FRPs, the sampling weight is a constant (1/plot size) for each tree. For VRPs, the sampling weight is a function of DBH and is calculated as individual-tree BA divided by BAF (Avery and Burkhart, 1994).

3.3.4. LiDAR data and processing

LiDAR data for the area was collected in June 2011 by Aerometric, Inc. (Sheboygan, WI, U.S.A.) using a RIEGL LMS-Q680i airborne laser scanner onboard a helicopter flown at an altitude of 457 m with a ground speed of 60 kts. The LiDAR system operated at 1550 nm near infrared wavelength with pulse frequency of 400 kHz and scan angle of ±30^0 from nadir, and generated a point density of 18 pulses per square meter and captured up to 9 returns per pulse. The sensor has ability to scan up to 200 lines per second with effective measurement rate (of coordinates) upto 266 kHz. The vendor provided data as discrete return point cloud in numerous tiles in *.las* format. The multi-return dataset was analyzed and classified in FUSION software (McGaughey, 2014) to produce information about above ground forest structure as well as the bare-earth surface. The ‘ground filter’ tool was used to separate ground returns out of all returns in the high-density point cloud. The default coefficients for the weight function (described in McGaughey, 2014) and a tolerance value of 0.03 m (0.1 ft) after 10 iterations were specified in the filtering process to properly screen out non-ground returns. A high resolution (1.5 m) grid surface (i.e. DEM) was then created out of the filtered ground returns and ultimately applied to normalize the raw LiDAR point cloud so that the remaining points represent the elevations of canopy elements above the ground. Since LiDAR acquires three-dimensional information on forest structure at all possible strata including tops and understory, all non-ground returns from each pulse were used to derive predictor metrics. The non-ground LiDAR returns were clipped for 24 different radii ranging from 7 to 38 m (Tables 3.2 and 3.4) including a fixed radius of 11.3 m at each sample plot location to derive a suite of area-based predictor metrics. Based on the returns above 1.5 m from the ground surface, numerous candidate predictors characterizing canopy structural attributes as mentioned in Falkowski et al. (2010), Hudak et al. (2008) and McGaughey (2014) were derived. Altogether ninety metrics representative of canopy (fractional) cover, height distributional statistics, relative vegetation density (percentage returns by height strata), proxy leaf area index, gap fraction (Wulder et al., 2008; McGaughey, 2014), and texture characteristics were developed (see Appendix 1). The same metrics were calculated across the entire study area at a grid size equivalent to the optimal size of VRP for the two most suitable BAFs, and also at a grid size of 22.6 m corresponding to the FRP. Note that the ratio of average stand basal area (on per unit area basis) and the desired number of tally trees per plot gives the optimal BAF; at least 4 trees per plot was the desired number in this study.

3.3.5. Optimal plot size

The optimal plot size for point sampling was estimated with the reference of average and maximum limiting distances that depend on the DBH of sample trees and BAF of the angle gauge. The average and maximum limiting distances were calculated for each plot by respectively using the average and maximum DBH in the equation:

$$R = 8.6962 * DBH / \sqrt{BAF}$$

If average limiting distance is used as the optimum radius to extract the LiDAR data, it is likely that a tree above average DBH may get excluded from the sample. Hence, a large number of radii in the range of average to maximum limiting distance (Table 3.4) were considered to extract the LiDAR data and processed to develop the predictor metrics. The impact of varying resolution of LiDAR metrics on the response variable (gross standing volume) was evaluated to identify the optimal plot size that was eventually adopted for development of LiDAR metric grids to spatially extend the VRP model over the area of interest (AOI). The plot size leading to the highest correlation between LiDAR metrics and the VRP attribute was taken as the optimum.

Table 3.2. Point sampling with different BAF and the summary of total number of tally trees in the sample plots (total 47), minimum and maximum number of tally trees per plot, number of plots having less than 4 tallies, and average and maximum limiting distances.

BAF	Total no. of tally trees	Min. no. of tallies per plot	Max. no. of tallies per plot	No. of plots with ≤ 4 tallies	Av. DBH (cm)	Max. DBH (cm)	Av. limiting dist. (m)	Max. limiting dist. (m)
5	840	6	33	0	23.9	81.6	11.15	38.12
7	627	5	27	0	24.3	81.6	9.61	32.21
9	479	2	22	1	24.3	81.6	8.46	28.41
10	441	2	21	1	24.5	81.6	8.09	26.95
12	367	2	18	5	24.7	81.6	7.44	24.60
14	317	2	14	9	24.6	81.6	6.88	22.78
15	291	2	14	11	24.5	81.6	6.62	22.00

3.3.6. Modelling and mapping

Random Forest (RF) (Breiman, 2001) based *k* nearest neighbors (kNN) imputation (Crookston and Finley, 2008; Hudak et al., 2008; Falkowski et al., 2010) was used to establish relationships between standing volume and LiDAR metrics for spatial prediction. RF is a non-parametric modeling approach that is dependent on the summary of many classification and regression trees where each tree is built up in a special way from a bootstrap sample such that each node split depends on the best predictor out of a random subset of all predictors (Liaw and Wiener, 2002; Cutler et al., 2007; Crookston and Finley, 2008). The RF algorithm provides a noble proximity metric to identify nearest neighbors for a target point from the list of reference points, depending on the covariates of the feature space. This RF-kNN based modeling assumes that the LiDAR metrics of any location are related to the forest structural attributes and the response estimate at a target point is the weighted average value of spectrally nearest neighbors in the reference set. In contemporary studies, RF-kNN models have shown better performance compared to other methods (Hudak et al., 2008; Powell et al., 2010; Evans et al., 2011) and also avoid attending to parametric assumptions inherent in traditional regression, particularly in spatial prediction with multivariate remote sensing data (Brosofske et al., 2014).

For the fixed area plot (diameter 22.6 m) and the various plot sizes (Table 3.4) at each BAF of VRP, separate reference data frames (with all predictors and the response) were created and separate models were formulated. The RF modeling procedure was executed in the R statistical software (R Core Team, 2013) which required selection of an optimum set of predictors for each resolution of LiDAR data via application of the QR-decomposition method (Cížková and Cížek, 2012) followed by a RF model selection function optimized on the RF model improvement ratio (MIR) (Falkowski et al., 2009). The QR-decomposition method applies multivariate variable screening process to prune multi-collinear candidate predictors. The importance ranking of predictors in RF follow an iterative procedure in which observed values of one variable is permuted at a time with random numbers while other predictors are left unchanged and percentage increase in mean squared error is estimated (Liaw and Weiner, 2002; Brosofske et al, 2014). The RF algorithm is useful to narrow down the predictors to an optimum set based on standardized importance values, %variance explained, and mean squared error (MSE). The optimum set of predictors were then used to develop the RF based imputation models separately for each reference frame corresponding to a plot size by using the randomForest package (Liaw and Wiener, 2002) in the R software. The amount of variation (and MSE) explained by the models for each BAF was plotted against the size of LiDAR sample to identify the best models. The size that yielded highest variance explanation or lowest MSE was considered as the optimum size for LiDAR grid metrics preparation. Two best VRP models corresponding to two optimum BAFs and respective optimum sample sizes were identified based on the goodness-of-fit statistics. The model formulated out of the FRP data was taken as the validation model. Next, wall-to-wall LiDAR grid metrics at the resolutions of the optimal sizes for the two BAFs and the exact size of FRP (22.6 m diameter) were developed for all the predictor variables in the models. Finally, the two VRP models and the FRP model were extended spatially using the yaImpute package (Crookston and Finley, 2008) of the R software. The outputs were three rasters of standing volume corresponding to the three models.

3.3.7. Accuracy assessment

The accuracy of imputation estimates by the two best VRP models at the plot level were evaluated based on the FRP inventory measurements through calculation of bias and root mean square error (RMSE). Graphical analysis of equivalence tests following Robinson et al. (2005) and Robinson and Froese (2004) was also carried out to verify whether the imputation estimates differ significantly from the fixed radius plot measurements. The stand level volume estimates from the two best VRP models were also compared with the estimates from the FRP model. In addition, stand level summaries of standing volume were also generated based on the extrapolation of plot level inventory estimates from the FRP sampling. However, the equivalence test could not be performed at the stand level as the number of stands was less than required (only six).

3.4. Results

The minimum and maximum numbers of tally trees per plot with VRP sampling at different BAF levels are given in Table 3.2. It reveals that BAF 10 is most suitable for effective inventory of the entire area as it tallies at least four trees per plot (except in one) as suggested in several inventory guidelines (NRIS, 2014). The BAF 10 encounters fewer tally trees compared to the excessively large numbers of trees with BAF 5.

The coinciding plot-level volume estimates based on the FRP and the seven different VRP sampling schemes applied at all the sample locations revealed that VRP sampling with BAF 9 produces the closest estimates (in terms of bias) on the reference data from the FRP cruising (Table 3.3; Figure 3.2). The correlation and error statistics for the plot level FRP versus VRP estimates are given in the Table 3.3 while the individual plot level estimates by those methods are given in the Appendix 2. When the sample plots were grouped into plots with younger stand ages and older stand ages, the analysis of residuals imply that the BAF 9 produces lower bias and error in the younger stands while BAF 10 performs better with the older stands (Tables 3.3a and 3.3b). It is also evident that younger stands suffer larger relative bias in the VRP based inventory compared to the older stands. However, the estimates are negatively biased (i.e. under estimated) with smaller BAFs in high biomass areas of the older stands (Table 3.3b). The VRP based plot volume estimates also show that a higher BAF gauge results in overestimation in low biomass areas while a lower BAF gauge results in underestimation in high biomass areas (Figure 3.2).

Table 3.3. Analysis of residual errors and correlation statistics of the VRP based plot level volume estimates compared with the FRP based estimates for all the plots

Statistics	BAF 5	BAF 7	BAF 9	BAF 10	BAF 12	BAF 14	BAF 15
Bias ($m^3.ha^{-1}$)	-2.4539	3.7707	1.5466	4.5152	4.7710	5.6569	3.5307
Rel. bias (%)	-2.3412	3.3959	1.4213	4.0393	4.2584	5.0095	3.1867
RMSE ($m^3.ha^{-1}$)	35.7435	35.4053	36.1586	38.5704	41.0013	45.0732	44.7076
Rel. RMSE (%)	34.1024	31.8860	33.2301	34.5052	36.5961	39.9149	40.3509
Correlation coefficient	0.8749	0.8808	0.8778	0.8734	0.8666	0.8395	0.8315

Table 3.3a. For the low volume plots in younger stands (ID 10, 12, 17 and 19)

Statistics	BAF 5	BAF 7	BAF 9	BAF 10	BAF 12	BAF 14	BAF 15
Bias ($m^3.ha^{-1}$)	10.9971	11.0520	7.0611	7.6846	6.9532	7.8970	8.0048
Rel. bias (%)	15.9667	16.0336	10.8734	11.7209	10.72513	12.0061	12.1501
RMSE ($m^3.ha^{-1}$)	21.4711	24.8439	22.6463	26.8394	27.0924	29.3860	30.9537
Rel. RMSE (%)	31.1737	36.0421	34.8728	40.9367	41.7889	44.6763	46.9827
Correlation coefficient	0.5974	0.5743	0.4572	0.4441	0.3755	0.3052	0.3151

Table 3.3b. For the high volume plots in older stands (ID 6 and 24)

Statistics	BAF 5	BAF 7	BAF 9	BAF 10	BAF 12	BAF 14	BAF 15
Bias ($m^3.ha^{-1}$)	-26.191	-9.0786	-8.1849	-1.0777	0.9199	1.7037	-4.3645
Rel. bias (%)	-15.568	-4.8983	-4.3949	-0.5574	0.4709	0.8686	-2.2964
RMSE ($m^3.ha^{-1}$)	52.1406	48.7487	49.3879	53.3083	57.9005	63.9755	61.9289
Rel. RMSE (%)	30.9936	26.3020	26.5190	27.5718	29.6407	32.6198	32.5844
Correlation coefficient	0.6059	0.6531	0.6507	0.6579	0.6796	0.5933	0.5909

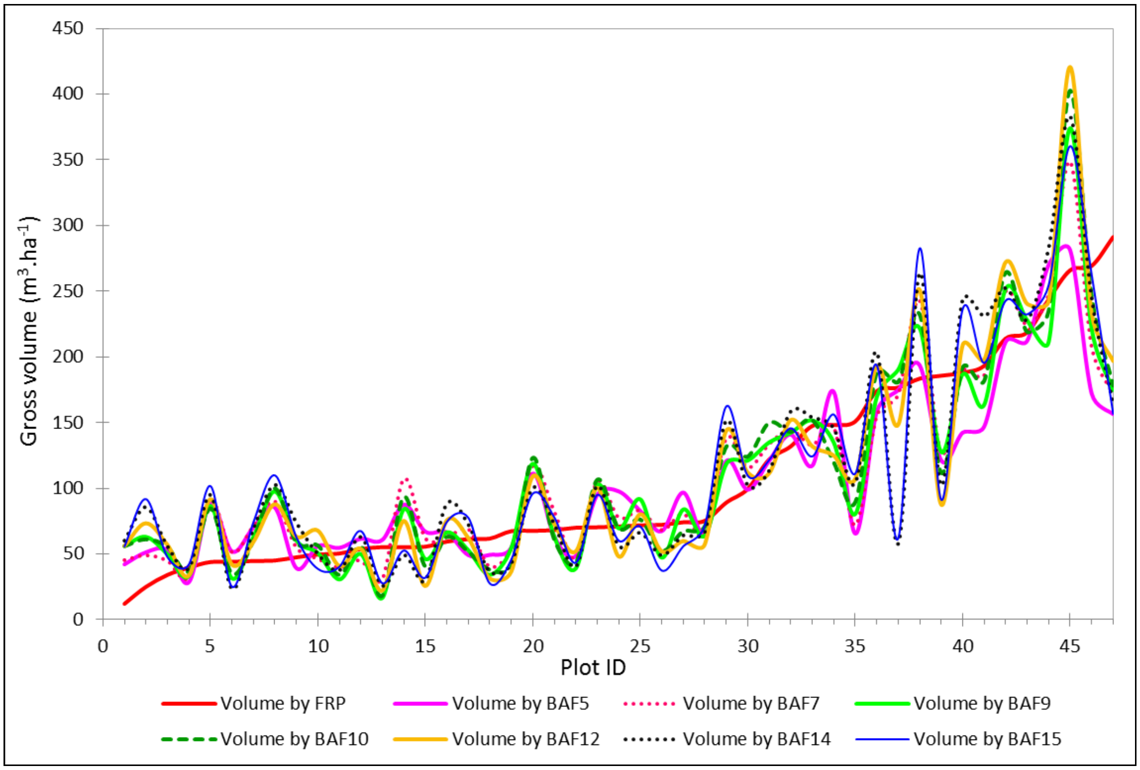


Figure 3.2. Comparison of FRP based plot level volume estimates with VRP based estimates at the seven different BAF levels.

The percentage variance explained by the formulated RF inventory models for the plot level volume prediction using numerous spatial extents (radius in m) of LiDAR samples are given in Table 3.4. The size of LiDAR samples were in the range of 7 to 38 m radius which corresponds to the average to maximum limiting distances derived respectively from the average and maximum DBH encountered for each BAF. Table 3.4 shows that the optimum LiDAR sample size (plot radius) decreases with increasing BAF. The best models from VRP data were obtained from LiDAR samples with a 9 m radius, corresponding to BAFs 9 and 10 (Tables 3.4 and 3.5; Figure 3.3). The model based on FRP data explained the highest variance (upto 83.32 %) among all other models and hence was taken as the reference model. The RMSE of the FRP, BAF9 VRP and BAF 10 VRP imputation models were 31.80, 37.97, 45.75 m^3.ha^{-1} respectively (Table 3.5).

Table 3.4. Variance explained by the RF models constructed from different sized VRP data and LiDAR metrics at different resolution (i.e. radius of LiDAR samples)

Radius of LiDAR samples (m)	% variance explained by the RF models at the different BAFs of VRP						
	BAF 5	BAF 7	BAF 9	BAF 10	BAF 12	BAF 14	BAF 15
7					59.72	60.13	60.42
8				64.5	**61.52**	**61.31**	**61.23**
9			**72.94**	**65.59**	59.91	59.28	59.33
10		59.93	69.87	61.27	58.25	57.41	58.27
11	63.36	**61.63**	68.60	62.22	58.79	56.08	57.56
12	64.56	60.48	66.01	61.16	57.92	53.36	52.91
13	61.65	59.42	65.32	61.35	58.06	55.3	57.57
14	62.22	58.82	64.36	61.80	58.88	56.45	56.57
15	**65.50**	58.74	63.90	61.06	56.70	54.28	53.63
16	64.19	57.99	63.40	59.44	53.72	57.62	55.11
17	61.53	58.70	63.19	60.85	55.73	60.29	55.16
18	60.52	57.70	62.73	60.58	56.17	60.14	54.29
20	60.62	57.01	62.66	61.54	56.42	59.38	57.03
22	61.66	56.78	60.99	62.35	57.10	60.28	59.63
23	62.43	57.73	61.92	62.55	58.98	58.17	
24	61.71	58.39	59.53	62.79	58.85		
26	60.52	58.76	60.03	63.81			
28	60.93	56.83	60.32				
29	58.72	55.67					
30	59.06	56.76					
32	57.28	57.06					
35	59.26						
38	59.50						

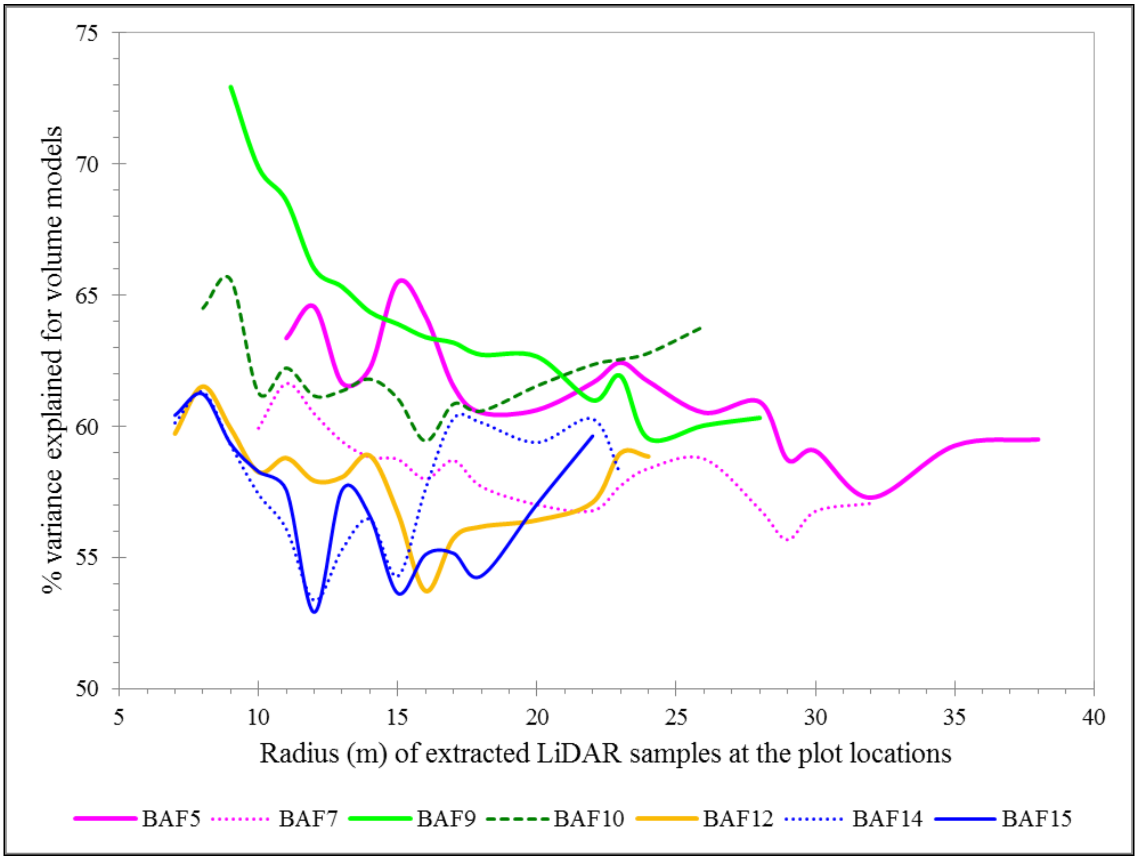


Figure 3.3. Percent variance explained by the RF prediction models built from VRP based plot volume estimates at different scales of BAFs and predictors at different resolution of LiDAR samples.

The summary of RF based relationships between field plot inventory and co-located LiDAR derived metrics are given in the Table 3.5. FRP data produced the best model when different combinations of the response from variable or fixed radius plots and the predictors derived from corresponding LiDAR samples were analyzed in the modeling exercise. The FRP model constructed using all plots, except the 4 plots from the old reserve-stand, produced the highest degree of determination (83.32%). However, inclusion of the reserve-stand plots resulted in slightly lower explanation of variance (81.12%) which can largely be attributed to the fact that many snags were present in the reserve plots but not recorded during the field inventory. Among the models based only on VRPs and corresponding LiDAR metrics, the most efficient model was with BAF 9 at the optimum radius of 9 m for the LiDAR samples. When the inventory data (response variable) from the VRP sampling and the predictors derived from the fixed area (11.33 m radius) LiDAR samples corresponding to the points were used in model building, the results were inferior as compared to when an optimum radius was used for each BAF. The combination of inventory data from FRP and VRP, in different proportion from both young and old stands, and LiDAR metrics at respective resolutions revealed that inclusion of a higher proportion of FRPs from older stands produces better training data for modeling. Similar results were obtained when VRP data from the two levels of BAF (9 and10) were mixed with FRP data in the model building process. This implies that VRP data from younger stands and FRP data from older stands, or only VRP data from all stands can be combined to formulate a generalized model with some compromise in accuracy. However, mixing different sized plots in the training dataset has practical limitations while spatially extending the model across the entire acquisition area (see discussion section). The predictor metrics selected by the RF approach are also listed for each model in Table 3.5.

Table 3.5. A summary of random forest based key models from the different combinations of field plot inventory data and LiDAR derived predictors

Description of model inputs	Selected explanatory variables	% Variance explained	RMSE (m^3ha^{-1})	LiDAR sample radius (m)
Only FRP samples (response and predictors from 11.33 m radius)				
All 47 FRP samples	See List 1	**81.12**	31.80	11.33
43 FRP samples (reserve stand excluded)	See List 2	**83.32**	30.30	11.33
30 FRP samples (young stands only)	See List 3	6.43	17.25	11.33
17 FRP samples (old stands only)	See List 4	23.13	42.46	11.33
Only VRP samples (response from VRP and predictors from an optimum radius)				
47 plots @BAF 5	See List 5	65.62	35.20	15
47 plots @BAF 7	See List 6	61.63	43.82	11
47 plots @BAF 9	See List 7	**72.94**	37.97	9
47 plots @BAF 10	See List 8	**65.59**	45.75	9
47 plots @BAF 12	See List 9	61.52	50.52	8
47 plots @BAF 14	See List10	61.31	50.95	8
47 plots @BAF 15	See List11	61.23	49.26	8
VRP samples (response from VRP and predictors from 11.33 m radius)				
47 plots @BAF 5	See List12	61.69	37.16	11.33
47 plots @BAF 7	See List13	61.77	43.74	11.33
47 plots @BAF 9	See List14	**68.18**	41.17	11.33
47 plots @BAF 10	See List15	**65.62**	45.73	11.33
47 plots @BAF 12	See List16	60.03	51.49	11.33
47 plots @BAF 14	See List17	57.31	53.52	11.33
47 plots @BAF 15	See List18	55.66	52.68	11.33

Table 3.5 (continued).

Description of model inputs	Selected explanatory variables	% Variance explain-ed	RMSE (m^{3}ha^{-1})	LiDAR sample radius (m)
Mixed FRP and BAF9 VRP samples (response and predictors from respective plots with FRP size 11.33 m and VRP optimum size 9 m)				
30 FRPs in young + 17 VRPs in old stands	See List19	74.26	37.50	11.33, 9
30 VRPs in young + 17 FRP in old stands	See List20	**78.66**	33.29	11.33, 9
Less FRPs and more VRPs (32:68) for all	See List21	70.52	40.55	11.33, 9
More FRPs and less VRPs (68:32) for all	See List22	**76.51**	34.62	11.33, 9
50% FRPs & 50% VRPs per stand	See List23	72.65	37.70	11.33, 9
Mixed FRP and BAF10 VRP samples (response and predictors from respective plots with FRP size 11.33 m and VRP optimum size 9 m)				
30 FRPs in young + 17 VRPs in old stands	See List24	69.61	43.58	11.33, 9
30 VRPs in young + 17 FRP in old stands	See List25	**76.81**	34.67	11.33, 9
Less FRPs and more VRPs (32:68) for all	See List26	68.06	44.86	11.33, 9
More FRPs and less VRPs (68:32) for all	See List27	**76.54**	34.76	11.33, 9
50% FRPs & 50% VRPs per stand	See List28	69.58	40.45	11.33, 9
Mixed FRP and BAF9 and BAF10 VRP samples (response and predictors from respective plots; VRPs being measured half-half with the two BAFs)				
30 FRPs in young + 17 VRPs in old stands	See List29	72.27	41.04	11.33, 9
30 VRPs in young + 17 FRP in old stands	See List30	**77.52**	34.39	11.33, 9
Less FRPs and more VRPs (32:68) for all	See List31	68.93	42.57	11.33, 9
More FRPs and less VRPs (68:32) for all	See List32	**77.14**	34.36	11.33, 9
50% FRPs & 50% VRPs per stand	See List33	71.50	38.99	11.33, 9

List 1: PropT, ElevMax, ElevMean, ElevMode, ElevSD, ElevIQR, Strata3, Strata5
List 2: ElevMax, ElevMean, ElevMode, ElevSD, ElevVar, Strata3, Strata5
List 3: PropT, Prop4, ElevL3, IntMode
List 4: PropT, ElevMax, ElevMean
List 5: PropT, ElevMax, ElevMean, ElevMode, ElevSD, ElevVar, ElevIQR, IntMode,
……...Strata3, Strata5
List 6: PropT, ElevMax, ElevMean, ElevMode, ElevSD, ElevIQR, EMADmed,
……...ElevL3, IntL4, Strata3, Strata5
List 7: PropT, ElevMean, ElevMode, Strata5
List 8: PropT, ElevMax, ElevMean, ElevMode, ElevSD, ElevVar, ElevSkew, ElevL3,
……...Strata3, Strata5
List 9: PropT, ElevMax, ElevMean, ElevMode, ElevSD, ElevVar, ElevL4, IntL4,
……...Strata3, Strata5
List 10: ElevMax, ElevMean, ElevMode, ElevVar, ElevL4

List 11: PropT, ElevMax, ElevMean, ElevMode, ElevSD, ElevVar, ElevSkew,
………ElevKurt, ElevL4, ElevLkurt, IntL4, Strata2, Strata3, Strata5
List 12: ElevMax, ElevMean, Strata5
List 13: ElevMax, ElevMean, Strata5
List 14: PropT, ElevMax, ElevMean, ElevSD, Strata5
List 15: ElevMax, Strata5
List 16: PropT, ElevMax, ElevMean, ElevMode, ElevSD, ElevIQR, EMADmed,
………ElevL3, ElevL4, IntL4, Strata3, Strata5
List 17: Prop4, ElevMax, ElevMean, ElevMode, ElevSD, ElevIQR, ElevKurt,
………ElevL3, ElevL4, IntSkew, IntL4, Strata3, Strata5
List 18: ElevMax, ElevMean, ElevMode, ElevSD, ElevIQR, ElevKurt, ElevL3,
………ElevL4, IntSkew, IntL4, Strata3, Strata5
List 19: PropT, ElevMax, ElevMean, ElevMode, ElevSD, Strata5
List 20: PropT, ElevMax, ElevMean, ElevSD, ElevVar, Strata5
List 21: PropT, Prop4, Prop5, ElevMax, ElevMean, ElevMode, ElevSD, ElevVar,
………ElevIQR, ElevSkew, ElevKurt, EMADmed, ElevL3, ElevL4, ElevP05,
………IntMin, IntMode, IntVar, IntL3, IntL4, Strata2, Strata3, Strata4, Strata5
List 22: PropT, ElevMax, ElevMean, ElevMode ElevSD, ElevVar, ElevL3, IntL4,
……....Strata3, Strata5
List 23: PropT, ElevMax, ElevMean, ElevMode, ElevSD, ElevVar, Strata5
List 24: PropT, ElevMax, ElevMean, ElevMode, ElevSD, ElevVar, ElevL3, IntL4,
………Strata3, Strata5
List 25: PropT, ElevMax, ElevMean, ElevSD, ElevVar, ElevIQR, Strata3, Strata5
List 26: ElevMax, ElevMean, ElevSD, ElevVar, Strata5
List 27: PropT, ElevMax, ElevMean, ElevMode, ElevSD, ElevVar, ElevL3, IntL4,
………Strata3, Strata5
List 28: PropT, ElevMax, ElevMean, ElevMode, ElevSD, ElevVar, EMADmed,
………ElevL3, Strata0, Strata3, Strata5
List 29: PropT, ElevMax, ElevMean, ElevMode, ElevSD, ElevVar, ElevL3, IntL4,
………Strata0, Strata3, Strata5
List 30: PropT, ElevMax, ElevMean, ElevSD, ElevVar, ElevIQR, Strata3, Strata5
List 31: PropT, Prop3, Prop4, ElevMax, ElevMean, ElevMode, ElevSD, ElevVar,
………ElevIQR, ElevSkew, ElevL3, ElevL4, ElevP05, IntMin, IntVar, IntL4 ,
………Strata2, Strata3, Strata4, Strata5
List 32: PropT, ElevMax, ElevMean, ElevMode, ElevSD, ElevVar, ElevL3, IntL4,
………Strata3, Strata5
List 33: PropT, ElevMax, ElevMean, ElevMode, ElevSD, ElevVar, ElevL3, Strata0,
………Strata3, Strata5

These results show that VRP sampling with BAF 9 and 10, both matching an optimal plot size of 9 m radius, produced the top two models after the FRP based model (if we disregard models based on mixed VRP and FRP data). The FRP model is taken as the reference to mainly evaluate the plot-level prediction accuracy of the VRP based models. For these three models, the selected LiDAR metrics and their importance ranking based on the percentage increase in mean square error when a particular variable is dropped from a model are given in Figure 3.4. The intensity metrics were not found to be significantly important in any of the formulated models.

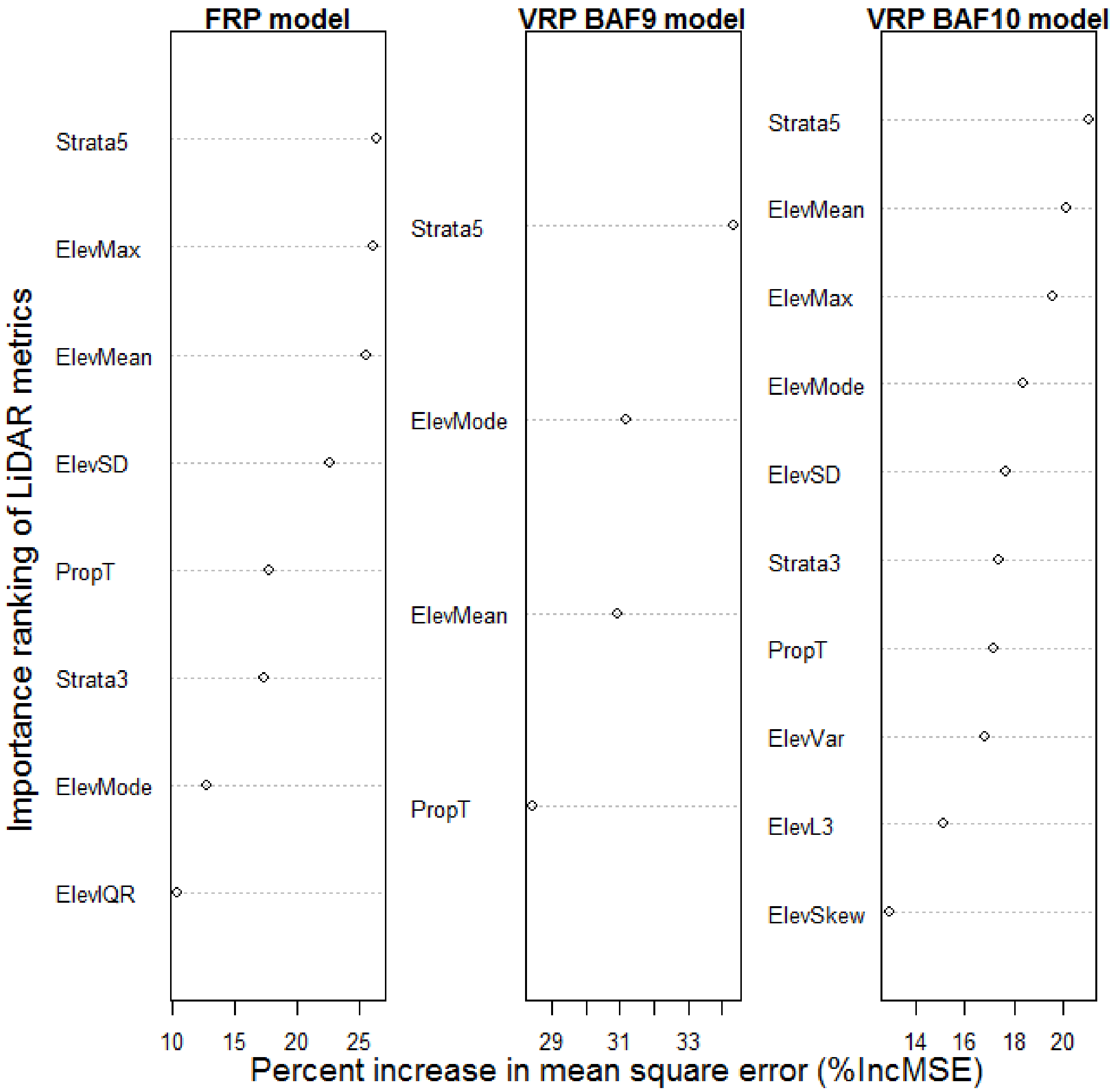


Figure 3.4. Importance ranking of LiDAR metrics selected in the fixed radius model and the variable radius models.

Comparisons of the plot-level volume estimates by the BAF 9 and BAF 10 imputation models against the FRP measurements are given in the equivalence plots in Figure 3.5. The equivalence test uses the null hypothesis of dissimilarity of two target datasets being compared (Robinson and Froese, 2004; Robinson et al., 2005). The test assumes that if two-one-sided confidence interval (at a given alpha level) of slope and intercept of the line of best fit lie within a specified region of similarity for the slope and intercept then the slope of observed-predicted regression is similar to 1 and the two datasets hold equivalency. Since the lines of best fit in Figure 3.5 lie within the region of similarity (set at 25% for both slope and intercept) and the confidence intervals for slopes and intercepts lie within the respective regions of similarity, it is concluded that the VRP model based imputation estimates are equivalent to the FRP based measurements.

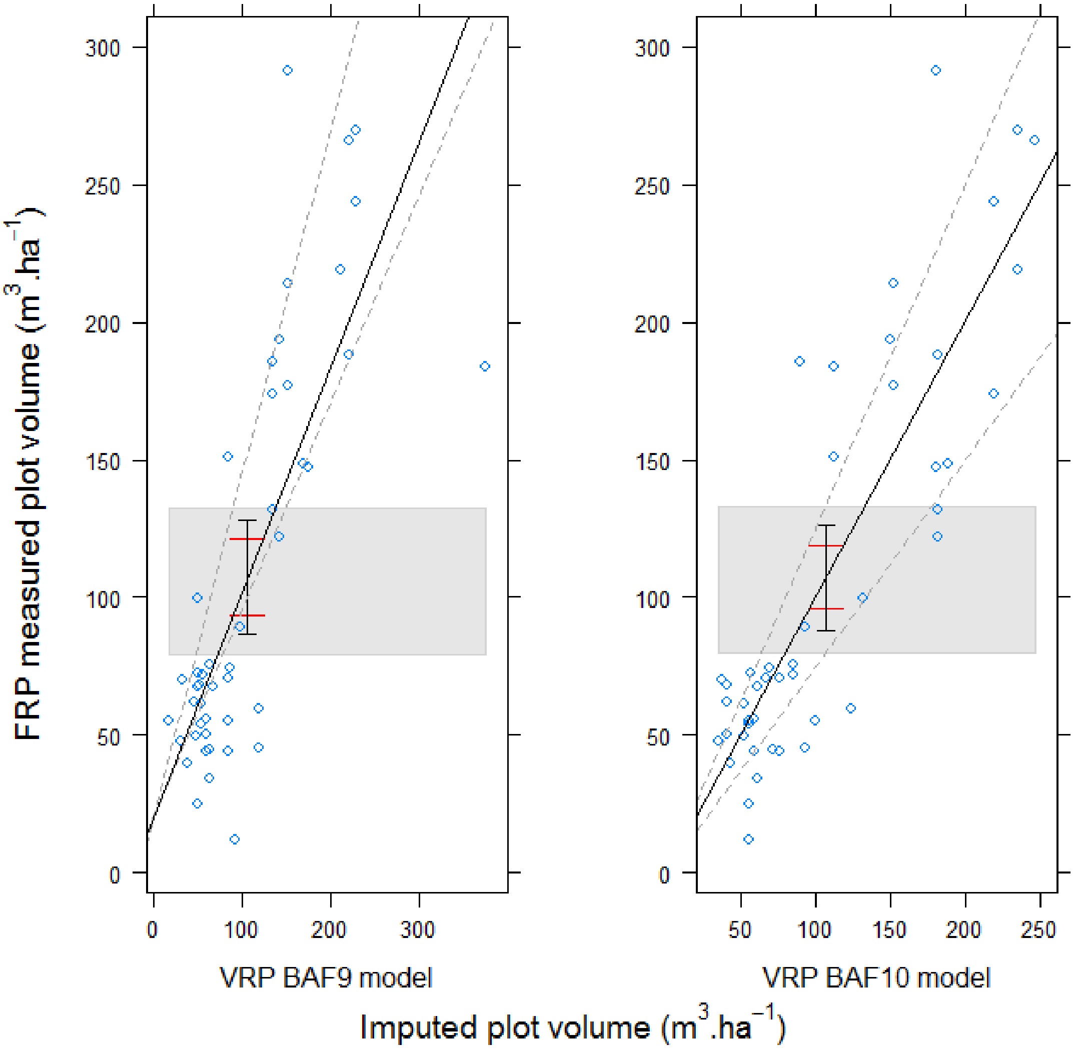


Figure 3.5. Equivalence plot of the measured and imputed plot level volumes by the two VRP models. The black inclined line represents the line of best fit, the dashed gray lines represent the 25% region of similarity for the slope, the shaded gray polygon represents the 25% region of similarity for the intercept, black vertical bar represents a confidence interval (at 5% alpha level) for the slope of the line of best fit, and red vertical bar indicates the confidence interval for the intercept.

The spatially explicit standing volume raster outputs were at 18 m resolution for BAF 9 and BAF 10 VRP models, and at 22.6 m resolution for the FRP model (Figure 3.6). The stand level gross volume estimates by the three models and directly based on extrapolation of the field inventory are given in Table 3.6. It can be argued that the volume estimates by the FRP model is close to the actual volumes compared to the extrapolation based or the VRP model based estimates.

Table 3.6. Comparison of stand level gross volume (m^3) estimates by the different methods.

Stand ID	Area (ha)	Design-based estimate from FRP inventory (% std. error)	Total volume by FRP imputation model	Total volume by VRP BAF9 imputation model	Total volume by VRP BAF10 imputation model
6	50.29	10301.66 (6.89)	9714.17	8418.49	9257.23
10	10.15	474.14 (7.18)	472.34	424.25	432.71
12	14.23	577.73 (17.34)	711.36	764.40	886.40
17	33.01	2154.64 (6.55)	2203.85	2505.80	2631.99
19	32.12	2191.80 (6.92)	2092.09	1989.73	2200.37
24	4.08	655.55 (9.40)	766.61	435.79	492.47

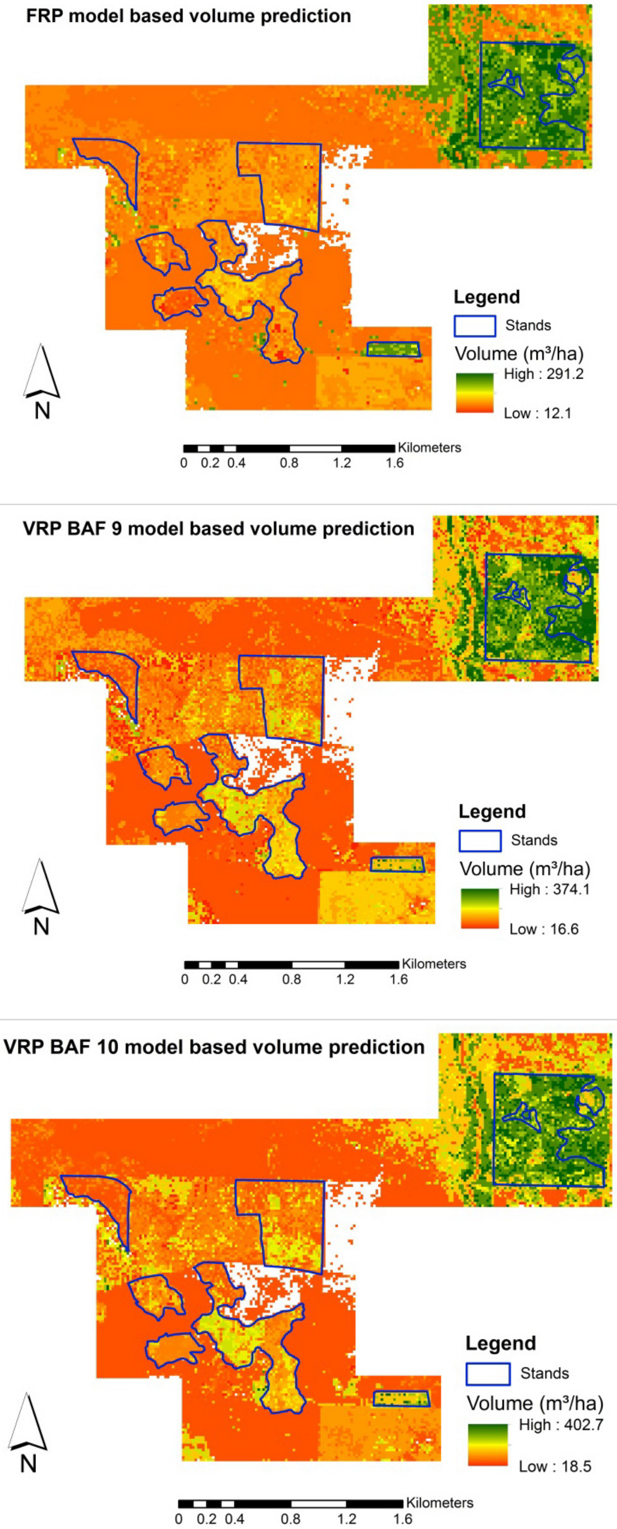


Figure 3.6. Volume prediction maps by the FRP (top) and BAF 9 (middle) and BAF 10 (bottom) based VRP models.

3.5. Discussion

A prime concern of forest managers is to obtain inventory estimates at a standard accuracy and minimum cost. A variety of inventory designs that include variable radius plots and fixed area plots are used to estimate forest structural parameters. The choice of a design depends on forest type, target variables, and required inventory accuracy. Often, managers want to increase sampling intensity for intensive management. As the plot size increases, the sampling effort and cost increases but the variation among plots diminishes. An optimum plot size depends on spatial pattern and variation in sampling units. In case of stands with clumps and gaps, a large plot size is useful to minimize variance among plots. Thus the sampling choice involves a trade-off between accuracy and cost-efficiency.

The sampling design impacts the accuracy, particularly bias of estimates. The probability-based selection of sampling units avoids bias and provides correct estimates of sampling error. A sampling method is efficient when it estimates the target variable with a probability proportional to the quantity (Avery and Burkhart, 1994). In an inventory using randomly-located fixed area plots, each tree is selected with equal probability, whereas in variable radius plots each tree has a selection probability proportional to DBH. VRP sampling which samples larger trees with greater probability, is more precise and cost-effective for stand timber volume estimation in most circumstances compared to fixed plots that require significantly more cruising time (Scott, 1990). Fixed plots are efficient in the examination of younger stands with smaller diameter trees where point samples suffer due to issues associated with missing trees. Point samples perform better in older stands with larger trees and a wide range of diameters. Fixed plot sampling requires more sample trees than point samples to yield the same precision (Matern, 1972). Martin (1983) reported that if the target variable is independent of stand basal area (e.g., stem density, cover type) then fixed plots are more accurate.

Selection of an optimum BAF that yields a desired number of trees per plot is important to improve inventory accuracy. A smaller BAF tallies more trees but miss (or double count) some as the distance between the point and the tree increases. A larger BAF on the other hand tally fewer trees per plot but cause increased coefficient of variation (CV) among plots. Hence, search for an optimum BAF is required to minimize the number of missing trees and the CV. In eastern U.S. a BAF of 5, 10, and 20 (sighting angles 73.66', 104.18', 147.34' respectively) are commonly applied to yield the desired number of “in” trees. The literature provides a wide range for the desired numbers. For example, Avery and Burkhart (1994) reported 5-12 trees, Schreuder et al. (1993) reported 6-12 trees, and NRIS (2014) recommended an average of 4-8 trees per plot. Avery and Burkhart (2002) noted that BAF 10 is commonly used for second growth saw timber and dense pole size stands in the eastern United States. NRCS (2011) reports that BAF 10 is the most common in Michigan. Avery and Newton (1965) found that BAF 10 VRPs are roughly equivalent to 0.04 ha and 0.02 ha FRPs in terms of tree tally in stands with average dbh of 34.2 cm and 24.1 cm respectively; average dbh of the stands in this study ranged from 14.4 cm to 24.1 cm. The Michigan Integrated Forest Monitoring Assessment and Prescription (IFMAP) project also applied BAF 10 in the stage 2 inventory. If a site estimate of stand basal area is available, an appropriate BAF can be obtained from the ratio of stand basal area and the desired number of tally trees per plot. Once a BAF is selected and applied on the first plot, conventionally it is used throughout the stand exam.

The angle gauges with BAF in multiple of 5 are commonly available in the market and associated expansion factors can be readily obtained from available inventory guidelines. The BAF 5 prism in this study encountered too many trees per plot (up to 33) and required much effort to avoid issues of missing, double counting and occluding trees in the dense stands. The BAF 10 is the most suitable gauge for inventory in the Ford Forest area since the gauge adequately sampled the number of trees per plot which ranged from 4-21. Avery and Newton (1965) found BAF 10 plots and 0.04 ha plots perform equally well in hardwood stands in Georgia. The individual fixed plots in existing set-up at the Ford Forest have 0.04 ha area and hence an initial assumption was that BAF 10 device would work better in the study area. This assumption is validated from the observations that the models based on LiDAR samples of 0.04 ha and inventory data from the BAF 9 and BAF 10 plots yielded high variance explanation (Table 3.5). Since BAF 9 devices are not easily available and requires additional work to apply in an inventory, we focused more on BAF 10.

The analysis of the design-based estimates from the VRP inventory compared with the FRP inventory revealed that BAF 9 performed best (i.e. least bias and RMSE) followed by BAF 10 for all the stands combined (Table 3.3). The LiDAR based models similarly revealed that BAF 9 was the most accurate (in terms of % variance explained and MSE) followed by BAF 10 (Table 3.5). This result signifies the strength of LiDAR remote sensing in forest structural characterization.

The model developed from younger stands only explained a small portion of the variance. This is not surprising because younger stands have small volumes with many trees below 10 cm DBH which were not counted in the field inventory, but included in LiDAR samples. On the other hand, the model based on older stands only explained variance marginally better, despite the fact that there were fewer plots and the stands included many snags and smaller live trees below 10 cm DBH. Fixed plots in the older stands against variable plots improved model accuracy by a small amount, which implies that

VRP can work well in older stands that have large volumes. In contrast, using fixed plots in the younger stands didn't improve accuracy since the young stands have low standing volume.

Application of data from mixed plot sizes (fixed and variable plots or only variable plots at different levels of BAF) from separate locations in model building also seems to work well in terms of the goodness-of-fit statistics. However, use of multiple BAFs in LiDAR based multi-stand inventory is work intensive since multiple LiDAR grids at different resolutions corresponding to different BAF need to be prepared in order to spatially extend the model spatially. The results also reveal that the VRP data in some cases can be substituted for FRP as long as VRP contain a desired number of trees per plot and the optimum radius identified for the VRP is comparable to the size of FRP. In this study, the optimum radius identified was 9 m for both BAF 9 and the BAF 10. The search of optimal radius for LiDAR samples for integration with VRP data at the different levels of BAF showed that the optimum radius decreases with increasing BAF. This observation is consistent with the fact that increasing BAF means increasing sighting angle and closer location of tally trees (i.e. a tree of given size to be "in" must lie closer to the plot center with increasing BAF).

Volume is a three dimensional metric, so information on LiDAR returns from various canopy height strata and horizontal coverage is essential for estimation of wood content or biomass (Wulder et al., 2008). Selection of LiDAR metrics is important for effective inventory modeling of structural parameters as many remote sensing derived predictors are linearly dependent. The canopy height distribution, percent cover and vertical strata density metrics were the most influential predictors in the selected models by the RF technique. The intensity metrics were not found to be significant in any of the models which parallels the note by McGaughey (2014) that "in aerial discrete return lidar technology the intensity values are not normalized, so they are not ideally suited for analytical work".

The errors in the LiDAR based models can be attributed to several factors. A prominent factor is the mismatch between LiDAR and field data (Gleason and Im, 2011). The extracted LiDAR samples corresponding to the field plots may represent canopy elements that are part of trees outside the plot. A tree just outside the plot or a leaning tree can contribute a large amount of returns in the LiDAR samples. Another reason for mismatch was the ignorance of dead trees in the field inventory and inclusion in the LIDAR samples. There may also be flaw in the field data as the allometric equations used for the volume calculation was regional based which is adopted by the FIA program in the Lake States. Nonetheless, the models are generalizations of reality and provide a valuable means to improve understanding of the complex interactions between interdependent ecosystem components.

3.6. Conclusions

i. VRP sampling of younger stands involve larger relative bias compared to older stands.

ii. BAF 9 VRP performed best (in terms of bias) compared to the reference field inventory obtained from FRP sampling at the target stands combined; however, BAF 10 VRP was the best for older stands.

iii. BAF 10 is most suitable for effective inventory in the Ford Forest area as it tallies at least four trees per plot, overcomes practical difficulties associated with the use of several BAFs, and the device with associated expansion factors is easily available.

iv. As expected, integration of FRP data with LiDAR data provided the best model. Nonetheless, VRP data can also be integrated with LiDAR data for inventory prediction with some compromise in accuracy but equivalent estimates as with the FRP based model.

v. VRP data can also be combined with FRP data (or substituted for FRP data) for spatial inventory with LiDAR derived metrics at an optimum grid resolution. Fixed plots in the older stands against variable plots improved model accuracy by a small amount, which implies that VRP can work well in older stands that have large volumes. A combination of VRP data from younger stands and FRP data from older stands, or only VRP data from all stands can be used to formulate a generalized model with some compromise in accuracy.

vi. Canopy height distribution, strata density, and cover density are the most influential LiDAR predictors but intensity is not.

3.7. Suggested further study

Similar study can also be extended to include broadleaved stands and explore suitability of a common BAF for the entire Ford Forest.

We have found optimum radii for seven different BAFs to guide integration of VRP data with LiDAR data for spatial inventory. It is a question whether the radius can also be applied at other sites with differing forest structure and composition.

Further research can be directed at identifying an optimum BAF for any target area based on LiDAR data. Such studies will help grouping of stands into cohorts of similar structure, composition, and quality, for silvicultural treatments.

Instead of measuring DBH of individual tally trees in each plot, the DBH of only the largest tree per plot can be measured to obtain the maximum limiting distance to vary the LiDAR sample size corresponding to the VRP. The average DBH or individual tree DBH for each plot can be derived from LiDAR data. That will potentially be most appropriate to map stand basal area.

4. Performance evaluation of imputed site index and biogeoclimatic spatial data in diameter growth modeling of selected tree species in the Great Lakes States[3]

4.1. Introduction

Forest managers require tree growth, yield and productivity models to project stand development and summarize growing stock status for decision making (Miner et al., 1988; Lessard et al., 2001). Tree growth models are usually based on tree size and periodic expansion of size (lateral or vertical) driven by a combination of edaphic, climatic, and biotic elements influencing ecological site and competition of trees (Wykoff, 1990; Valentine, 1997). Formulation of reliable growth models with careful selection of predictors is necessary for operational applications such as estimation of wood production for sustainable harvesting (Vanclay, 1994). The way site quality is represented in growth models has been shown to have important influence on prediction accuracy as site relates to productivity (Pokharel and Froese, 2009). In principle, site and site productivity are often distinguished in the sense that site refers to the combination of local physical, biological, and climatic factors, while site productivity refers to the synoptic effect of the biogeoclimatic characteristics on the quantitative production of plant biomass (Skovsgaard and Vanclay, 2008).

The Forest Vegetation Simulator (FVS) is a widely used growth and yield modeling framework that generates stand statistics for current and future management scenarios (Dixon, 2002). FVS based projections are dependent on empirical aspatial individual tree growth, mortality, and volume equations (Lacerte et al., 2004). FVS has been adopted nationally by the US Forest Service and is implemented via twenty variants throughout the U.S.A., across all forest types. All the FVS variants are continuously modified and updated by improving the embedded models (Dixon, 2002). For example, the Lake States variant was reformulated in 2006, except for the diameter growth models. The most important component in the FVS based projection and simulation of stand development constitutes the large tree diameter increment models that are empirically derived from observed periodic growth of sample trees from geo-referenced locations (Wykoff et al., 1982; Froese and Robinson, 2007). The diameter growth models can be formulated with either diameter increment (ΔD) or basal area increment as dependent variable since the variables are algebraically related and variance of the response distribution can be made

[3] This chapter is ready to submit in a forestry journal with Ram K. Deo as the first author and Robert E. Froese and Michael J. Falkowski as the second and third authors respectively.

homogeneous through transformations. Wykoff et al. (1982; 1990) used natural log transformed ten years difference in diameter squared (lnDDS) as the response variable to emphasize basal area increment. Cole and Stage (1972), West (1980), and Zhao et al. (2004) noticed better performance of lnDDS model over lnΔD. Vanclay (1994) argues that basal area increment and diameter increment are algebraically related, and therefore either can be used as response variable. The DDS models have biological rationale and better correlate with tree volume increment. Any empirical model for diameter growth should express higher increments at smaller size with a peak at intermediate size and then diminish slowly as the photosynthetic capacity of larger trees go down, and finally approach zero asymptotically (Leary, 1997).

The commonly measured attributes from routine inventory programs are generally selected as explanatory variables in any growth modeling. Diameter at breast height (DBH) is the prime predictor in most growth models. In addition, stand basal area per acre (SBA), quadratic mean diameter (QMD), crown ratio (CR), trees per hectare (TPH), site index (SI), and derivatives from DBH measurements such as percentile ranking of tree size are predominantly used in the growth models. The factors such as SBA, TPH, and cumulative basal area of all trees larger than a subject tree (BAL) signify competition that affects tree growth. The BAL is included in the prognosis growth model (Wykoff et al., 1982) and the Central States TWIGS growth model (Shifley, 1987; Miner et al., 1988) to represent relative advantage of the individual trees for site resources (Stage, 1973; Monserud and Sterba, 1996). SBA reflects two-sided competition of a tree for moisture and nutrients. The competition effect of stem density (i.e., TPH) on growth can be understood from the general observation that tree growth in open area is slower than in stands of similar site quality (Burkhart et al., 1987). In general, the competition factors reduce theoretical potential growth (Holdaway, 1984), and microsite or genetic differences produce random (stochastic) effect on tree growth. Tree CR is included in growth models to reflect a tree's vigor, photosynthetic potential, and effect of past competition (Wykoff, 1990).

An accurate metric of site productivity is important for forest growth modeling (Carmean et al., 1989). Site index (SI), defined as the height of dominant and co-dominant trees in competition free environment at a given base age (e.g., 50 years in the Lake States), is a proxy for forest productivity (Rehfeldt et al., 2006; Skovsgaard and Vanclay, 2008; Crookston et al., 2010; Weiskittel et al., 2011; Skovsgaard and Vanclay, 2013). This index is derived from the dominant tree height at a selected reference age which may be the age of culmination of mean annual increment, or a common rotation age. Tree height is commonly used as a proxy of site quality because of the recognized association of the height growth with volume growth, and the indifference of dominant height to thinning (stem density). A classical concept (called Eichhorn rule) is that volume production for a given species at a given stand height is identical for all sites (Skovsgaard and Vanclay, 2013).

Accurately estimating SI depends on accurate estimates of total height and age of sample trees that are free of past competition and damage. Thus, the method is most suited to fully stocked even-aged stands of known or measurable age. Although tree height for SI calculation can be measured with higher accuracy, tree age estimation is often difficult or impossible to obtain, especially for diffuse-porous hardwood tree species that grow slowly. Further, the total height estimates may also be erroneous when tree tops are broken. In regions like the Lake States, many stands are characteristically composed of shade tolerant species in uneven-aged conditions, and it is not surprising that substantial error exists in SI estimation. Since finding sample trees of dominant quality in competition free niche is difficult at many sites, development of spatially explicit map of SI may be useful for many applications.

SI depends on interaction of several biogeoclimatic variables and shift in management regimes such as fertilization and genetic improvement (Stage et al., 2001; Sharma et al., 2012). Spatial variability in topography, soil, climate, and complex biotic interactions leads to variations in site conditions. The moisture gradient across a landscape, soil depth, soil nutrient, and temperature characteristic can influence site productivity since physiological systems of vascular plants are rooted to these factors. The spatial and temporal variation in forest site productivity can be modeled dependent on measures of climate, soil moisture, soil nutrients, land cover type, canopy density, canopy height, topographic variables, and other satellite imagery derived digital metrics (Klinka and Carter, 1990; Monserud et al., 2006; Monserud et al., 2008; Waring et al., 2010; Beaulieu et al., 2011; Weiskittel et al., 2011; Sharma et al., 2012; Skovsgaard and Vanclay, 2013). Since site productivity depends on climate and climate is changing, integration of climatic spatial data is essential to make SI prediction models sensitive to climate.

An important data source for tree growth modeling and derivation of spatially explicit SI maps is the Forest Inventory and Analysis (FIA) Program of the US Forest Service. FIA has been conducting periodic national forest inventories on a state-by-state basis since 1960s. FIA has adopted a coherent national plot design under the standard annual inventory system since 1999 in which the whole country is divided into regular hexagons (2,428 ha) and at least one permanent plot (0.067 ha) is established per hexagon. The total plots in a state are divided into 5 to 10 panels and a single panel is measured each year and thus plots are re-visited in 5 to 10 years. FIA computes species-specific SI for every tree (SITREE) in the sample plots for a reference age (usually 25 or 50 years), based on measurements of one or more dominant and co-dominant site-trees per plot during the inventory (Woudenberg et al., 2010). The SI is most commonly calculated using a family of curves dependent on total height and age of sample trees (Carmean et al., 1989). A given species can have entirely different SI curves in different geographic regions, and each set of curves may use a different reference age.

The changing paradigm in forestry to holistic management justifies the search for alternatives to traditional SI (Pokharel and Froese, 2009). A number of spatial layers of biogeoclimatic variables are freely available through public web-portals. These spatial predictors can be coupled with the FIA database in order to formulate SI models, and of course the predicted SI can be incorporated into growth models to analyze the potential for broader application. The likelihood that FIA plots are evenly distributed over all age and site classes make the database more appropriate for regional scale SI modeling; an uneven proportion of sample trees from younger stands (high growth rate, so better sites) or older stands (slow growth, so poor sites) can introduce bias (Avery and Burkhart, 1994). The FIA data can be co-registered with biogeoclimatic spatial layers and utilized for geospatial inventory of SI under a modeling framework such as Random Forest based *k* nearest neighbor (RF-kNN) imputation (Falkowski et al., 2010). The RF-kNN imputation has been profoundly applied in contemporary forestry research for regional and local scale species distribution and structural attributes mapping (Rehfeldt et al., 2006; Weiskittel et al., 2011). The imputed SI may be applicable for large scale resource inventory and growth assessment especially in uneven-aged mixed species stands where site trees are difficult to identify and measure for total height and age. The spatial mapping has the advantage that users can estimate site quality even for the areas that are presently devoid of forests but need afforestation. Imputed SI can also be helpful for forest managers in selecting suitable crop, and planning operations.

Direct application of biogeoclimatic variables instead of measured SI in a growth model is also an option to account the influence of site quality on tree growth. This approach can be justified if we could replace measured SI with readily available spatial biogeoclimatic predictors, or imputed SI. As measured SI may involve error accumulated from tree age and height measurements, my hypothesis is that direct inclusion of bioclimatic variables in the growth model will perform better in large tree growth modeling.

4.2. Objectives

This study was aimed at evaluating alternative ways of including site factors in the formulation and application of large-tree diameter growth models in the Lake States. Three sets of proxies characterizing site productivity namely measured SI, imputed SI and a combination of biogeoclimatic variables were used separately in calibrating species-specific growth models and their performance was evaluated in terms of growth prediction. The study examined the strength of imputed SI and environmental variables in growth modeling and projection (i.e., whether or not the derived SI can capture site differences in the increment model). In addition, the implications and tradeoffs of applying a growth model based on measured SI in areas where only imputed SI exists and vice versa were explored.

Research question

- How well does imputed SI or the combination of biogeoclimatic variables compare when substituted for measured SI in the large-tree diameter growth models?

4.3. Methods

4.3.1. Reference data

The FIA inventory database (FIA, 2013) based on the annual inventory design was utilized for the preparation of species-specific SI maps and formulation of large tree diameter growth models for the five major species of the Lake States (LS). The selected species, belonging to tolerant and intolerant categories of both conifer and broadleaf, included red pine (*Pinus resinosa*), northern white cedar (*Thuja occidentalis*), sugar maple (*Acer saccharum*), quaking aspen (*Populus tremuloides*), and northern red oak (*Quercus rubra*). These species were found be the dominant in terms of the number of trees in the database and spatial coverage of site classes (niche). These species also have high commercial importance in this region. Among the two selected conifers, red pine is shade intolerant that mostly inhabit plain or gently rolling sandy ground or low ridges adjacent to swamps, while northern white-cedar is shade tolerant that prefers cool, moist, nutrient-rich sites, particularly organic soils near streams. Among the three broad leaves, sugar maple is very tolerant to shade, quaking aspen is very intolerant, and northern red oak is intermediate in shade tolerance. Sugar maple grows best on well-drained loams (stunted growth on dry shallow soils or swamps) but quaking aspen grows on a great variety of soils ranging from shallow rocky to deep loamy sands or heavy clays. The northern red oak grows on cool moist glacial soils, preferably deep well-drained loams.

The inventory data of FIA plots were obtained from the online database by using the FIA Data Mart Tool (FIA, 2013). The database was downloaded separately for Michigan (MI), Wisconsin (WI) and Minnesota (MN) which are the states covered by the LS variant of FVS. FIA started annual inventory system in the states in 1999 (for MN) and 2000 (for MI and WI). The selected inventory years with 10 years gap between two points in time for re-measurement of sample trees are given in the Table 4.1. All the three states have five inventory panels and five-year cycle to revisit the plots. The inventory data from two panels of WI and three panels each of MI and MN were used in the species-specific diameter growth modeling while the remaining panels were used for the species-specific SI mapping (Table 4.1). Separate panels were used for SI mapping and diameter growth modeling to avoid circularity bias (using measured SI of a plot to impute SI for the same plot and using those in growth modeling).

Table 4.1. Measurement years and cycles of sample trees from FIA plots used for growth modeling and site index imputation

State	Time 1 inventory year	Time 2 inventory year	Time 1 FIA cycle-subcycle	Time 2 FIA cycle-subcycle	Sample data application
MI	2000	2010	6-1	8-1	diameter growth modeling
	2001	2011	6-2	8-2	diameter growth modeling
	2002	2012	6-3	8-3	diameter growth modeling
	2003		6-4		site index modeling
	2004		6-5		site index modeling
WI	2000	2010	6-1	8-1	diameter growth modeling
	2001	2011	6-2	8-2	diameter growth modeling
	2002		6-3		site index modeling
	2003		6-4		site index modeling
	2004		6-5		site index modeling
MN	1999	2009	12-1	14-1	diameter growth modeling
	2000	2010	12-2	14-2	diameter growth modeling
	2001	2011	12-3	14-3	diameter growth modeling
	2002		12-4		site index modeling
	2003		12-5		site index modeling

The FIA database encompasses detailed information on biophysical parameters under separate tables including plot and tree tables. The tree level information includes tree-ID, species code, status code (live or dead), DBH, total height, crown ratio, calculated SI (SITREE), biomass, and many other variables. In addition, location reference of each tree is signified in terms of the hosting subplot, plot, county, and state. However, only fuzzed and swapped coordinates of the plots are available to general public (due to privacy regulations set forth by the federal government and plot integrity concerns of FIA). The fuzzing and swapping creates random shift in plot coordinates and exchange of plot attributes. For large area growing stock estimation, Coulston et al. (2006) found that spatial models based on perturbed plot locations did not differ significantly from the models based on unperturbed plot locations.

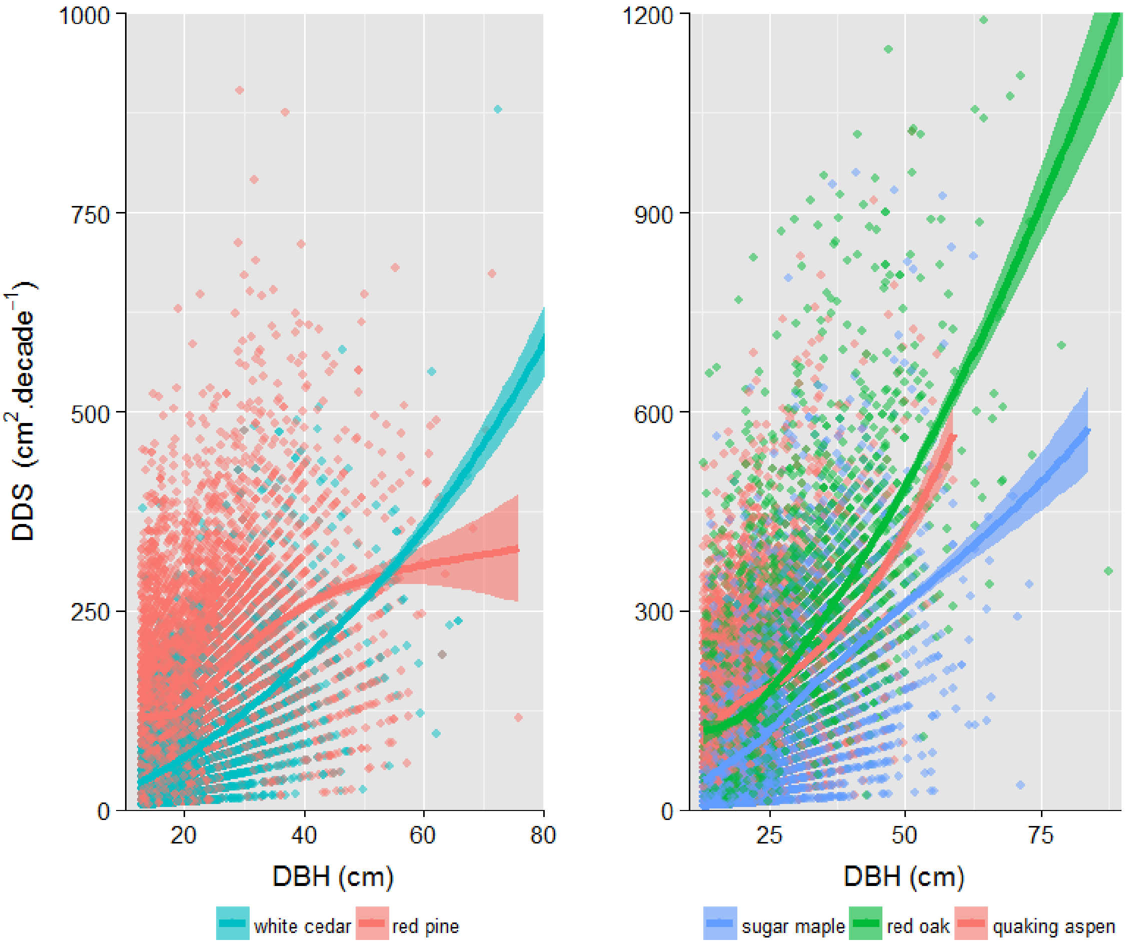
Live sample trees (STATUSCD = 1), measured for over-bark diameter at breast height (DIAHTCD = 1) in the last three consecutive cycles, (Table 4.1) were selected for analysis in this study. Only sound trees of merchantable standards (TREECLCD = 2, that excludes rough or rotten cull trees), above 12.7 cm DBH, and re-measured accurately for DBH (DIACHECK = 0) in the inventories were selected for diameter growth modeling. The sample trees selected for SI modeling were additionally subjected to the criteria of being actually measured in the field for total height (HTCD=1). The sample tree measurements at time 1 and time 2 were related by matching the county, plot, subplot, tree, and species codes of the database using Microsoft Access query language. The actual diameter increments in 10 years (ΔD) as well as 10 years difference in squared diameters (DDS) were calculated for each sample tree. Any sample trees with negative, zero, or above 15 cm diameter increment in 10 years were discarded from the reference frame for model fitting (only 3 trees were found to have above 15 cm diameter increment). In ArcGIS display, some plots that were found to fall outside of the spatial extent of the LS boundary, due to the fuzzed and swapped coordinates, were also excluded from the reference set (25 plots with 475 trees). Based on the above criteria of sample selection, the number of trees for SI and growth modeling and their distribution over plots are given by species in the Table 4.2 and Table 4.3. Since the trees were selected via stringent criteria from a large number of FIA plots, the samples were representative of virtually all stand age, structures, composition, and site across the entire LS. The diameter distribution of sample trees used in growth modeling and the scatter plots of DDS against DBH are given in the Figure 4.1. The reverse J-shaped diameter distribution of sample trees signifies the representation of heterogeneous stand structure and diverse site qualities.

Table 4.2. Number of sample trees for growth modeling, size distribution and growth characteristics over the range of FIA sites/plots

Species	No. of trees	No. of plots	DBH range (cm)	Mean DBH (cm)	10-yr DBH growth range (cm)	SI range (m)	Correlation of DBH and DDS
Red pine	7,923	710	12.7-75.9	22.9	0.2-12.7	4.8-32.3	0.3842
N. white-cedar	9,905	754	12.7-83.1	20.3	0.2-10.4	3.9-30.4	0.6124
Sugar maple	10,575	1,403	12.7-83.8	22.1	0.2-12.9	6.4-32.0	0.585
Quaking aspen	9,269	1,766	12.7-58.9	20.2	0.2-13.2	6.7-34.7	0.4376
N. red oak	3,104	883	12.7-99.5	28.3	0.2-15.2	7.6-36.3	0.6325

Table 4.3. Number of sample trees for SI modeling, size distribution and site index ranges

Species	No. of trees	No. of plots	DBH range (cm)	Mean DBH (cm)	SI range (m)	Mean SI (m)
Red pine	6,384	949	12.7- 71.1	23.5	6.4- 32.3	20.6
N. white-cedar	7,991	1,134	12.7- 87.8	20.9	4.3- 31.4	11.4
Sugar maple	13,089	2,263	12.7- 93.2	22.4	9.1- 37.2	19.2
Quaking aspen	12,283	2,808	12.7- 81.3	21.1	6.4- 36.6	21.4
N. red oak	4,810	1,455	12.7- 96.5	28.8	7.9- 35.0	20.2



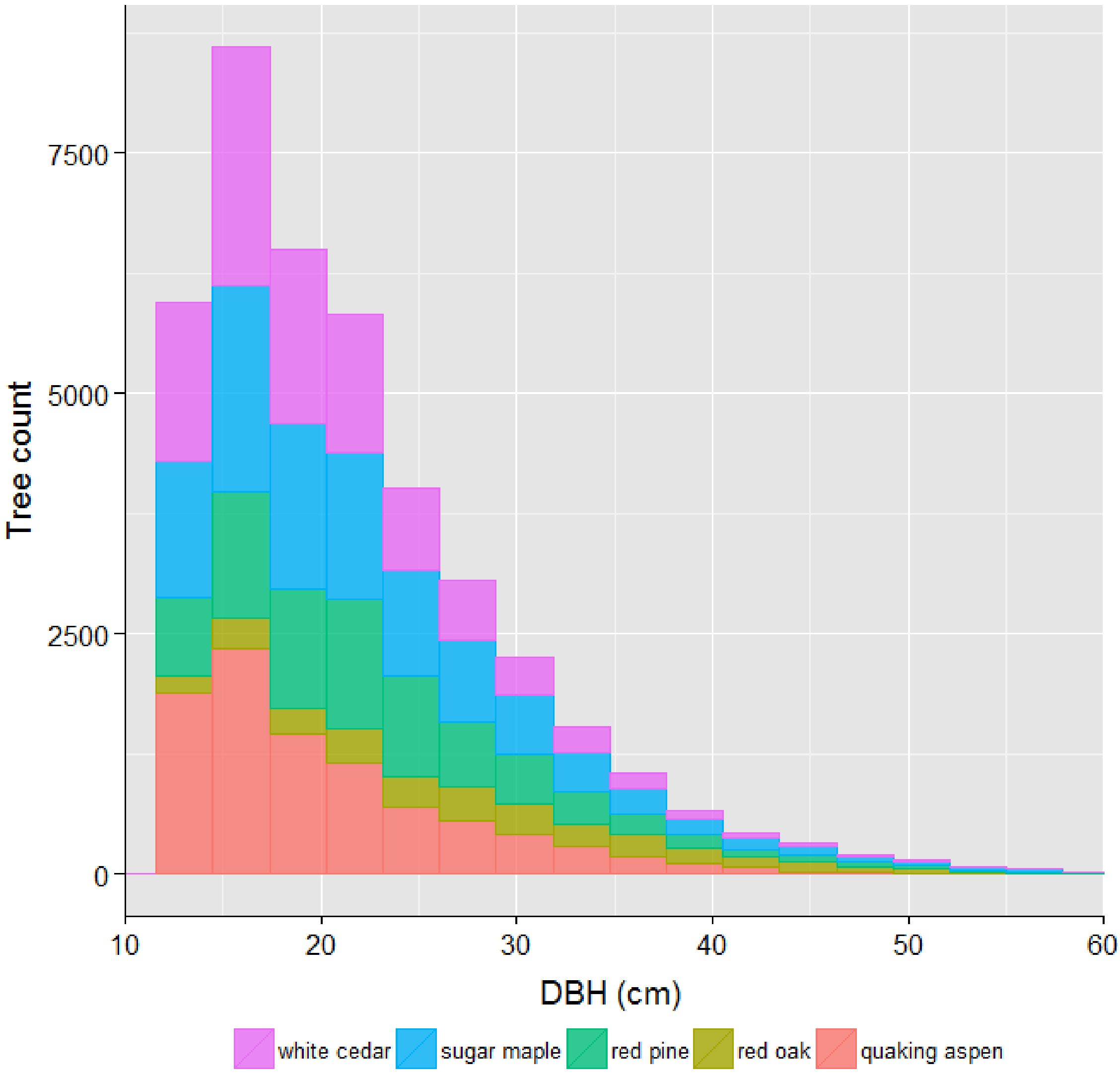


Figure 4.1. Locally weighted regression (*loess*) curve with 95% confidence band fitted to the scatter plot of DDS against DBH (top), separately for conifer and broadleaf, along with the relative frequency of sample trees by diameter classes (bottom) for the growth model fitting.

The DBH (≥ 12.7 cm) data of all live trees in each plot, measured at the start of the 10 year growth period (i.e. time 1 data), were summarized to obtain plot characteristics namely stand basal area per acre (SBA), basal area of larger trees (BAL) than the subject tree, trees per hectare (TPH), and quadratic mean diameter (QMD). These variables were eventually used as predictors in the growth modeling. The smaller ingrowths were ignored with an assumption that they have relatively little influence on growth of the over-story trees. Additional tree variables considered for direct inclusion in growth modeling were CR and SI, both measured at the start of the growth period. I also tested the potential of imputed SI and biogeoclimatic variables to represent site quality since they could aid in decoupling the model from the need for accurate SI measurements. The fuzzed and swapped coordinates of the FIA plots were used to intersect and attach the auxiliary variables to each tree. My assumption is that biogeoclimatic variables do not change significantly within the fuzzed-swapped zone (about 0.8 km) of the plots.

The biogeoclimatic variables considered in the study can be grouped into three categories: climate, soil, and satellite. The contemporary climate data consisting of raster grids for frost-free degree-days above 5^0C (DD5), growing season precipitation (GSP), mean annual precipitation (MAP), mean annual temperature (MAT), and mean temperature in the warmest month (MTWM) were obtained from the Moscow Forest Sciences Laboratory (RMRS, 2013). These climatic rasters were derived at the source by fitting Hutchinson's spline surfaces to 30 year (1961-1990) normalized average monthly data from meteorological stations throughout the North America (Rehfeldt et al., 2006; Crookston et al., 2010). Accuracy of these layers was tested by Weiskittel et al. (2011) through comparison with DAYMET (http://daymet.org/) derived daily temperature and precipitation datasets. The USDA system of soil taxonomy based two spatial layers, namely soil drainage index (DI) and productivity index (PI), were additionally identified for inclusion in the growth modeling because these layers are primarily developed to indicate long term soil wetness, soil volume available for plant rooting, soil productivity ranks, and likely tree stress areas (Schaetzl et al., 2009; Schaetzl et al., 2012). The DI and PI layers were downloaded from the forest health protection, mapping and reporting portal (USDA Forest Service, 2013a). The layers were originally developed from the most detailed digital soil survey geographic database, SSURGO (NRCS, 2013), by spatially joining its soil map units (MUKEY) field with an empirically produced master table for soil drainage and productivity indexes (USDA Forest Service, 2013a). Both DI and PI are on ordinal scale: DI ranges from 0 to 99 with the higher values representing more water, and PI ranges from 0 to 19 with higher values representing more productive sites. The satellite dependent predictor layers included MODIS/Terra sensor derived 16-day composite NDVI image (see https://lpdaac.usgs.gov) in the peak growing season (July) of 2010, and the landcover dataset from the National Gap Analysis Program (GAP, 2013). The GAP landcover is produced from multi-season Landsat (ETM+) imageries from 1999-2001. The GAP landcover map is available at six tiered levels of vegetation details based on physiognomy (FGDC, 2008) but we considered only the macrogroup level of national vegetation hierarchies for ease in interpretation and analysis. The macrogroup classes (total 59) were further aggregated into broader categories (total 20) by merging similar vegetation cover types. All the biogeoclimatic layers were clipped to the LS boundary in ArcMap and the grids were resampled to a common spatial resolution of 250 m with exactly overlapping orientation of pixels of all the layers. The GAP landcover data was not directly employed in diameter growth modeling (but in SI imputation) to avoid complex model structure with the addition of categorical variables with many classes.

4.3.2. Validation data

A validation dataset with permanent locations of the sample plots established under the Continuous Forest Inventory (CFI) design, were obtained from the Bureau of Indian Affairs (BIA), Midwest Regional Office. There were altogether 739 permanent plots (radius 16 m) within seven reservations (404, 409.1, 409.2, 409.3,432, 434, and 438) with true coordinates of plot centers and measured tree and plot level variables. The data were available from 1965 to 2006, however, only the last two inventory cycles (Table 4.4) were considered for the validation, in the similar fashion used in the growth modeling. The Microsoft Access query command was utilized to join database tables of different inventory years by matching the reservation, plot, tree, and species codes. The numbers of sample trees by target species are given in the Table 4.4. Since the sample tree re-measurement intervals varied from 13 to 17 years, the diameter increment data was normalized to 10 years for consistency in the analysis. The dataset contained species-specific calculated SI only for some reservations (404, 432, and 434) but site tree measurements were available in other reservations (409.1, 409.2, 409.3, and 438). The coefficients and equations from Carmean et al. (1989) were used to calculate species-specific SI, denoted by SI_{bia}, for each plot based on age and height data of site trees. The analysis revealed that BIA has applied Carmean et al. (1989) equations for SI calculation. All the biogeoclimatic variables were attached to the sample trees using actual coordinates of the BIA plots in ArcGIS.

Table 4.4. BIA validation dataset description

Reservations	No. of plots	Time 1 inventory year	Time 2 inventory year	No. of sample trees by target species, and DBH and SI ranges
404	143	1993	2006	Red pine: 1179 trees; 77 plots; 12.7 to 62 cm dbh; 14 to 31 m SI W. cedar: 1767 trees; 47 plots; 12.7 to 58 cm dbh; 6 to 17 m SI S. maple: 481 trees; 31 plots; 12.7 to 53 cm dbh; 16 to 23 m SI Q. aspen: 780 trees; 85 plots 12.7 to 46 cm dbh; 15 to 34 m SI Red oak: 299 trees; 39 plots; 12.7 to 55 cm dbh; 15 to 26 m SI
409.1	65	1992	2008	
409.2	141	1992	2007	
409.3	55	1992	2009	
432	154	1991	2005	
434	65	1990	2005	
438	116	1989	2003	

4.3.3. Site index modeling

The FIA measured SI, denoted by SI_{fia}, was the response variable, and ten biogeoclimatic variables were attached as potential explanatory variables in SI modeling. The panels and numbers of sample trees used for SI modeling are given in the Tables 4.1 and 4.3 respectively. Separate SI models for each of the five species was developed using the RF-kNN imputation technique in which the proximity of target and reference locations in the feature space of covariates are first calculated using the RF algorithm (Crookston and Finley, 2008) and then the response variable from the nearest reference plot is assigned to the target location. The RF-kNN modeling and spatially explicit mapping was implemented in ArcMap using the Marine Geospatial Ecology Tools (Roberts et al., 2010; MGET, 2013). The mean square error, proportion of variance explained, and variable importance metrics produced by the algorithm were analyzed to select optimal values for the required parameters in the modeling. The number of trees (*ntree*) in RF and number of variables (*mtry*) for node split of each tree are the key parameters to implement the modeling. The model predictions depend on summary (average or majority voting) of many trees and their internal structure (each tree is built from a bootstrap sample; and tree growth depends of node splitting by the best out of a random subset of predictors). It was found that 1500 or more trees and 2 variables at each node (*mtry*) produced stable error. The SI imputation accuracy was corroborated with the BIA data for SI.

4.3.4. Diameter growth modeling

Natural log transformed ten years difference in diameter squared (lnDDS) of sample trees was the response variable in the species-specific large tree diameter growth models. The logarithmic transformation was necessary as the histogram of DDS was found to be positively skewed which is against the principles of linear least square regression. The panels and numbers of sample trees used for the growth modeling are given in the Tables 4.1 and 4.2, respectively. The sample trees were also intersected with the spatially explicit layers of imputed SI for respective species. That means each sample tree was attached to measured SI, imputed SI, and ten different biogeoclimatic values. Keeping the tree size and competition parameters as the common predictors, diameter growth models were formulated by separately using the three different proxies of site productivity: measured SI, imputed SI, and a direct combination of the biogeoclimatic variables. The size effect was accounted in the model by including DBH (D), CR, and transformed DBH namely 1/D, and D^2. The competition indices as the ratio of DBH to QMD and interaction of DBH and relative diameter (i.e. D^2/QMD) were considered as predictors; the rationale is that a tree of given size has less competition in younger stand than in an older stand with similar stem density (Wykoff, 1986). The interaction terms such as SI×QMD (referred as anabolic terms by Hahn and Leary, 1979) and BAL× lnDBH were also evaluated with the rationale that the former can be a proxy for crown (respiring) surface and the later can describe the impact of diameter distribution on tree size.

In an independent research comparing several composite linear model forms as formulated in Andreassen and Tomter (2003), Cole and Stage (1972), Wykoff (1990), Froese (2003), Weiskittel et al. (2007), and Zhao et.al. (2004), it was observed that the form as in Equation-1 provided best fit statistics (adjusted R^2 and standard error) with the same reference data.

$$lnDDS = \beta_1 + \beta_2 \cdot \frac{1}{D} + \beta_3 \cdot D + \beta_4 \cdot D^2 + \beta_5 \cdot \frac{D}{QMD} + \beta_6 \cdot \frac{D^2}{QMD} + \beta_7 \cdot SBA + \beta_8 \cdot BAL + \beta_9 \cdot CR + \beta_{10} \cdot CR^2 + \beta_{11} \cdot SI_{fia} \quad \text{...................} Equation\ 1$$

$$lnDDS = \beta_1 + \beta_2 \cdot \frac{1}{D} + \beta_3 \cdot D + \beta_4 \cdot D^2 + \beta_5 \cdot \frac{D}{QMD} + \beta_6 \cdot \frac{D^2}{QMD} + \beta_7 \cdot SBA + \beta_8 \cdot BAL + \beta_9 \cdot CR + \beta_{10} \cdot CR^2 + \beta_{11} \cdot SI_{impt} \quad \text{.................} Equation\ 2$$

$$\begin{aligned} lnDDS = {} & \beta_1 + \beta_2 \cdot \frac{1}{D} + \beta_3 \cdot D + \beta_4 \cdot D^2 + \beta_5 \cdot \frac{D}{QMD} + \beta_6 \cdot \frac{D^2}{QMD} + \beta_7 .SBA + \beta_8 \cdot BAL + \beta_9 \cdot CR + \beta_{10} \cdot CR^2 \\ & + \beta_{11} \cdot DD5 + \beta_{12} \cdot MAP.DI + \beta_{13} \cdot MAP + \beta_{14} \cdot DI + \beta_{15} \cdot GSP + \beta_{16} \cdot PI + \beta_{17} \cdot MNDVI \\ & + \beta_{18} \cdot MTWM + \beta_{19} \cdot MAT + \beta_{20} \cdot BAWHT \quad \text{.................} Equation\ \ 3 \end{aligned}$$

Where *DDS*= ten years difference in over-bark diameter squared (cm^2); *D* = diameter at breast height (cm); *QMD*= quadratic mean diameter (cm); *CR* = crown ratio; *SI* = site index (m); *BAL*= basal area of larger tree than the subject tree ($m^2\ ha^{-1}$); *SBA* = stand basal area ($m^2\ ha^{-1}$); *MAP* = mean annual precipitation (mm); *DI* = soil drainage index; *PI*= soil productivity index; *GSP* = growing season precipitation (mm); *DD5* = degree-days above 5^0C accumulating within frost-free period; *MNDVI*= MODIS sensor derived normalized difference vegetation index; *MTWM*= mean temperature in warmest month (^{0}C); *MAT*= mean annual temperature (^{0}C); *BAWHT*= basal area weighted canopy height (m); *MAP.DI* = interaction of mean annual precipitation and soil drainage index; β_i are species dependent regression coefficients.

The *Equation-1* was modified to two additional versions: one with imputed SI (SI_{impt}) replacing measured SI (SI_{fia}) (*Equation 2*) and the other with a combination of biogeoclimatic variables completely substituting the site factor (*Equation 3*). Thus three model forms were examined for each of the species. The linear model function '*lm*' in the R statistical software (R Core Team, 2013) was used to formulate the growth models with the best subset of available predictors. A stepwise variable selection technique followed by the best-subsets regression was applied for identifying the most influential variables and also to reduce multicollinearity among predictors. The Akaike Information Criterion (AIC) for stopping the both direction stepwise process was implemented with the '*stepAIC*' function in the 'MASS' package of R (Venables and Ripley, 2002). The stepwise algorithm iteratively adds or removes individual predictors into the model to attain an optimum subset of candidate predictors beyond which the function no longer invoke improvement in the model with additional variables (Sakamoto et al., 1986). This implies that each parameter is assigned an AIC value and the model with the minimum AIC total is taken as the best among candidate models. The AIC is defined as *AIC* = - *2.log(L) + k*edf*, where *L* is the maximum likelihood of the candidate model, *edf* is equivalent degree of freedom, and *k* is a numeric weight for *edf* (Adler, 2010). Since a model with large number of parameters better fits the data, possibly with smaller residual, the best choice require a balance between goodness of fit and model size. The second term in the AIC formula favors model parsimony and penalizes for addition of more parameters that might lead to overfitting. AIC offers a relative estimate of information loss when a selected model is used to predict the data obtained from true process (model). A model with the lowest AIC are supposed to perform best when used for prediction outside the dataset.

The models screened from the stepwise regression were further refined for the best-subset of parameters using '*regsubsets*' function in R with the '*leaps*' package (Lumley, 2009) so that the initial adjusted R^2 and AIC values remained at the similar level and the refined models provided biologically justified growth trend with increasing DBH. The median values of the predictors (except DBH) extracted from the reference data frame were used in the refined models to portray species-specific diameter growth surfaces against DBH and only the combination of predictors that yielded satisfactory unimodal curves were selected as the final best-subset model. The best-subsets function requires an argument (namely '*nvmax*') specifying maximum size of subsets to examine and returns separate best models of all sizes up to '*nvmax*' via exhaustive search of all possible combinations of predictors in contrast to the stepwise function that identify only one combination (Hudak et al., 2006). I set '*nvmax*' equal to the number of predictors retained after stepwise operation of full models. I considered adjusted R^2 statistic for the best model selection because R^2 overestimates the strength of association between response and predictors (i.e. R^2 would never decline even when irrelevant *x*-variables are added at the cost of loss of degree of freedom). The adjusted R^2 statistic accounts for the number of x variables by penalizing excessive use of unimportant variables.

Adjusted $R^2 = 1-[(1-R^2)\ (n-1)/\ (n-k-1)]$, where n is sample size and *k* is number of independent variables in the model.

The importance ranking of predictor variables of each model was done through a sensitivity analysis by using the Sampling and Sensitivity Analysis Tool (Hoare et al., 2008). The analysis requires sampling of the input parameter space and the samples are used in an external model to predict the target response variable. The variations in the predicted response caused by the variations in the predictors are scrutinized to rank the importance of each input in terms of their contribution to uncertainty in the prediction. The Latin hypercube sampling (LHS) approach was used to characterize the input parameter space. LHS is a standard sampling technique in which a probability density function is assigned to each predictor factor and the distribution is divided into N equal probability areas, so that only one value is randomly selected from every interval of each predictor (Hoare et al., 2008). The two-parameter (scale and shape) Weibull probability density function (because of its flexibility) was fitted to the histograms of each factor. A variety of metrics are available for conducting sensitivity analysis but we used the most fascinating approach of factors prioritization by reduction of variance.

4.3.5. Model validation

The diagnostic measures such as root mean square error (RMSE), bias, and R^2 values and also the signs of coefficients were analyzed for model evaluation. The best model form (out of the three versions) for each of the five species was identified by evaluating the prediction error obtained when applying the models to the independent inventory data set from BIA. The fundamental notion that better measures of site quality should reduce mean square error on the training and validation data was favored to evaluate model performance. The values of explanatory variables from the BIA data frame were used in the calibrated models to predict diameter increment which was then compared with the measured increments. The five scenarios as in the Table 4.5 were tested to evaluate the relative performance of SI_{fia}, $SI_{impt,}$ and biogeoclimatic variables. The equivalence test as suggested by Robinson and Froese (2004) and Robinson et al. (2005) was performed to determine the performance of proxy of site productivity variables in diameter growth prediction. In addition, Tukey's honest significant differences (TukeyHSD) *post hoc* test was performed to compare the mean estimates of diameter by each model with the field measurements.

Table 4.5. Possible cases of diameter growth model calibration and verification of predicted diameter growth with measured values at BIA plot locations

No.	Proxy of site quality in model calibration	Proxy of site quality in model verification with BIA data
1.	measured site index (SI_{fia})	measured site index (SI_{bia})
2.	imputed site index (SI_{impt})	imputed site index (SI_{impt})
3.	combination of biogeoclimatic variables	combination of biogeoclimatic variables
4.	measured site index (SI_{fia})	imputed site index (SI_{impt})
5.	imputed site index (SI_{impt})	measured site index (SI_{bia})

4.4. Results

The site index imputation models reasonably explained the variance in the training dataset; however, the importance ranking (sensitivity) of auxiliary predictors varied with the species. The productivity index (PI) was found to be the most important while the mean temperature in the warmest month least affected the prediction of species-specific SI. The growing season precipitation (GSP), MODIS derived NDVI (MNDVI), mean annual precipitation (MAP), landcover, and normal degree-days above 5 ^{0}C (DD5) were the bands of second most important predictors whose ranking varied with species. The other predictors have intermediate roles in determining the site quality of any location as in the Figure 4.2. The RMSE and prediction powers of the individual RF-kNN imputation models are given in the Table 4.6.

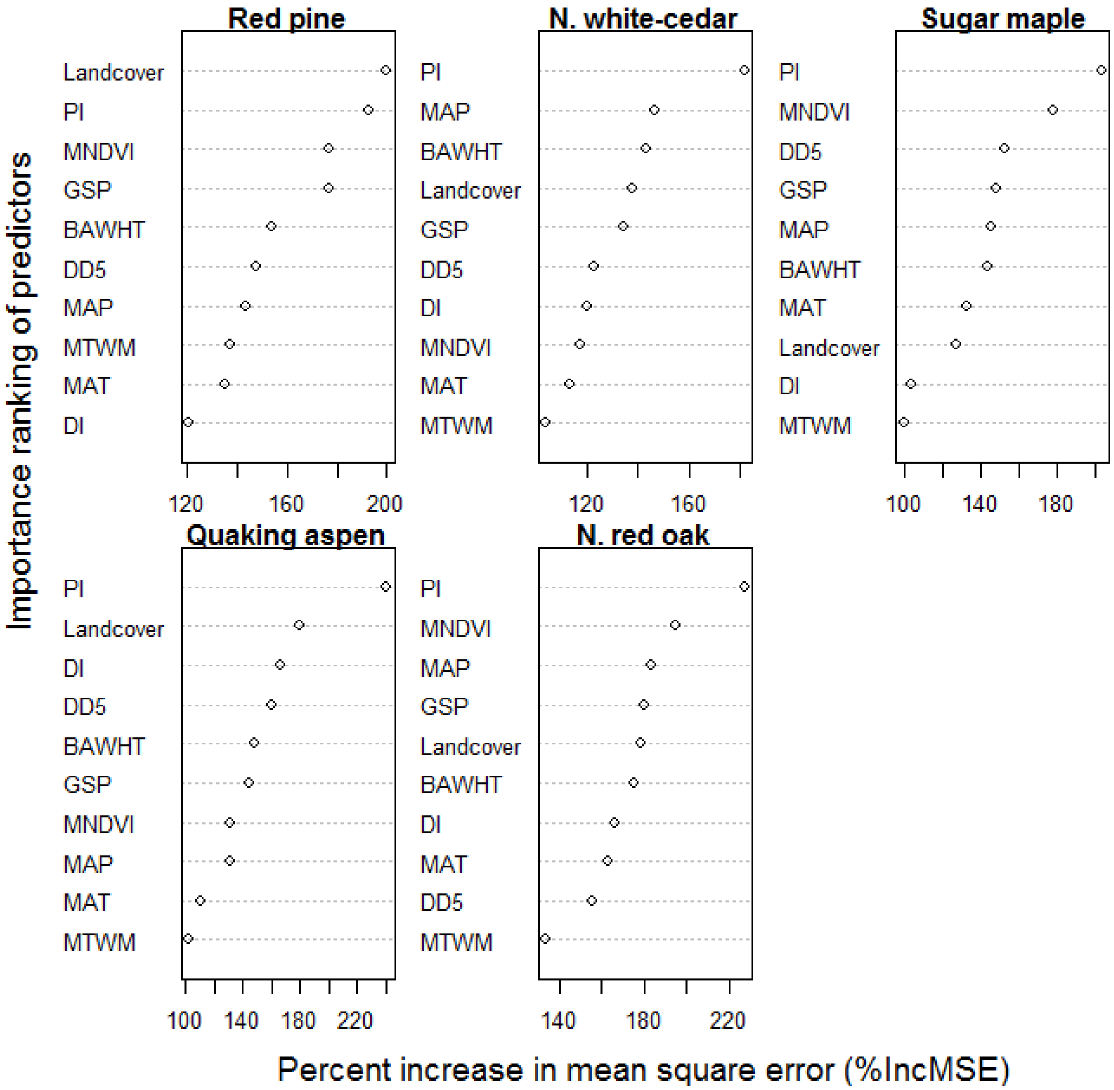


Figure 4.2. The ranking of relative importance of variables in site index prediction mapping using the random forest technique. The ranking is determined based on the percentage increase in mean square error if a particular variable in not included in the model.

Table 4.6. Fit statistics of random forest based site index imputation model

			RF model parameters	
Species	% var explained	RMSE (m)	mtry	ntree
Red pine	93.9	0.89	2	3200
N. white-cedar	91.5	1.22	2	2500
Sugar maple	92.8	0.81	2	1500
Quaking aspen	90.1	1.19	2	1700
N. red oak	86.4	1.37	2	4000

The comparisons of species-specific imputed site index with calculated (measured) site index at FIA plot locations show reasonable trend as can be expected with imputation techniques (Figure 4.3).

The Pearson's correlation (*r*) analysis of the individual predictors (including transformations and interaction terms) with the DDS indicates that *D,* D^2*, D/QMD,* D^2*/QMD, CR, QMD*, and *SI* are positively correlated to the response while *SBA, BAL*, and *TPH* (trees per hectare) are negatively correlated (Table 4.7). This result indicates positive effect of site quality and negative effect of competition elements on tree growth. CR has positive influence on growth despite the fact that tree density, stand age, management regimes, and growth habits of species affect CR (over-stocked stands supports low CR and vice versa). Further, the scatter plot of DDS against D plus D^2 in general revealed a linear trend (graph not shown) which implies a quadratic relationship of the growth with DBH (or linear relationship with initial tree basal area).

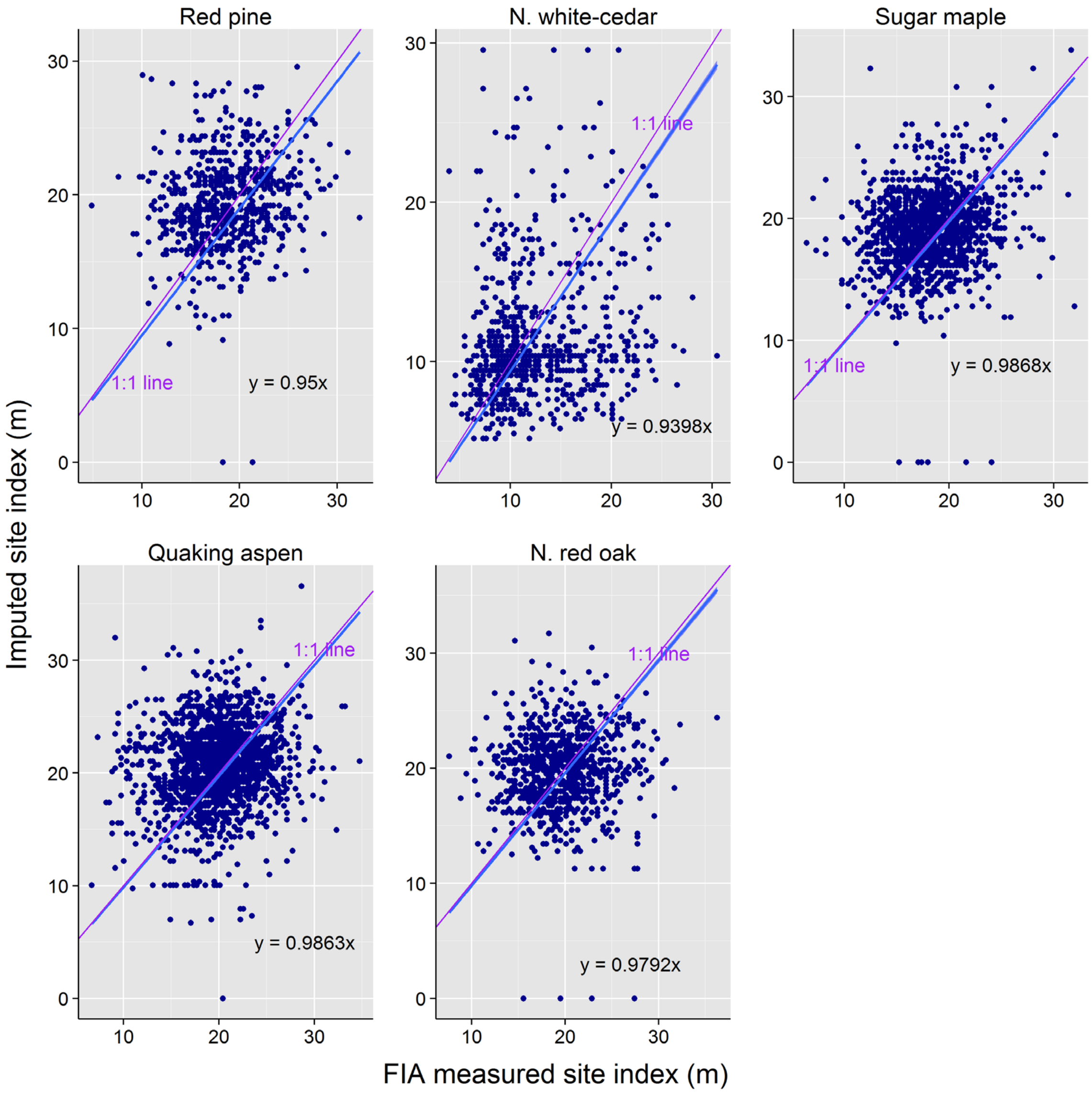


Figure 4.3. Comparison of species-specific calculated (measured) site index in the FIA plots with the imputed values at respective locations.

Table 4.7. Correlation coefficient (*r*) of predictor variables with DDS by species

Predictors	Red pine	N. white-cedar	Sugar maple	Quaking aspen	N. red oak
1/D	-0.351	-0.547	-0.536	-0.4123	-0.521
D	0.384	0.612	0.585	0.4377	0.633
D^2	0.367	0.611	0.568	0.4355	0.645
QMD	0.082	0.409	0.19	0.1849	0.246
D/QMD	0.452	0.466	0.533	0.4069	0.561
D^2/QMD	0.407	0.573	0.569	0.4406	0.625
SBA	-0.347	-0.081	-0.148	0.0023	-0.072
BAL	-0.473	-0.282	-0.346	-0.2227	-0.356
CR	0.429	0.397	0.247	0.2607	0.185
TPH	-0.365	-0.31	-0.295	-0.1407	-0.239
SI_{fia}	-0.062	0.204	0.131	0.0515	0.138
BAWHT	-0.024	0.047	-0.062	-0.0523	-0.124
DI	0.059	-0.168	-0.016	0.0648	0.099
PI	0.07	-0.13	0.052	0.0725	0.135
DD5	-0.198	-0.118	0.203	0.1147	0.25
GSP	0.177	0.146	0.046	-0.013	0.161
MAP	-0.193	0.094	-0.032	-0.0659	0.082
MAT	-0.329	-0.134	0.193	0.0318	0.201
MNDVI	0.042	0.082	-0.016	0.0027	-0.015
MTWM	-0.103	-0.107	0.197	0.1205	0.252

Table 4.8. Coefficients and fit statistics of species-specific diameter growth models in the three forms (Equations-1, 2, and 3) with lnDDS as the response variable and measured SI (SI_{fia}), imputed SI (SI_{impt}) and biogeoclimatic (*BGC*) predictors successively substituting the site variable.

Para-	Red pine models			N. white-cedar models			Sugar maple models		
meters	~SI_{fia}	~SI_{impt}	~*BGC*	~SI_{fia}	~SI_{impt}	~*BGC*	~SI_{fia}	~SI_{impt}	~*BGC*
Intercept	5.01935	5.24531	2.86382	1.29009	1.42301	2.97122	4.05131	3.97291	2.63501
1/D	-27.7165	-27.1814	-27.7361				-20.4084	-20.2670	-20.1392
D				0.10555	0.11452	0.10399	0.078745	0.082584	0.07803
D^2							-0.00060	-0.00064	-0.00062
D/QMD	0.56589	0.60674	0.66951				-0.56119	-0.59852	-0.48859
D^2/QMD	-0.00965	-0.01050	-0.01333	-0.02509	-0.02899	-0.02465			
SBA	-0.01506	-0.01338	-0.00913	-0.00383	-0.00402	-0.00303	-0.02781	-0.02656	-0.02459
BAL	-0.00820	-0.00923	-0.00814				-0.00203	-0.00237	-0.00255
CR	0.023804	0.024117	0.02794	0.029853	0.029167	0.028076	0.022762	0.022778	0.023408
CR^2	-5.0E-05	-6.1E-05	-0.00011	-0.00016	-0.00015	-0.00015	-0.00011	-0.00011	-0.00014
SI	0.018899	0.004485		0.017876	-0.00259		0.017193	0.017937	
DD5			-0.00035			0.001662			
MAP.DI			1.63E-05			-3.3E-06			-2.4E-05
DI			-0.01315						0.01791
MTWM			0.019925			-0.02647			
MAT			-0.01949			-0.00668			0.009011
GSP						0.003453			0.002449
BAWHT						0.000594			0.000266
Adj. R^2	0.4734	0.4669	0.5244	0.4087	0.4028	0.4266	0.4043	0.4048	0.4165
RSS	2845.92	2880.99	2568.91	4104.41	4144.93	3978.07	5727.13	5722.40	5608.33
RSE	0.5996	0.6033	0.5698	0.6439	0.6471	0.6341	0.7362	0.7359	0.7287
DF	7914	7914	7910	9898	9898	9893	10565	10565	10561
F-stat	891.22	868.33	728.96	1142.1	1114.8	670.96	798.61	800.24	581.59
AIC	14392.29	14489.31	13588.92	19399.09	19496.4	19099.42	23547.14	23538.4	23333.47

Table 4.8 (continued). RSS: residual sum of square; RSE: residual standard error; and DF: degrees of freedom

Para-meters	Quaking aspen models			N. red oak models		
	$\sim SI_{fia}$	$\sim SI_{impt}$	$\sim BGC$	$\sim SI_{fia}$	$\sim SI_{impt}$	$\sim BGC$
Intercept	5.04331	5.20084	0.42506	2.45192	3.03057	5.60123
1/D	-18.8786	-18.2105	-17.9302			
D				0.080475	0.079799	0.067672
D^2				-0.00029	-0.00034	-0.00022
D/QMD				0.44995	0.33575	0.63932
D^2/QMD				-0.01105	-0.00748	-0.01172
SBA	-0.00924	-0.00789	-0.00438	-0.02139	-0.01771	-0.01212
BAL	-0.00463	-0.00531	-0.00629			-0.00406
CR	0.022837	0.024561	0.029455	0.007506		0.006722
CR^2	-9.8E-05	-0.00012	-0.00016		7.54E-05	
SI	0.020625	0.008793		0.037313	0.015498	
DD5			-0.00036			0.000988
MAP.DI			1.10E-05			2.33E-05
DI			-0.00652			-0.01682
MAP						-0.00296
MTWM			0.028794			-0.02068
MAT			-0.01125			0.009863
GSP			-0.00129			0.003414
MNDVI			0.85502			
PI						0.012064
Adj. R^2	0.2545	0.2447	0.2827	0.4500	0.4277	0.4807
RSS	3170.98	3212.75	3049.03	1218.02	1267.28	1147.12
RSE	0.5851	0.5889	0.5739	0.6272	0.6397	0.6094
DF	9262	9262	9256	3096	3096	3088
F-stat	528.4	501.46	305.44	363.72	332.4	192.49
AIC	16378.05	16499.34	16026.54	5923.10	6046.16	5752.94

Table 4.9. Factors prioritization in the three forms of growth models through reduction of variance approach in sensitivity analysis; the numbers in the parentheses represent sensitivity index (i.e., importance ranking of the predictors in the model).

Spp	Models	Ranking of predictor variables (in descending order from left to right) by sensitivity analysis											
Red pine	$\sim SI_{fia}$	CR (0.3891)	1/D (0.3863)	D/QMD (0.0660)	SBA (0.0575)	D^2/QMD (0.0503)	BAL (0.0238	CR^2 (0.0138)	SI_{fia} (0.0131)				
	$\sim SI_{impt}$	CR (0.3982)	1/D (0.3703)	D/QMD (0.0756)	D^2/QMD (0.0595)	SBA (0.0453)	BAL (0.0301)	CR^2 (0.0204)	SI_{impt} (0.0006)				
	$\sim BGC$	CR (0.2485)	1/D (0.1791)	MAPDI (0.1525)	DI (0.1239)	MAT (0.0854)	MTWM (0.0634)	D^2/QMD (0.0445)	D/QMD (0.0428)	CR^2 (0.0307)	BAL (0.0109)	SBA (0.0098)	DD5 (0.0084)
N. white cedar	$\sim SI_{fia}$	D (0.5997)	CR (0.2327)	D^2/QMD (0.1015)	CR^2 (0.0591)	SI_{fia} (0.0051)	SBA (0.0019)						
	$\sim SI_{impt}$	D (0.6312)	CR (0.1989)	D^2/QMD (0.1213)	CR^2 (0.0466)	SBA (0.0019)	SI_{impt} (0.0001)						
	$\sim BGC$	D (0.4971)	CR (0.1759)	DD5 (0.0999)	D^2/QMD (0.0837)	MTWM (0.0718)	CR^2 (0.0443)	GSP (0.0159)	MAT (0.0053)	BAWHT (0.0027)	MAPDI (0.0024)	SBA (0.0010)	
Sugar maple	$\sim SI_{fia}$	D (0.5860)	1/D (0.1197)	CR (0.0995)	D^2 (0.0769)	SBA (0.0599)	D/QMD (0.0383)	CR^2 (0.0143)	SI_{fia} (0.0045)	BAL (0.0009)			
	$\sim SI_{impt}$	D (0.6030)	1/D (0.1105)	CR (0.0933)	D^2 (0.0822)	SBA (0.0511)	D/QMD (0.0407)	CR^2 (0.0133)	SI_{impt} (0.0047)	BAL (0.0011)			
	$\sim BGC$	D (0.4097)	MAPDI (0.1507)	DI (0.1299)	1/D (0.0830)	CR (0.0750)	D2 (0.0585)	SBA (0.0334)	D/QMD (0.0207)	CR^2 (0.0164)	MAT (0.0111)	GSP (0.0092)	BAWHT (0.0014)
Quaking aspen	$\sim SI_{fia}$	1/D (0.4838)	CR (0.3953)	SI_{fia} (0.0394)	SBA (0.0368)	CR^2 (0.0338)	BAL (0.0108)						
	$\sim SI_{impt}$	CR (0.4546)	1/D (0.4478)	CR2 (0.0505)	SBA (0.0267)	BAL (0.0141)	SI_{impt} (0.0063)						
	$\sim BGC$	CR (0.2688)	MTWM (0.2668)	1/D (0.1771)	MAPDI (0.0937)	DI (0.0590)	MAT (0.0560)	CR^2 (0.0366)	DD5 (0.0172)	GSP (0.0083)	BAL (0.0081)	MNDVI (0.0050)	SBA (0.0034)
N. red oak	$\sim SI_{fia}$	D (0.7979)	D^2/QMD (0.0667)	D^2 (0.0382)	SBA (0.0327)	D/QMD (0.0304)	SI_{fia} (0.0248)	CR (0.0093)					
	$\sim SI_{impt}$	D (0.8562)	D^2 (0.0573)	D^2/QMD (0.0334)	SBA (0.0245)	D/QMD (0.0185)	CR^2 (0.0068)	SI_{impt} (0.0033)					
	$\sim BGC$	D (0.4145)	MAPDI (0.1474)	DI (0.1287)	DD5 (0.0600)	D^2/QMD (0.0551)	MTWM (0.0536)	D/QMD (0.0452)	MAP (0.0251)	GSP (0.0228)	D^2 (0.0161)	MAT (0.0145)	SBA (0.0077)

The coefficients and fit statistics of the three model forms for each of the species are given in the Table 4.8. Only the coefficients that were statistically significant at 95% confidence level (*p*-values ≤ 0.05) are retained in the model. Table 4.8 shows that for each of the species, the model form including biogeoclimatic variables have the highest adjusted R^2, least variance (residual sum of square), least error, and least F-statistic. The large values of F-statistic compared to the tabulated values at the specified model and error degrees of freedom (p-1, n-p) implies that each of the model coefficients are significantly different from zero. Signs of the coefficients too are noteworthy and make sense for the allometric models. For example, positive coefficients for D and negative coefficients for 1/D, and D^2 corroborate the typical unimodal allometric growth pattern such as mean annual increment (MAI). The negative coefficients with SBA, BAL, and interaction of diameter and relative diameter (D^2/QMD) terms signify the suppression effect of competition on tree growth. The negative coefficients of relative diameter (D/QMD) in the case of sugar maple indicates retarding growth with increasing tree size (DBH) and this behavior can be attributed to the very high shade tolerance characteristics of the species. The CR and measured SI have positive influence on tree growth. The negative coefficient of imputed SI for northern white-cedar is strange; this indicates inefficiency of spatial model of SI for such species that can grow over a wide range of sites, remain dominated for several years, and respond quickly to release at any age. If we compare the error statistics, the models based on measured and imputed SI perform similarly for red pine, sugar maple and quaking aspen. But the model for northern red oak dependent on imputed SI causes largest drop in adjusted R^2 compared to the others. It can also be noticed that the growth of northern red oak and quaking aspen is influenced by relatively larger number of biogeoclimatic variables.

The sensitivity analysis for factors prioritization using reduction of variance approach shows that DBH is the most important factor for red oak, white cedar and sugar maple, whereas crown ratio is more important in red pine and quaking aspen (Table 4.9). As expected, the importance of imputed SI was either similar to or worse than the measured SI. The site index (measured) was found to have more influence on growth compared to the competition parameters such as SBA and BAL for white cedar and quaking aspen. For northern red oak, the SI was even more influential than CR.

The validation and performance evaluation of the three model forms and their variants with switched application of measured and imputed values of SI in the respective models is shown in Table 4.10. The projected diameters, 10 years after the initial measurements, when compared via equivalence test with the measured (interpolated) diameters in the BIA plots, it was found that the measured and projected values by all model forms are similar at 25% region of similarity for slopes and intercepts (graphs not shown). The Tukey's Honest Significant Difference (TukeyHSD) *post hoc* test to compare the difference between the means of each pair of measured and projected diameters as well as among the means of each pair of projected diameters indicated that the model including biogeoclimatic variables are superior (closer to the measurements) and involved least RMSE and bias. For white cedar, sugar maple and red oak, the means of measured and predicted diameters did not differ significantly at 95% level of confidence (Table 4.10). However, in the case of quaking aspen, the mean diameter estimates by all the models differed significantly from the mean of measured diameters at 95% level of confidence. In the case of red pine, the models with measured and imputed SI poorly estimated the mean diameter growth per decade. Poor correspondence of mean diameter growth obtained from the BIA data and the predictions by the three model forms can also be attributed to the quality of BIA data, especially the long gap for remeasurements and ocular estimation of some variable such as crown ratio.

Table 4.10. RMSE and bias of estimates along with the evaluation of the difference between the means of each pair of diameter projection methods* by Tukey's Honest Significant Difference (TukeyHSD) *post hoc* test. The adjusted *p*-values in bold case (less than 0.05) imply that the related pair of methods differs significantly

Pair of Methods	Red pine			N. white-cedar			Sugar maple			Quaking aspen			N. red oak		
	TukeyHSD p-adjusted	RMSE	Bias	TukeyHSD p-adjusted	RMSE	Bias	TukeyHSD p-adjusted	RMSE	Bias	TukeyHSD p-adjusted	RMSE	Bias	TukeyHSD p-adjusted	RMSE	Bias
B-A	**<0.05**	2.01	-1.43	0.76	0.80	-0.25	0.08	1.79	-1.23	**<0.05**	1.96	-1.09	0.75	1.91	-1.01
C-A	**<0.05**	2.04	-1.45	0.76	0.81	-0.25	0.07	1.80	-1.24	**<0.05**	2.10	-1.31	0.78	1.95	-0.98
D-A	0.41	1.63	-0.74	0.93	0.78	-0.18	0.11	1.75	1.17	**<0.05**	1.97	-1.10	0.76	1.95	-1.00
E-A	**<0.05**	2.07	-1.50	0.85	0.83	-0.22	0.08	1.79	-1.23	**<0.05**	2.08	-1.28	0.77	1.88	-0.99
F-A	**<0.05**	2.02	-1.44	0.77	0.82	-0.24	0.07	1.80	-1.24	**<0.05**	2.05	-1.24	0.77	1.96	-0.99
C-B	1.00	0.13	-0.02	1.00	0.06	0.00	1.00	0.08	-0.02	0.98	0.30	-0.22	1.00	0.21	0.03
D-B	0.49	0.77	0.69	1.00	0.14	0.07	1.00	0.10	0.05	1.00	0.41	-0.01	1.00	0.34	0.01
E-B	1.00	0.19	-0.07	1.00	0.12	0.03	1.00	0.07	-0.01	0.99	0.30	-0.19	1.00	0.25	0.03
F-B	1.00	0.10	-0.01	1.00	0.06	0.01	1.00	0.02	-0.01	1.00	0.19	-0.14	1.00	0.18	0.03
D-C	0.45	0.79	0.71	1.00	0.14	0.07	1.00	0.10	0.07	0.99	0.42	0.21	1.00	0.33	-0.02
E-C	1.00	0.11	-0.04	1.00	0.13	0.03	1.00	0.02	0.01	1.00	0.11	0.03	1.00	0.23	-0.01
F-C	1.00	0.04	0.02	1.00	0.02	0.00	1.00	0.08	0.01	1.00	0.13	0.08	1.00	0.10	-0.01
E-D	0.38	0.85	-0.76	1.00	0.17	-0.04	1.00	0.10	-0.06	0.99	0.43	-0.18	1.00	0.43	0.02
F-D	0.48	0.77	-0.70	1.00	0.14	-0.07	1.00	0.10	-0.06	1.00	0.38	-0.13	1.00	0.30	0.03
F-E	1.00	0.15	0.02	1.00	0.11	-0.03	1.00	0.07	0.00	1.00	0.18	0.04	1.00	0.30	0.00

*Methods: A: time-2 diameter measured (interpolated) 10 years later from the first measurement; B: time-2 diameter predicted by SI$_{fia}$ model; C: time-2 diameter predicted by SI$_{impt}$ model; D: time-2 diameter predicted by BGC model; E: time-2 diameter predicted by SI$_{fia}$ model when imputed SI values are used instead of measured SI; F: time-2 diameter predicted by SI$_{impt}$ model when measured SI values are used instead of imputed SI.

4.5. Discussion

The biogeoclimatic approach of site productivity mapping fundamentally involves measurements of SI as response and biotic, climate, soil, and physiographic variables as explanatory variables for the locations of sample trees (site trees) from stands distributed throughout the region of interest. The measured SI is then traditionally related to the explanatory variables by means of regression analysis. The practical application of SI maps in forest growth and yield mapping require analytical evaluation of accuracy with respect to modeling approaches and inclusion of explanatory biogeoclimatic variables. Because of the complex nature of ecological systems particularly in regional level studies with limited availability of spatially explicit auxiliary predictors in public domain, it is always challenging to identify and select variables for large area SI mapping. The scale dependent correlation and interaction among the predictor biogeoclimatic variables and the inherent assumptions regarding independence, homocedasticity, and normality in the traditional multivariate modeling approach has opened ways for sophisticated machine learning approaches. Hence, I have selected freely available spatial variables believed to directly influence site quality and the non-parametric RF-kNN modeling approach which is assumed to be free from the general assumptions of parametric regression. The accuracies of species-specific imputed site index layers are satisfactory as the developed individual models possessed more than 86% explanatory power for the variations (Table 4.6). Although Avery and Burkhart (1994) have mentioned limited success in SI prediction mapping, Klinka and Carter (1990) reported strong relationship ($R^2 = 0.84$) between SI of Douglas fir and spatial metrics of climate, soil moisture, and nutrients. Similarly, Sharma et al. (2012) developed SI models for Norway spruce and Scots pine by using national forest inventories and different combinations of site and climatic variables that explained a large part of the total variations (adjusted R^2 of 0.86 and 0.72 for the spruce and pine respectively). Despite some reported deficiencies, spatially explicit SI maps have the intuitive appeal since the productivity models based on data from sites of known/ measurable quality can be applied to the sites where site parameters are difficult to measure.

No model can perfectly portray the growth pattern of trees since the complex ecological systems are dynamic, interactive and dependent on several environmental and socio-economic factors (Zeide, 1993). It is difficult to incorporate the effects of numerous factors including disturbance in a growth model. We have not accounted for the effect of catastrophic mortality or excessive harvesting on the growth, primarily because of the unavailability of true coordinates of the FIA plots. As the disturbance component is not considered, the assumption that the initial stand condition prevailed during the growth period suffers from a drawback.

The scatter plot of DDS against D^2 revealed an approximately linear trend which implies quadratic relationship of the growth with DBH (or linear relationship with initial tree basal area). I did not use polynomial growth forms which lack biological interpretation (Zeide, 1993). The developed diameter growth models explained the variability to varying degrees. The models possessed the desirable statistical characteristics of homogenous residual variance with increasing DBH. The adjusted R^2 statistic obtained in this study are similar to or better than the reported values by Lessard et al. (2001); they obtained the fit indexes (analogous to R^2) of 0.438, 0.36, 0.363, 0.246, and 0.227, respectively for red pine, white-cedar, soft maples, quaking aspen, and red oak by using FIA data from undisturbed, mixed species, and mixed age stands in Minnesota. The fit statistics in this study are also better than the ones reported by Shifley (1987).

The varying degrees of dependency of diameter growth on site and competition elements is justified from the notion that SI represents a only fraction of potential growth of an individual tree, and the effect of other parameters depend on competition (e.g., trees have larger CR in understocked stands with abundant nutrients, that creates positive effect on growth). In addition, the growth rate for a given DBH and SI decreases as BAL increases, and reach zero asymptotically. Teck and Hilt (1991) have reported positive correlations of tree diameter and diameter growth, and also site index and diameter growth. This study indicates relatively low ranking of SI in the models (Table 4.9), as also noticed by Wykoff (1990) who relates the cause to the selection of sample trees from irregular stands spread over large areas. The positive coefficients of CR and negative coefficients of SBA and BAL in the calibrated models suggest that growth increments are larger for dominant trees with large crown from low density stands in contrast to suppressed trees of short crowns from high density stands.

The diameter growth patterns for each of the target species are biologically justified as shown by the unimodel positively skewed shapes in the Figure 4.4. The under estimation of diameter growth (revealed by negative biases in the Table 4.10) by each of the models is in line with the previous study by Zhao et al. (1988) who also applied a similar model form; however, Froese and Robinson (2007) and Holdaway and Brand (1983) have observed overpredictions with different model forms.

The individual tree level general growth trend against size maintains a certain degree of rigidity (Figure 4.4), which is an asset for applications in simulation frameworks. Since growth equations serve as building blocks for simulation programs and commercial operational applications largely apply timber simulations/projections, the large tree growth equations are formulated and alternative versions of the models are evaluated to see how well the predictions matches the observed growths.

Discrepancies in the inventory designs/systems of FIA and BIA have likely introduced higher error in the validation of growth predictions at BIA locations. Tree CR (which is the most important predictor for red pine and quaking aspen) was inconsistently measured, on different scales (e.g. coding) in different reservations, in the BIA plots. The long gaps between two successive measurements (up to 17 years) in the BIA plots have further consequence in the calculation of 10-year periodic diameter growth. The SI data was not available for all BIA plots/reservations and the BIA measurement methods (particularly age and height) of site trees may have been different from the ones adopted in the FIA design. The distributions of BIA validation plots are over narrow areas/ pockets in MN and WI, and the application of developed models to local conditions still suffer shortfalls since the predicted growth rates represent average rates over the entire LS region (despite the fact that SI or BGC variables account for the site variations).

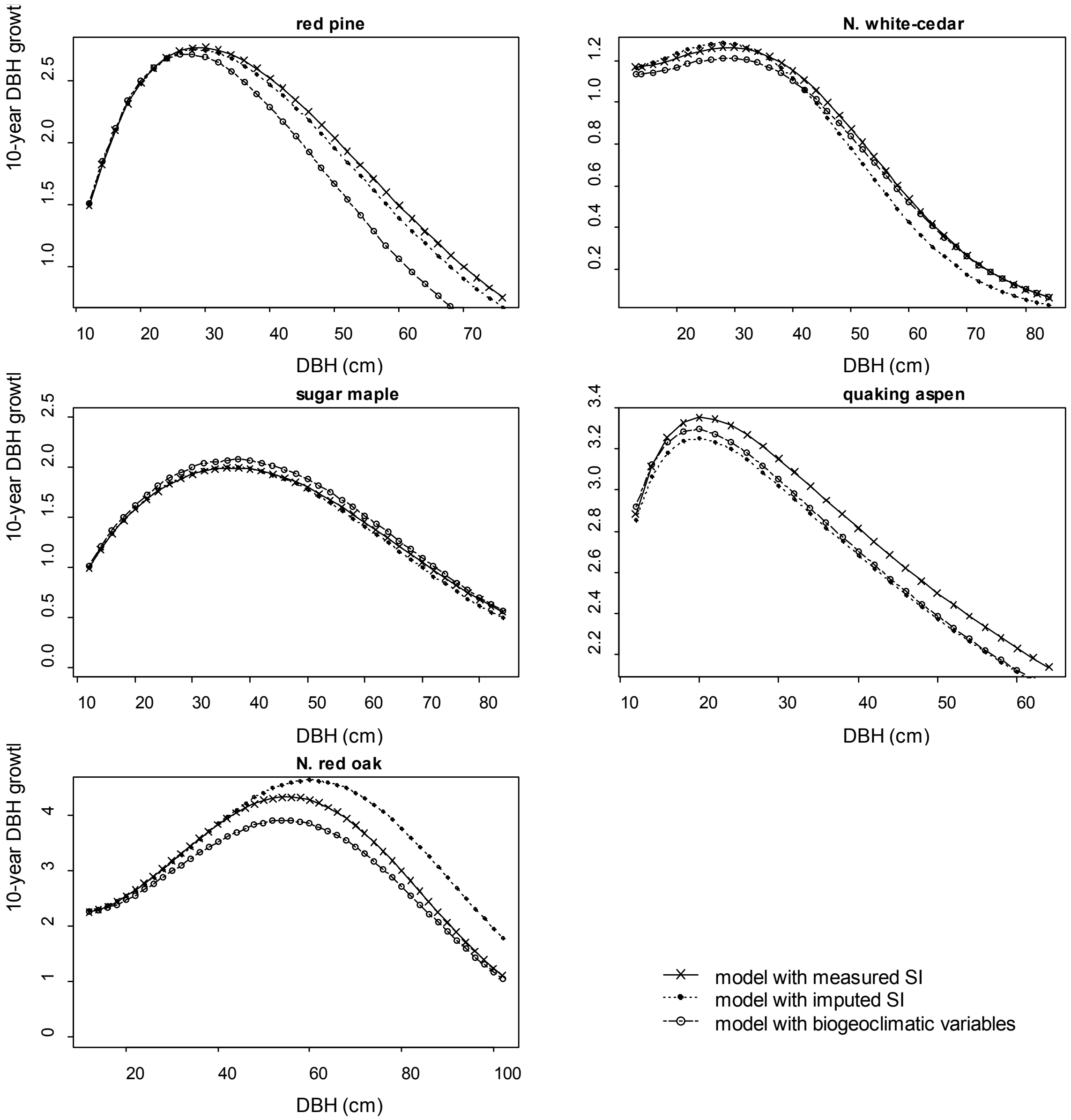


Figure 4.4. Comparisons of 10-year predicted diameter growth surfaces based on models consecutively consisting of measured site-index, imputed site-index, and biogeoclimatic variables for each of the target species. The decadal growths are derived from lnDDS by varying only the initial DBH and using the median values of other predictors.

4.6. Conclusions

i. The site index imputation models dependent on biogeoclimatic variables strongly explained the variance in the training dataset; however, sensitivity of the predictors varied with the species.

ii. Per decade diameter (or basal area) growth of the sample trees were found to be positively related to initial DBH, crown ratio, and site index and negatively related to stem density, stand basal area, and cumulative basal area of larger trees. This indicates positive effect of site quality and negative effect of competition elements on tree growth.

iii. Diameter growth models based on biogeoclimatic variables better explained the variance compared to the models based on measured or imputed site index.

iv. The spatial model of site index for northern white cedar was inefficient because it revealed a negative coefficient of the imputed variable in the growth model. This indicates inability of site index models for species that grow on a wide range of sites, remain dormant for several years, and respond quickly to release at any age.

v. The success of imputed site index in diameter growth projection was either similar to or worse than the measured site index and varied with species.

vi. Diameter growth models based on biogeoclimatic variables were superior in predicting the diameter growth as verified with the independent dataset from BIA.

vii. Site index models (except for some species) have intuitive appeal since site data from a reference sample can be extended to areas where sample site trees are not available.

viii. Since this study applied fuzzed-swapped coordinates of the FIA plots, better spatial models of site index can be prepared with actual coordinate of the plots.

5. Conclusions

Foundation of this work is built on the strengths of the Random Forest based k-Nearest Neighbor (RF-kNN) imputation algorithm that combines a sample of geo-referenced ground inventory data with geospatial datasets for spatially explicit prediction of forest structural attributes. I applied the novel RF-kNN imputation approach to generate forest inventory across large spatial extents by coupling remote sensing and geospatial data with sample inventory collected by different sampling methods. The algorithm is highly acknowledged in contemporary forestry research focused on large-area assessment of inventory attributes to guide operational management and strategic planning. The strength particularly lies in the distribution free assumptions and the algorithm's capability to provide relative importance of selected variables while simultaneously predicting multiple inventory attributes for large (even inaccessible) target area. This study evaluated accuracy of imputation estimates produced by using optical and LiDAR remote sensing and other publicly available geospatial layers combined with the forest inventory data at multiple spatial scales (regional and local). The accuracy, particularly for small-area operational requirements, was found to be dependent on the characteristics of geospatial datasets as well as the sample field inventory. The training dataset developed from high resolution geospatial layers intersected with the actual coordinates of sample inventory plots were found to be efficient and precise in generating resource stock and distribution information. The imputation products are useful to forest managers and policy makers for enhanced production of goods and ecosystem services.

To evaluate the impact of data-driven modeling approaches and optimization criteria on the accuracy of biomass estimates at small and large spatial scales, two new imputation models were developed and two extant models produced by USFS (Blackard et al., 2008) and NBCD (Kellndorfer et al., 2012) were considered in this study. Using publicly available remote sensing and other biogeoclimatic spatial layers coupled with the national forest inventory (FIA) data from a large part of Michigan, the new models were built in two contrasting ways: (i) a limited number (total five) of spatial predictors were attached to the FIA data under the policy restrictions on disclosing actual plot locations and a model called *Actu.imput* was formulated to develop a high resolution map (30 m pixel); (ii) a large number (total eleven) of spatial predictors were related to FIA data via the fuzzed-swapped plot coordinates available in the online database and a model called *Fuzz.imput* was formulated to develop a coarse resolution map (250 m pixel). The biomass estimates of the four imputation models at plot, stand and county scales, validated against separate datasets revealed that the prediction accuracy improves with increasing size of the target area. The actual coordinate based new model (*Actu.imput*) relatively performed better for the plot (pixel) level prediction but none of the models were reliable since the estimates were not statistically equivalent to the field plot observations. The stand level estimates by the *Actu.imput* model were best in terms of RMSE; however, the USFS and NBCD models also generated estimates statistically equivalent to the field observations. This implies that an imputation model based on limited number of sensible spatial predictors attached to the inventory data via the actual plot coordinates can provide reliable biomass estimates at a stand or larger spatial extent. Since the model based on fuzzed-swapped plots provided statistically equivalent results as with the model derived from true coordinate data, it can be concluded that a high degree of sophistication or adjustments to offset the spatial mismatch of the plot data and corresponding spatial predictors is not necessary for large area estimation. As in many published works, insensitivity of optical remote sensing data was obvious from the validation analysis since the models produced under estimation in large biomass plots and over estimation in low biomass plots. Among the predictors selected, basal area weighted height (BAWHT) was found to be the most influential in the two new models. In the *Actu.imput* model with five predictors, the ranking in the order of decreasing importance followed BAWHT, land cover, Landsat image derived normalized difference vegetation index (NDVI), and MODIS time-series images derived disturbance (MODIS-slope), and elevation. In the *Fuzz.imput* model including 11 different predictors, climatic variables (precipitation and temperature) were the next important predictors after BAWHT. The performance of models varied with the size of target area, choice of statistical measure to test goodness-of-fit, and the quality of calibration and validation data.

The potential of combining the strength of LiDAR data with inexpensively collected sample inventory data via variable-radius plot (VRP) or point sampling has long been realized in remote sensing community to enhance cost-effective geospatial inventory. Accuracy of inventory estimates by the RF-kNN imputation model based on the integration of indeterminate size VRP sampling data and LiDAR derived metrics was evaluated against the estimates by a similar model developed from the integration of fixed radius plot (FRP) sampling data and LiDAR derived metrics. The FRP sampling data was used as reference to compare the coinciding plot level standing volume estimates by seven different VRP models developed on the basis of sampling of six conifer stands in the Ford Forest area of Michigan Tech. It was found that the VRP data based models are capable of estimating volume statistically equivalent to the FRP data based model predictions. The most efficient VRP model in terms of bias was associated with the basal area factor (BAF) 9 for all the stands together; however, BAF 10 was the best for older stands only. BAF 10 was concluded to be the most effective for the target area inventory since it tallied at least four trees per plot, and the sampling device (i.e. prism) along with associated expansion factors is easily available in the market. The study revealed that VRP and FRP data from separate stands can also be combined to formulate a spatial model at an optimal grid resolution of LiDAR metrics. The suitability of VRP model in older stands (with large volume) was supported by the observations that use of fixed plots instead of variable plots for model training improved the accuracy only by a small margin. Further a combination of VRP data from younger stands and FRP data from older stands, or only VRP data from all stands can be used to formulate a generalized model with some compromise in accuracy. However, use of a single BAF (especially BAF 10) for all stands is useful as this practice overcomes the practical difficulties associated with the use of several BAFs.

The characteristic of imputation to generate an estimate of target inventory attribute at unsampled points based on the observations from a sample of reference points is particularly attractive for the attributes such as forest site index (SI) which require meticulous efforts in parameters measurement (e.g., tree age and height) and are sometime difficult or impossible to measure. The comprehensive FIA database with species-specific SI estimated per plot was combined with biogeoclimatic spatial layers (linked via fuzzed-swapped coordinates) in the RF-kNN framework to develop spatially explicit maps of SI for five major species (red pine, northern white cedar, sugar maple, quaking aspen, and northern red oak) of the Lake States (MI, WI, and MN). Accuracy of the imputed SI (produced as raster at 250 m resolution) was validated against measured SI at the plots other than the ones used for model training. Analysis showed that utility of the SI models vary with species, and specially models for shade tolerant species or others that grow over a wide range of sites are less reliable. Additionally, when large tree diameter growth models were formulated by using three proxies for site quality namely measured SI, imputed SI, and a combination of biogeoclimatic variables, negative coefficient of imputed SI was found for the white cedar model; this is not reasonable because all SI coefficients for other species were positive. Tree diameter growth predictions based on models using measured SI and imputed SI when compared with the field observation (from BIA plots), it was found that statistically significant difference prevail in the predictions of red pine and quaking aspen. As expected the sensitivity of imputed SI in the growth projection are either similar to or poorer than measured SI. The spatial maps of SI have intuitive appeal since one can estimate site quality even for the areas that are presently devoid of forests; it can guide crop selection for plantation.

Appendix 1. Description of 90 different LiDAR metrics used in this study

Predictor	Description
PropT	proportion of total return >1.5 m (total returns >1.5 m /total returns)
Prop1	proportion of first return >1.5 m (first returns >1.5 m /total returns >1.5
Prop2	proportion of second return >1.5 m (second returns >1.5 m /total
Prop3	proportion of third return >1.5 m (third returns>1.5 m /total returns>1.5
Prop4	proportion of fourth return >1.5 m (fourth returns>1.5 m /total return
Prop5	proportion of fifth return >1.5 m (fifth returns>1.5 m /total return
ElevMin	Elevations minimum
ElevMax	Elevations maximum
ElevMean	Elevations mean
ElevMode	Elevations mode
ElevSD	Elevations standard deviation
ElevVar	Elevations variance
ElevCV	Elevations coefficient of variation
ElevIQR	Elevations interquartile range
ElevSkew	Elevations skewness
ElevKurt	Elevations kurtosis
ElevAAD	Elevations average absolute deviation
EMADmed	Median of the absolute deviations from the overall median of
EMADmod	Mode of the absolute deviations from the overall mode of elevations
ElevL1	Elevations first L-moment
ElevL2	Elevations second L-moment
ElevL3	Elevations third L-moment
ElevL4	Elevations fourth L-moment
ElevLCV	Elevations L-moment coefficient of variation
ElevLskew	Elevation L-moment skewness
ElevLkurt	Elevation L-moment kurtosis
ElevP01	Elevations 1st percentile
ElevP05	Elevations 5th percentile
ElevP10	Elevations 10th percentile
ElevP20	Elevations 20th percentile
ElevP25	Elevations 25th percentile
ElevP30	Elevations 30th percentile
ElevP40	Elevations 40th percentile
ElevP50	Elevations 50th percentile
ElevP60	Elevations 60th percentile
ElevP70	Elevations 70th percentile
ElevP75	Elevations 75th percentile
ElevP80	Elevations 80th percentile
ElevP90	Elevations 90th percentile
ElevP95	Elevations 95th percentile

ElevP99	Elevations 99th percentile
CRR	Canopy relief ratio ((mean-min)/(max-min))
EQM	Elevation quadratic mean
ECM	Elevation cubic mean
IntMin	Intensity minimum
IntMax	Intensity maximum
IntMean	Intensity mean
IntMode	Intensity mode
IntSD	Intensity standard deviation
IntVar	Intensity variance
IntCV	Intensity coefficient of variation
IntIQR	Intensity interquartile range
IntSkew	Intensity skewness
IntKurt	Intensity kurtosis
IntAAD	Intensity average absolute deviation
IntL1	Intensity first L-moment
IntL2	Intensity second L-moment
IntL3	Intensity third L-moment
IntL4	Intensity fourth L-moment
IntLCV	Intensity L-moment coefficient of variation
IntLskew	Intensity L-moment skewness
IntLkurt	Intensity L-moment kurtosis
IntP01	Intensity 1st percentile
IntP05	Intensity 5th percentile
IntP10	Intensity 10th percentile
IntP20	Intensity 20th percentile
IntP25	Intensity 25th percentile
IntP30	Intensity 30th percentile
IntP40	Intensity 40th percentile
IntP50	Intensity 50th percentile
IntP60	Intensity 60th percentile
IntP70	Intensity 70th percentile
IntP75	Intensity 75th percentile
IntP80	Intensity 80th percentile
IntP90	Intensity 90th percentile
IntP95	Intensity 95th percentile
IntP99	Intensity 99th percentile
Density1	overstory canopy density as % of first return >3m(Ist returns >3m/total
Density2	overstory canopy density as % of all return >3m (all returns > 3m/total
Density3	Percentage first returns above mean
Density4	Percentage first returns above mode
Density5	Percentage all returns above mean
Density6	Percentage all returns above mode

Strata0	proportion of ground return
Strata1	proportion of above-ground returns below 1.5 m
Strata2	proportion of vegetation returns above 1.5 m and below 6 m
Strata3	proportion of vegetation returns above 6 m and below 10.6 m
Strata4	proportion of vegetation returns above 10.6 m and below 15.2 m
Strata5	proportion of vegetation returns above 15.2 m and below 19.8 m
Strata6	proportion of vegetation returns above 19.8 m

Appendix 2. Plot level volume estimates based on the FRP and coinciding VRP sampling schemes in the field

Stand ID	Plot ID	FRP volume $(m^3.ha^{-1})$	BAF 5 volume $(m^3.ha^{-1})$	BAF 7 volume $(m^3.ha^{-1})$	BAF 9 volume $(m^3.ha^{-1})$	BAF 10 volume $(m^3.ha^{-1})$	BAF 12 volume $(m^3.ha^{-1})$	BAF 14 volume $(m^3.ha^{-1})$	BAF 15 volume $(m^3.ha^{-1})$
6	1	193.32	147.12	186.04	163.71	181.90	197.77	230.74	195.48
6	2	218.76	212.32	224.23	227.96	218.98	240.40	228.49	232.45
6	3	131.95	140.94	144.43	142.32	143.01	152.01	158.50	145.75
6	4	188.04	142.43	189.26	186.85	192.06	209.04	243.88	236.43
6	5	269.35	172.11	207.68	221.50	246.11	233.17	245.17	262.68
6	6	173.96	158.30	155.01	169.36	188.18	193.01	203.60	193.78
6	7	176.75	174.93	171.54	190.34	181.01	148.53	57.83	61.96
6	8	291.27	156.46	172.07	174.79	178.27	196.95	164.06	156.74
6	9	266.00	281.26	348.61	374.15	402.71	420.73	383.08	360.48
6	10	148.24	173.41	146.23	134.93	119.90	124.95	145.77	156.18
6	11	147.50	117.28	132.93	151.69	151.99	132.00	154.00	124.33
6	12	243.81	270.08	247.56	211.44	234.94	242.57	283.00	254.68
6	13	214.03	211.50	250.05	251.91	264.20	272.19	252.50	242.89
10	1	47.52	39.26	54.96	60.36	58.52	63.26	73.80	62.72
10	2	44.04	52.00	51.40	31.51	35.01	42.02	23.53	25.21
10	3	39.77	28.55	32.71	33.36	37.07	33.75	39.37	42.18
10	4	55.61	67.07	62.04	46.15	40.45	25.69	29.97	32.11
12	1	34.13	52.90	44.85	49.91	55.45	56.99	56.01	47.66
12	2	67.69	111.25	116.96	118.82	123.21	109.24	100.39	96.03
12	3	55.29	60.83	31.22	16.65	18.50	22.21	25.91	27.76
12	4	44.77	70.26	71.39	63.03	60.29	59.34	69.23	74.17
12	5	25.07	51.32	49.09	63.11	61.22	73.47	85.71	91.83
12	6	12.12	42.19	45.39	56.05	56.01	56.24	60.24	55.82
12	7	45.08	85.38	90.02	97.64	99.87	87.94	102.59	109.92
17	1	71.89	82.76	83.07	91.65	76.25	80.34	66.18	70.91
17	2	74.20	96.84	80.64	83.95	67.44	59.93	64.23	55.94
17	3	61.37	49.22	68.90	52.86	51.71	62.06	72.40	77.57
17	4	89.16	120.79	138.09	118.55	131.72	143.92	151.53	162.35
17	5	55.30	87.54	107.16	84.18	93.54	75.21	49.17	52.69
17	6	75.27	64.61	65.24	63.95	71.06	57.37	66.93	71.72
17	7	70.07	48.54	46.37	39.07	43.41	52.10	40.55	43.45
17	8	59.50	66.48	62.63	66.51	63.06	75.67	88.28	74.50
17	9	50.43	54.72	39.67	30.72	34.13	40.96	37.63	40.32
17	10	43.93	89.02	91.44	86.58	84.68	91.76	95.01	101.79
19	1	99.60	99.07	115.66	121.20	124.25	111.94	102.28	109.59

19	2	70.42	95.27	104.96	102.69	106.59	100.45	101.96	95.02
19	3	61.81	49.43	40.26	34.22	38.02	31.38	36.61	27.61
19	4	72.22	67.46	68.29	47.24	52.49	52.08	49.03	37.46
19	5	70.93	97.72	77.62	69.53	69.56	48.12	56.14	60.15
19	6	67.45	54.90	54.29	54.70	40.48	36.74	42.87	45.93
19	7	68.26	63.20	82.15	63.10	60.82	72.98	72.67	77.86
19	8	49.65	56.06	46.41	50.84	56.49	67.79	50.03	38.73
19	9	53.81	61.62	45.03	50.02	55.58	54.02	63.02	67.52
24	1	185.73	121.31	126.52	127.74	111.84	87.63	102.23	91.46
24	2	150.87	65.58	71.89	80.44	89.38	107.25	103.88	111.30
24	3	122.06	121.21	133.85	135.06	150.06	111.61	113.56	121.67
24	4	183.53	193.65	242.92	221.82	232.29	250.99	263.85	282.69

References

Adler, J. 2010. R in a Nutshell. O'Reilly Media, Inc.

Ahamed, T.; Tian, L.; Zhang, Y.; Ting, K.C. 2011. A review of remote sensing methods for biomass feedstock production. *Biomass and Bioenergy* **35**(7): 2455-2469.

Anaya, J.A.; Chuvieco, E.; Palacios-Orueta, A. 2009. Aboveground biomass assessment in Colombia: A remote sensing approach. *Forest Ecology and Management* **257**(4): 1237-1246.

Avery, G.; Newton, R. 1965. Plot sizes for timber cruising in Georgia. *Journal of Forestry* **63**(12): 930-932.

Avery, T.E.; Burkhart, H.E. 1994. Forest Measurements, Fourth Edition. McGraw-Hill, Inc

Baccini, A.; Friedl, M.A.; Woodcock, C.E.; Warbington, R. 2004. Forest biomass estimation over regional scales using multisource data. *Geophysical Research Letters* **31**(L10501): doi:10.1029/2004GL019782.

Beaulieu, J.; Raulier, F.; Pregent, G.; Bousquet, J. 2011. Predicting site index from climatic, edaphic and stand structural properties for seven plantation-grown conifer species in Quebec. *Canadian Journal of Forest Research* **41**(4): 682-693.

Berndt, L.W. 1988. Soil survey of Baraga county. USDA, Soil Conservation Service. Washington, D.C.

Blackard, J.A.; Finco, M.V.; Helmer, E.H.; Holden, G.R.; Hoppus, M.L.; Jacobs, D.M.; Lister, A.J.; Moisen, G.G.; Nelson, M.D.; Riemann, R.; Ruefenacht, B.; Salajanu, D.; Weyermann, D.L.; Winterberger, K.C.; Brandeis, T.J.; Czaplewski, R.L.; McRoberts, R.E.; Patterson, P.L.; Tymcio, R.P. 2008. Mapping U.S. forest biomass using nationwide forest inventory data and moderate resolution information. *Remote Sensing of Environment* **112**(4): 1658-1677.

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

Breiman, L. 2001a. Statistical modeling: The two cultures. *Statistical Science* **16**(3): 199-231.

Bropleh. 1967. Applicability of point sampling to the forests of Liberia. *MS Thesis*, Department of Forest Management. Oregon State University.

Brosofske, K.; Froese, R.E.; Falkowski, M.J.; Banskota, A. 2014. A review of methods for mapping and prediction of inventory attributes for operational forest management. *Forest Science* **60**(2): 1-24.

Burkhart, H.E.; Farrar, K.D.; Amateis, R.L.; Daniels, R.F. 1987. Simulation of individual tree growth and stand development in loblolly pine plantations on cutover, site-prepared areas. Publication No. FWS-1-87. Department of Forestry, Virginia Tech, Blacksburg, Virginia 24061.

Burkhart, H.E.; Tomé, M. 2012. Modeling forest trees and stands. Dordrecht Heidelberg New York London, Springer.

Carmean, W.H.; Hahn, J.T.; Jacobs, R.D. 1989. Site index curves for forest tree species in the eastern United States. Gen. Tech. Rep. NC-128. USDA Forest Service, North Central Forest Experiment Station, St. Paul, MN. .

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**(5): 893-903.

Chen, Q.; Laurin, G.V.; Battles, J.J.; Saah, D. 2012. Integration of airborne LiDAR and vegetation types derived from aerial photography for mapping aboveground live biomass. *Remote Sensing of Environment* **121**: 108-117.

Chirici, G.; Barbati, A.; Corona, P.; Marchetti, M.; Travaglini, D.; Maselli, F.; Bertini, R. 2008. Non-parametric and parametric methods using satellite images for estimating growing stock volume in alpine and Mediterranean forest ecosystems. *Remote Sensing of Environment* **112**(5): 2686-2700.

Cížková, L.; Cížek, P. 2012. Handbook of computational statistics: concepts and methods. J. E. Gentle, W. K. Hardle and Y. Mori. Heidelberg Dordrecht London New York, Springer.

Cole, D.M.; Stage, A.R. 1972. Estimating future diameters of lodgepole pine trees. *Research Paper INT-131*, USDA Forest Service, Intermountain Forest and Range Experiment Statation, Ogden, Utah**:** 9.

Coulston, J.W.; Moisen, G.G.; Wilson, B.T.; Finco, M.V.; Cohen, W.B.; Brewer, C.K. 2012. Modeling percent tree canopy cover: A pilot study. *Photogrammetric Engineering and Remote Sensing* **78**(7): 715-727.

Coulston, J.W.; Riitters, K.H.; McRoberts, R.E.; Reams, G.A.; Smith, W.D. 2006. True versus perturbed forest inventory plot locations for modeling: a simulation study. *Canadian Journal of Forest Research* **36**(3): 801-807.

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

Crookston, N.L.; Rehfeldt, G.E.; Dixon, G.E.; Weiskittel, A.R. 2010. Addressing climate change in the forest vegetation simulator to assess impacts on landscape forest dynamics. *Forest Ecology and Management* **260**(7): 1198-1211.

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

D'Annunzio, R.; Lindquist, E.J.; MacDicken, K.G. 2014. Global forest land-use change from 1990 to 2010: an update to a global remote sensing survey of forests. Food and Agriculture Organization of the United Nations.

Dickmann, D.I.; Leefers, L.A. 2003. The forests of Michigan. Ann Arbor, The University of Michigan Press.

Dixon, G.E. 2002. Essential FVS: A user's guide to the Forest Vegetation Simulator. *Internal Report*, USDA Forest Service, Forest Management Service Centre, Fort Collins, CO**:** 244p (Revised: Aug, 2011).

Ek, A.R.; Birdsall, E.T.; Spears, R. 1981. Total and merchantable tree height equations for Lake States tree species. Technical Report 27. University of Minnesota, College of Forestry and Agricultural Experiment Station.

Ek, A.R.; Robinson, A.P.; Radtke, P.J.; Walters, D.W. 1997. Development and testing of regeneration imputation models for forests in Minnesota. *Forest Ecology and Management* **94**: 129-140.

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**: 235-246.

Evans, J.S.; Murphy, M.A.; Holden, Z.A.; Cushman, S.A. 2011. Modeling species distribution and change using Random Forests in Predictive species and habitat modeling. *Landscape ecology: concepts and applications*. C. Drew, Y. F. Wiersma and F. Huettmann. NY, Springer**:** 139-159.

Falkowski, M.J.; Evans, J.S.; Martinuzzi, S.; Gessler, P.E.; Hudak, A.T. 2009. Characterizing forest succession with LiDAR data: an evaluation for the inland northwest, U.S.A. *Remote Sensing of Environment* **113**(5): 946-956.

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. *Canadian Journal of Forest Research* **40**(2): 184-199.

Falkowski, M.J.; Smith, A.M.S.; Gessler, P.E.; Hudak, A.T.; Vierling, L.A.; Evans, J.S. 2008. The influence of conifer forest canopy cover on the accuracy of two individual tree measurement algorithms using lidar data. *Canadian Journal of Forest Research* **34**(2): 338-350.

Falkowski, M.J.; Smith, A.M.S.; Hudak, A.T.; Gessler, P.E.; Vierling, L.A.; Crookston, N.L. 2006. Automated estimation of individual conifer tree height and crown diameter via two-dimensional spatial wavelet analysis of lidar data. *Canadian Journal of Forest Research* **32**(2): 153-161.

FAO. 2009. Biomass: Essential Climate Variables, Food and Agriculture Organization of the United Nations, Global Terrestrial Observing System Secretariat.

FAO. 2010. Global Forest Resources Assessment 2010. Food and Agriculture Organization of the United Nations. Rome.

Fazakas, Z.; Nilsson, M.; Olsson, H. 1999. Regional forest biomass and wood volume estimation using satellite data and ancillary data. *Agricultural and Forest Meteorology* **98-99**(31 Dec): 417-425.

FGDC. 2008. National Vegetation Classification Standard (version 2). Federal Geographic Data Committee. FGDC Document number FGDC-STD-005-2008. [Online] http://usnvc.org/wp-content/uploads/2011/02/NVCS_V2_FINAL_2008-02.pdf (accessed on July 6, 2013).

FIA. 2013. Forest inventory and analysis (FIA) national program: Data and tools. USDA Forest Service. [Online] http://apps.fs.fed.us/fiadb-downloads/datamart.html (accessed on April 24, 2014).

FIA. 2014. Forest Inventory and Analysis National Program: Data and Tools. National Office, U.S. Forest Service, Arlington, VA 22209. [Online] http://www.fia.fs.fed.us/tools-data/ (accessed on May 5, 2014).

Foody, G.M.; Boyd, D.S.; Cutler, M.E.J. 2003. Predictive relations of tropical forest biomass from Landsat TM data and their transferability between regions. *Remote Sensing of Environment* **85**(4): 463-474.

Franco-Lopez, H.; Ek, A.R.; Bauer, M.E. 2001. Estimation and mapping of forest stand density, volume, and cover type using the *k*-nearest neighbors method. *Remote Sensing of Environment* **77**(3): 251-274.

Frelich, L.E. 2002. Forest dynamics and disturbance regimes: Studies from temperate evergreen- deciduous forests, Cambridge University Press.

Froese, R.E.; Robinson, A.P. 2007. A validation and evaluation of the Prognosis individual-tree basal area increment model. *Canadian Journal of Forest Research* **73**(8): 1438-1449.

Fuchs, H.; Magdon, P.; Kleinn, C.; Flessa, H. 2009. Estimating aboveground carbon in a catchment of the Siberian forest tundra: Combining satellite imagery and field inventory. *Remote Sensing of Environment* **113**: 518-531.

GAP. 2013. National Gap Analysis Program (GAP): Land Cover Data Portal. U.S. Department of the Interior | U.S. Geological Survey. [Online] http://gapanalysis.usgs.gov/gaplandcover/data/ (accessed on April 24, 2014).

Gleason, C.J.; Im, J. 2011. A review of remote sensing of forest biomass and biofuel: options for small-area applications. *GIScience & Remote Sensing* **48**(2): 141-170.

Gleason, C.J.; Im, J. 2012. Forest biomass estimation from airborne LiDAR data using machine learning approaches. *Remote Sensing of Environment* **125**: 80-91.

Gobakken, T.; Naesset, E. 2008. Assessing effects of laser point density, ground sampling intensity, and field sample plot size on biophysical stand properties derived from airborne laser scanner data. *Canadian Journal of Forest Research* **38**(5): 1095-1109.

Goerndt, M.E.; Monleon, V.J.; Temesgen, H. 2010. Relating forest attributes with area- and tree-based light detection and ranging metrics for western Oregon. *Western Journal of Applied Forestry* **25**(3): 105-111.

Golinkoff, J.; Hanus, M.; Carah, J. 2011. The use of airborne laser scanning to develop a pixel-based stratification for a verified carbon offset project. *Carbon Balance and Management* **6-9**(16 Nov): 1-17.

Haapanen, R.; Lehtinen, K.; Miettinen, J.; Bauer, M.E.; Ek, A.R. 2002. Progress in adapting k-NN methods for forest mapping and estimation using the new annual Forest Inventory and Analysis data. *Third Annual Forest Inventory and Analysis Symposium. Gen. Tech. Rep. NC-230.* R. E. McRoberts, G. A. Reams, P. C. Van Deusen and J. W. Moser (editors), USDA Forest Service, North Central Research Station, St. Paul, MN**:** 87-95.

Hahn, J.T. 1984. Tree volume and biomass equations for the Lake States. USDA, Forest Service, North Central Forest Experiment Station, St. Paul, MN. Research Paper NC-250.

Hall, F.H.; Bergen, K.; Blair, J.B.; Dubayah, R.; Houghton, R.; Hurtt, G.; Kellndorfer, J.; Lefsky, M.; Ranson, J.; Saatchi, S.; Shugart, H.H.; Wickland, D. 2011. Characterizing 3D vegetation structure from space: Mission requirements. *Remote Sensing of Environment* **115**(11): 2753-2775.

Hall, R.J.; Skakun, R.S.; Arsenault, E.J.; Case, B.S. 2006. Modeling forest stand structure attributes using Landsat ETM+ data: Application to mapping of aboveground biomass and stand volume. *Forest Ecology and Management* **225**(1–3): 378-390.

Hapfelmeier, A.; Ulm, K. 2013. A new variable selection approach using Random Forests. *Computational Statistics & Data Analysis* **60**: 50-69.

Hoare, A.; Regan, D.G.; Wilson, D.P. 2008. Sampling and sensitivity analyses tools (SaSAT) for computational modelling. *Theoretical Biology and Medical Modelling* **5:4**: 1-18. doi:10.1186/1742-4682-1185-1184.

Holdaway, M.R. 1984. Modeling the effect of competition on tree diameter growth as applied in stems. Gen. Tech. Rep. NC-94. USDA Forest Service, North Central Forest Experiment Station, St. Paul, MN 55108**:** 9p.

Holdaway, M.R.; Brand, G.J. 1983. An evaluation of STEMS tree growth projection system. Res. Pap. NC-234. St. Paul, MN. USDA Forest Service, North Central Forest Experiment Station. 22-26.

Hollaus, M.; Wagner, W.; Maier, B.; Schadauer, K. 2007. Airborne laser scanning fore forest stem volume in a mountainous environment. *Sensors* **7**: 1559-1577.

Hollaus, M.; Wagner, W.; Schadauer, K.; Maier, B.; Gabler, K. 2009. Growing stock estimation for alpine forests in Austria: a robust lidar-based approach. *Can. J. For. Res.* **39**: 1387-1400.

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

Hudak, A.T.; Crookston, N.L.; Evans, J.S.; Falkowski, M.J.; Smith, A.M.S.; Gessler, P.E.; Morgan, P. 2006. Regression modelling and mapping of coniferous forest basal area and tree density from discrete-return LiDAR and multispectral satellite data. *Canadian Journal of Forest Research* **32**(2): 126-138.

Hudak, A.T.; Crookston, N.L.; Evans, J.S.; Hall, D.E.; Falkowski, M.J. 2008. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. *Remote Sensing of Environment* **112**(5): 2232-2245.

Hudak, A.T.; Evans, J.S.; Smith, A.M.S. 2009. LiDAR utility for natural resource managers. *Remote Sensing* **1**: 934-951.

Hudak, A.T.; Strand, E.K.; Vierling, L.A.; Byrne, J.C.; Eitel, J.U.H.; Martinuzzi, S.; Falkowski, M.J. 2012. Quantifying aboveground forest carbon pools and fluxes from repeat LiDAR surveys. *Remote Sensing of Environment* **123**: 25-40.

Huete, A.; Didan, K.; Miura, T.; Rodriguez, E.P.; Gao, X.; Ferreira, L.G. 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment* **83**(1-2): 195-231.

Hummel, S.; Hudak, A.T.; Uebler, E.H.; Falkowski, M.J.; Megown, K.A. 2011. A comparision of accuracy and cost of LiDAR versus stand exam data for landscape management of the Malheur national forest. *Journal of Forestry* **July/Aug**: 267-273.

IFMAP. 2001. Review of Remote Sensing Technologies for the IFMAP Project. Michigan Department of Natural Resources, Integrated Forest Monitoring Assessment and Prescription Project. [Online] http://www.dnr.state.mi.us/spatialdatalibrary/sdl2/land_use_cover/2001/IFMAP_1p_landcover.htm (accessed on April 12, 2014).

Jenkins, J.C.; Chojnacky, D.C.; Heath, L.S.; Birdsey, R.A. 2003. National-scale biomass estimators for United States tree species. *Forest Science* **49**(1): 12-35.

Jensen, J.R. 2005. Introductory digital image processing: a remote sensing perspective (3rd edition), Prentice Hall series in geographic information science.

Jochem, A.; Hollaus, M.; Rutzinger, M.; Hofle, B. 2011. Estimation of aboveground biomass in alpine forests: A semi-empirical approach considering canopy transparency derived from airborne LiDAR data. *Sensors* **11**: 278-295.

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

Kellndorfer, J.; Walker, W.; Pierce, L.; Dobson, C.; Fites, J.A.; Hunsaker, C.; Vona, J.; Clutter, M. 2004. Vegetation height estimation from Shuttle Radar Topography Mission and National Elevation Datasets. *Remote Sensing of Environment* **93**(3): 339-358.

Kellndorfer, J.; Walker, W.S.; LaPoint, E.; Bishop, J.; Cormier, T.; Fiske, G.; Hoppus, M.L.; Kirsch, K.; Westfall, J. 2012. NCAP Aboveground biomass and carbon baseline dataset (NBCD 2000). [Online] http://daac.ornl.gov from ORNL DAAC, Oak Ridge, Tennessee, U.S.A. http://dx.doi.org/10.3334/ORNLDAAC/1081 (accessed on June 2, 2014).

Klinka, K.; Carter, R.E. 1990. Relationships between site index and synoptic environmental factors in immature coastal Douglas-fir stands. *Forest Science* **36**(3): 815-830.

Koch, B. 2010. Status and future of laser scanning, synthetic aperture radar and hyperspectral remote sensing data for forest biomass assessment. *Journal of Photogrammetry and Remote Sensing* **65**(6): 581-590.

Kronsender, K.; Ballhorn, U.; Bohm, V.; Siegert, F. 2012. Above ground biomass estimation across forest types at different degradation levels in Central Kalimantan using LiDAR data. *International Journal of Applied Earth Observation and Geoinformation* **18**: 37-48.

Labrecque, S.; Fournier, R.A.; Luther, J.E.; Piercey, D. 2006. A comparison of four methods to map biomass from Landsat-TM and inventory data in western Newfoundland. *Forest Ecology and Management* **226**(1–3): 129-144.

Lacerte, V.; Larocque, G.R.; Woods, M.; Parton, W.J.; Penner, M. 2004. Testing the Lake States variant of FVS (Forest Vegetation Simulator) for the main forest types of northern Ontario. *The Forestry Chronicle* **80**(4): 495-506.

Laes, D.; Reutebuch, S.E.; McGaughey, R.J.; Mitchell, B. 2011. Guidelines to estimate forest inventory parameters from lidar and field plot data. USDA. Forest Service. Companion document to the Advanced Lidar Applications.

Latifi, H.; Nothdurft, A.; Koch, B. 2010. Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: application of multiple optical/LiDAR-derived predictors. *Forestry* **83**(4): 395-407.

Le Maire, G.; Marsden, C.; Nouvellon, Y.; Grinand, C.; Hakamada, R.; Stape, J.; Laclau, P. 2011. MODIS NDVI time-series allow the monitoring of Eucalyptus plantation biomass. *Remote Sensing of Environment* **115**(10): 2613-2625.

Leary, R.A. 1997. Testing models of unthinned red pine plantation dynamics using a modified Bakuzis matrix of stand properties. *Ecological Modelling* **98**: 35-46.

Lefsky, M.A.; Cohen, W.B.; Parker, G.G.; Harding, D.J. 2002. LiDAR remote sensing for ecosystem studies. *BioScience* **52**(1): 19-30.

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): 109-119.

Lessard, V.C.; McRoberts, R.E.; Holdaway, M.R. 2001. Diameter growth models using Minnesota Forest Inventory and Analysis data. *Forest Science* **47**(3): 301-310.

Liaw, L.A.; Wiener, M. 2002. Classification and regression by randomForest. *R News* **2**(3): 18-22.

LP DAAC. 2013. NASA Land Processes Distributed Active Archive Centre. ASTER L1B. Technical report, USGS/Earth Resources Observation and Science (EROS) Centre, Sioux Falls, South Dakota.

Lu, D. 2006. The potential and challenge of remote sensing-based biomass estimation. *International Journal of Remote Sensing* **27**(7): 1297-1328.

Lu, D.; Chen, Q.; Wang, G.; Moran, E.; Batistella, M.; Zhang, M.; Laurin, G.V.; Saah, D. 2012. Aboveground forest biomass estimation with Landsat and LiDAR data and uncertainity analysis of the estimates. *International Journal of Forestry Research*(436537): 1-16.

Lumley, T. 2009. Thomas Lumley using Fortran code by Alan Miller @ leaps: regression subset selection. R package version 2.9. [Online] http://CRAN.R-project.org/package=leaps (accessed on June 2, 2014).

Luther, J.E.; Fournier, R.A.; Piercey, D.; Guindon, L.; Hall, R.J. 2006. Biomass mapping using forest type and structure derived from Landsat TM imagery. *International Journal of Applied Earth Observation and Geoinformation* **8**(3): 173-187.

Main-Knorn, M.; Moisen, G.G.; Healey, S.P.; Keeton, W.S.; Freeman, E.A.; Hostert, P. 2011. Evaluating the remote sensing and inventory based estimation of biomass in the western carpathians. *Remote Sensing* **3**: 1427-1446.

Martin, G.L. 1983. The relative efficiency of some forest growth estimators *Biometrics* **39**(3): 639-650.

Matern, B. 1972. The precision of basal area estimates. *Forest Science* **18**(2): 123-125.

McGaughey, R.J. 2014. FUSION/LDV: Software for LiDAR data analysis and visualization, version 3.21. USDA, Forest Service, Pacific Northwest Research Station, University of Washington, Seattle, WA 98195-2100. [Online] http://forsys.cfr.washington.edu/fusion/FUSION_manual.pdf (accessed on March 30, 2014).

McRoberts, R.E. 2009. Diagnostic tools for nearest neighbors techniques when used with satellite imagery. *Remote Sensing of Environment* **113**: 489-499.

McRoberts, R.E. 2012. Estimating forest attribute parameters for small areas using nearest neighbors techniques. *Forest Ecology and Management* **272**: 3-12.

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

McRoberts, R.E.; Tomppo, E.; 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.

MGET. 2013. Marine Geospatial Ecology Tools. Open source geoprocessing for marine research and conservation. Duke| Nicholas School of the Environment. [Online] http://mgel.env.duke.edu/mget/ (accessed on June 16, 2013).

Miles, P.D.; Hill, A.D. 2010. Volume equations for the northern research station's forest inventory and analysis program as of 2010. USDA, Forest Service, Northern Research Station. General Technical Report NRS-74.

Miner, C.L.; Walters, N.R.; Belli, M.L. 1988. A guide to the TWIGS Program for the North Central United States. *Gen. Tech. Rep. NC-125*, USDA Forest Service, North Central Forest Experimental Station, St. Paul, MN.

Mitchell, J.J.; Glenn, N.F.; Sankey, T.T.; Derryberry, D.R.; Anderson, M.O.; Hruska, R.C. 2011. Small-footprint lidar estimations of sagebrush canopy characteristics. *Photogrammetric Engineering and Remote Sensing* **77**(5): 1-10.

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**: 337-359.

Monserud, R.A.; Huang, S.; Yang, Y. 2006. Predicting lodgepole pine site index from climatic parameters in Alberta. *The Forestry Chronicle* **82**(4): 562-571.

Monserud, R.A.; Sterba, H. 1996. A basal area increment model for individual trees growing in even- and uneven-aged forest stands in Austria. *Forest Ecology and Management* **80**(1-3): 57-80.

Monserud, R.A.; Yang, Y.; Huang, S.; Tchebakova, N. 2008. Potential change in lodgepole pine site index and distribution under climatic change in Alberta. *Canadian Journal of Forest Research* **38**(2): 343-352. doi:310.1139/X1107-1166.

Muukkonen, P.; Heiskanen, J. 2007. Biomass estimation over a large area based on standwise forest inventory data and ASTER and MODIS satellite data: A possibility to verify carbon inventories. *Remote Sensing of Environment* **107**(4): 617-624.

Næsset, E.; Gobakken, T. 2008. Estimation of above- and below- ground biomass across regions of the boreal forest zone using airborne laser. *Remote Sensing of Environment* **112**(6): 3079-3090.

Næsset, E.; Gobakken, T.; Solberg, S.; Gregoire, T.G.; Nelson, R.; Ståhl, G.; Weydahl, D. 2011. Model-assisted regional forest biomass estimation using LiDAR and InSAR as auxiliary data: A case study from a boreal forest area. *Remote Sensing of Environment* **115**(12): 5539-3614.

Nelson, M.D.; Healey, S.P.; Moser, W.K.; Maser, J.G.; Cohen, W.B. 2011. Consistency of forest presence and biomass predictions modeled across overlapping spatial and temporal extents. *Mathematical and Computational Forestry & Natural-Resource Science* **3**(2): 102-113.

Nelson, R.; Gobakken, T.; Naesset, E.; Gregoire, T.G.; Stahl, G.; Holm, S.; Flewelling, J. 2012. Lidar sampling- Using an airborne profiler to estimate forest biomass in Hedmark County, Norway. *Remote Sensing of Environment* **123**: 563-578.

Nilsson, M. 1996. Estimation of tree heights and stand volume using an airborne lidar system. *Remote Sensing of Environment* **56**(1): 1-7.

Nilsson, M. 2002. Deriving nationwide estimates of forest variables for Sweden using Landsat ETM+ and field data. *ForestSAT2002 Symposium*. Heriot Watt University, Edinburgh.

NRCS. 2011. Michigan technical note: Conducting a forest inventory. Forestry # 29. USDA-Natural Resouces Conservation Service

NRCS. 2013. Description of SSURGO database. USDA Natural Resources Conservation Service. [Online] http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/ref/?cid=nrcs142p2_053631 (assessed on April 24, 2014).

NRIS. 2014. Common stand exam user guide: Chapter 2- preparation and design. U.S. Forest Service, Natural Resource Information System. Washington, DC 20250-0003. [Online] http://www.fs.fed.us/nrm/fsveg/index.shtml (accessed on April 1, 2014).

O'Connell, B.; LaPoint, E.; Turner, J.; Ridley, T.; Boyer, D.; Wilson, A.; Waddell, K.; Conkling, B. 2013. The Forest Inventory and Analysis Database: Database description and users' manual version 5.1.5 for Phase 2. USDA, Forest Service, Rocky Mountain Research Station. General Technical Report.

Ohmann, J.L.; Gregory, M.J. 2002. Predictive mapping of forest composition and structure with direct gradient analysis and nearestneighbor imputation in coastal Oregon, U.S.A. *Canadian Journal of Forest Research* **32**: 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): 660-676.

Pesonen, A.; Kangas, A.; Maltamo, M.; Packalen, P. 2010. Effect of auxiliary data source and inventory unit size on the efficiency of sample-based coarse woody debris inventory. *Forest Ecology and Management* **259**(10): 1890-1899.

Pokharel, B.; Froese, R.E. 2009. Representing site productivity in the basal area increment model for FVS-Ontario. *Forest Ecology and Management* **258**(5): 657-666.

Pond, N.C. 2012. Evaluating northern hardwood management using retrospective analysis and diameter distributions. *Ph.D. dissertation*, Michigan Technological University. School of Forest Resources and Environmental Science. Houghton.

Pond, N.C.; Froese, R.E.; Deo, R.K.; Falkowski, M.J. 2014. Multiscale validation of an operational model of forest inventory attributes developed with contrained remote sensing data. *Canadian Journal of Remote Sensing* **40**(1): 43-59.

Popescu, S.C. 2007. Estimating biomass of individual pine trees using airborne LiDAR. *Biomass and Energy* **31**: 646-655.

Popescu, S.C.; Wynne, R.H.; Nelson, R.F. 2003. Measuring individual tree crown diameter with lidar and assessing its influence on estimating forest volume and biomass. *Canadian Journal of Remote Sensing* **29**(5): 564-577.

Popescu, S.C.; Zhao, K.; Neuenschwander, A.; Lin, C. 2011. Satellite lidar *vs.* small-fooprint airborne lidar: Comparing the accuracy of aboveground biomass estimates and forest structure metrics at footprint level. *Remote Sensing of Environment* **115**: 2786-2797.

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 comparison of empirical modelling approaches. *Remote Sensing of Environment* **114**(5): 1053-1068.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ranson, K.J.; Sun, G.; Lang, R.; Chauhan, N.; Cacciola, R.; Kilic, O. 1997. Mapping of boreal forest biomass from spaceborne synthetic aperture radar. *Journal of Geophysical Research* **102**(D24): 29,599-529,610.

Reed, D.R.; Mroz, G.D. 1997. Resource assessment in forested landscapes. John Wiley & Sons, Inc.

Reese, H.; Nilsson, M.; Sandstrom, P.; 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.

Rehfeldt, G.E.; Crookston, N.L.; Warwell, M.V.; Evans, J.S. 2006. Empirical analysis of plant-climate relationships for the western United States. *International Journal of Plant Science* **167**(6): 1123-1150.

RMRS. 2013. Research on Forest Climate Change: Potential Effects of Global Warming on Forests and Plant Climate Relationships in Western North America and Mexico. USDA Forest Service - Rocky Mountain Research Station (RMRS) - Moscow Forestry Sciences Laboratory. [Online] http://forest.moscowfsl.wsu.edu/climate/ (accessed on April 24, 2014).

Roberts, J.J.; Best, B.D.; Dunn, D.C.; Treml, E.A.; Halpin, P.N. 2010. Marine Geospatial Ecology Tools: An integrated framework for ecological geoprocessing with ArcGIS, Python, R, MATLAB, and C++. *Environmental Modelling & Software* **25**(10): 1197-1207.

Robinson, A.P.; Duursma, R.A.; Marshall, J.D. 2005. A regression-based equivalence test for model validation: shifting the burden of proof. *Tree Physiology* **25**: 903-913.

Robinson, A.P.; Froese, R.E. 2004. Model validation using equivalence tests. *Ecological Modelling* **176**: 349-358.

Robinson, A.P.; Hamann, J.D. 2011. Forest analytics with R: An introduction. New York Dordrecht Heidelberg London, Springer.

Saatchi, S.S.; Houghton, R.A.; Alvala, R.C.D.S.; Soares, J.V.; Yu, Y. 2007. Distribution of aboveground live biomass in the Amazon basin. *Global Change Biology* **13**: 816-837.

Sakamoto, Y.; Ishiguro, M.; Kitagawa, G. 1986. Akaike Information Criterion Statistics. D. Reidel Publishing Company.

Salas, C.; Ene, L.; Gregoire, T.G.; Naesset, E.; Gobakken, T. 2010. Modeling tree diameter from airborne laser scanning derived variables: a comparison of spatial statistical models. *Remote Sensing of Environment* **114**(6): 1277-1285.

Schaetzl, R.J.; Krist, F.J.J.; Miller, B.A. 2012. A Taxonomically Based, Ordinal Estimate of Soil Productivity for Landscape-Scale Analyses. *Soil Science* **177**: 288-299.

Schaetzl, R.J.; Krist, F.J.J.; Stanley, K.E.; Hupy, C.M. 2009. The natural soil drainage index: An ordinal estimate of long-term soil wetness. *Physical Geography* **30**: 383-409.

Schreuder, H.T.; G, G.T.; B, W.G. 1993. Sampling methods for multiresource forest inventory. New York, John Wiley and Sons, Inc.

Scott, C.T. 1990. An overview of fixed versus variable-radius plots for successive inventories. *State-of-the-art methodology of forest inventory*. V. LaBau and T. Cunia. Randor, PA 19087. USDA Forest Service, Pacific Northwest Research Station, Portland, Oregon. General Technical Report 263.

Sharma, R.P.; Brunner, A.; Eid, T. 2012. Site index prediction from site and climate variables for Norway spruce and Scots pine in Norway. *Scandinavian Journal of Forest Research* **27**: 619-636.

Shifley, S.R. 1987. A generalized system of models forecasting Central States tree growth. Research Paper NC-279. US Forest Service, North Central Forest Experiment Station, St. Paul, MN. 10 p.

Skovsgaard, J.P.; Vanclay, J.K. 2008. Forest site productivity: a review of the evolution of dendrometric concepts for even-aged stands. *Forestry* **81**(1): 13-31.

Skovsgaard, J.P.; Vanclay, J.K. 2013. Forest site productivity: a review of spatial and temporal variability in natural site conditions. *Forestry* **86**: 305-315.

Solberg, S.; Brunner, A.; Hanssen, K.H.; Lange, H.; Naesset, E.; Rautiainen, M.; Stenberg, P. 2009. Mapping LAI in a Norway spruce forest using airborne laser scanning. *Remote Sensing of Environment* **113**(11): 2317-2327.

Song, C. 2012. Optical remote sensing of forest leaf area index and biomass. *Progress in Physical Geography.* pp.1-16. [Online] http://ppg.sagepub.com/content/early/2012/12/21/0309133312471367 (accessed on May 25, 2014)

Stage, A.R. 1973. Prognosis model for stand development *Res. Pap. INT-137*, USDA Forest Service**:** 32p.

Stage, A.R.; Moore, J.C.; Renner, D.L. 2001. Modeling silvicultural options in the context of uncertain climate: using the forest vegetation simulator and its fire and fuels extension. *J. Math. Model. Sci. Comp.* **13**(3-4): 249-259.

Steininger, M.K. 2000. Satellite estimation of tropical secondary forest above-ground biomass: data from Brazil and Bolivia. *International Journal of Remote Sensing* **21**(6-7): 1139-1157.

Straub, C.; Dees, M.; Weinacker, H.; Koch, B. 2009. Using airborne laser scanner data and CIR orthophotos to estimate the stem volume of forest stands. *Photogrammetrie* **3**: 277-287.

Straub, C.; Koch, B. 2011. Enhancement of bioenergy estimations within forests using airborne laser scanning and multispectral line scanner data. *Biomass and Energy* **35**(8): 3561-3574.

Sun, G.; Ranon, K.J.; Guo, Z.; Zhang, Z.; Montesano, P.; Kimes, D. 2011. Forest biomass mapping from LiDAR and radar synergies. *Remote Sensing of Environment* **115**(11): 2906-2916.

Teck, R.M.; Hilt, D.E. 1991. Individual tree diameter growth model for the Northeastern United States. Research Paper NE-649. US Forest Service, Northeastern Forest Experiment Station, Radnor, Pennsylvania.

Tesfamichael, S.G.; van Aardt, J.A.N.; Ahmed, F. 2010. Estimating plot-level tree height and volume of Eucalyptus grandis plantations using small-footprint, discrete return lidar data. *Progress in Physical Geography* **34**(4): 515-540.

Tomppo, E. 1991. Satellite image based national forest inventory of Finland. *International Archives of Photogrammetry and Remote Sensing* **27**(7.1): 419-424.

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

Tomppo, E.; Katila, M. 1991. Satellite image-based national forest inventory of finland. Geoscience and Remote Sensing Symposium 1991 (IGARSS '91), Espoo, Finland.

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

Tomppo, E.; Olsson, H.; Ståhl, G.; Nilsson, M.; Hagner, O.; Katila, M. 2008. Combining national forest inventory field plots and remote sensing data for forest databases. *Remote Sensing of Environment* **112**(5): 1982-1999.

Tuominen, S.; Eerikainen, K.; Schibalski, A.; Haakana, M.; Lehtonen, A. 2010. Mapping biomass variables with a multi-source forest inventory techniques. *Silva Fennica* **44**(1): 109-119.

Tuominen, S.; Pekkarinen, A. 2005. Performance of different spectral and textural aerial photograph features in multi-source forest inventory. *Remote Sensing of Environment* **94**(2): 256-268.

UN-REDD. 2010. The UN-REDD programme strategy 2011-2015. The United Nations Collaborative Programme on Reducing Emissions from Deforestation and Forest Degradation in Developing Countries. Geneva, Switzerland.

UNFCCC. 1992. United Nations Framework Convention on Climate Change. United Nations. New York.

USDA Forest Service. 2013a. Soil drainage and productivity index. [Online] http://foresthealth.fs.usda.gov/soils/ (assessed on April 24, 2014).

USFS. 2000. Timber cruising handbook. USDA Forest Service. Washington. FSH 2409.12. .

USFS. 2013. Forest inventory and analysis national program – data and tools - FIA data mart, FIADB Version 5.1. USDA Forest Service. [Online] http://apps.fs.fed.us/fiadb-downloads/datamart.html (31 July 2013).

Valentine, H.T. 1997. Height growth, site index, and Carbon metabolism. *Silva Fennica* **31**(3): 251-263.

Van Aardt, J.A.N.; Wynne, R.H.; Oderwald, R.G. 2006. Forest volume and biomass estimation using small-footprint lidar-distributional parameters on a per-segment basis. *Forest Science* **52**(6): 636-649.

Van Deusen, P.C. 1997. Annual forest inventory statistical concepts with emphasis on multiple imputation. *Canadian Journal of Forest Research* **27**: 379-384.

Vanclay, J.K. 1994. Modelling forest growth and yield: applications to mixed tropical forests, CAB International, Wallingford, U.K.

Vauhkonen, J.; Korpela, I.; Maltamo, M.; Tokola, T. 2010. Imputation of single-tree attributes using airborne laser scanning-based height, intensity, and alpha shape metrics. *Remote Sensing of Environment* **114**(6): 1263-1276.

Venables, W.N.; Ripley, B.D. 2002. Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

Vincent, G.; Sabatier, D.; Blanc, L.; Chave, J.; Weissenbacher, E.; Pelissier, R.; Fonty, E.; Molino, J.F.; Couteron, P. 2012. Accuracy of small footprint airborne LiDAR in its predictions of tropical moist forest stand structure. *Remote Sensing of Environment* **125**: 23-33.

Walker, W.S.; Kellndorfer, J.M.; LaPoint, E.; Hoppus, M.L.; Westfall, J. 2007. An empirical InSar-optical fusion approach to mapping vegetation canopy height. *Remote Sensing of Environment* **109**(4): 482-499.

Waring, R.H.; Coops, N.C.; Landsberg, J.J. 2010. Improving predictions of forest growth using the 3-PGS model with observations made by remote sensing. *Forest Ecology and Management* **259**(9): 1722-1729.

Waske, B.; Benediktsson, J.A.; Sveinsson, J.R. 2012. Random forest classification of remote sensing data. *Signal and Image Processing for Remote Sensing*. C. H. Chen, CRC Press, Taylor & Francis Group**:** 365-374.

Weiskittel, A.R.; Crookston, N.L.; Radtke, P.J. 2011. Linking climate, gross primary productivity, and site index across forests of the western Unites States. *Canadian Journal of Forest Research* **41**: 1710-1721.

West, P.W. 1980. Use of diameter increment and basal increment in tree growth studies. *Can. J. For. Res.* **10**(1): 71-77.

White, E.M. 2010. Woody biomass for bioenergy and biofuels in the United States: a briefing paper, USDA Forest Service, Portland, OR.

White, J.C.; Wulder, M.A.; Varhola, A.; Vastaranta, M.; Coops, N.C.; Cook, B.D.; Pitt, D.; Woods, M. 2013. A best practice guide for generating forest inventory attributes from airborne laser scanning data using an area-based aprroach (Version 2.0), Natural Resources Canada/ Canadian Forest Service/ Canadian Wood Fibre Centre. 506 West Burnside Road, Victoria, British Columbia.

Woodall, C.W.; Heath, L.S.; Domke, G.M.; Nichols, M.C. 2010. Methods and equations for estimating aboveground volume, biomas, and carbon for trees in the US forest inventory. USDA, Forest Service, Northern Research Station. General Technical Report NRS-88.

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*, USDA, Forest Service, Rocky Mountain Research Station, Ft. Collins, CO**:** 339p.

Wulder, M.A.; Bater, C.W.; Coops, N.C.; Hilker, T.; White, J.C. 2008. The role of LiDAR in sustainable forest management. *The Forestry Chronicle* **84**(6): 807-826.

Wulder, M.A.; White, E.M.; Nelson, R.F.; Naesset, E.; Orka, H.O.; Coops, N.C.; Hilker, T.; Bater, C.W.; Gobakken, T. 2012. Lidar sampling for large-area forest characterization: A review. *Remote Sensing of Environment* **121**: 196-209.

Wulder, M.A.; White, J.C.; Bater, C.W.; Coops, N.C.; Hopkinson, C.; Chen, G. 2012a. Lidar plots- a new large-area data collection option: context, concepts, and case study. *Canadian Journal of Remote Sensing* **38**(5): 600-618.

Wulder, M.A.; White, J.C.; Fournier, R.A.; Luther, J.E.; Magnussen, S. 2008a. Spatially explicit large area biomass estimation: Three approaches using forest inventory and remotely sensed imagery in a GIS. *Sensors* **8**: 529-560.

Wykoff, W.R. 1990. A basal area increment model for individual conifers in the Northern Rocky Mountains. *Forest Science* **36**(4): 1077-1104.

Wykoff, W.R.; Crookston, N.L.; Stage, A.R. 1982. Users' guide to the Stand Prognosis Model. *Gen. Tech. Rep. INT-122*, USDA, Forest Service, Intermountain Forest and Range Experiment Station, Ogden, UT 84401**:** 112p.

Yim, J.S.; Kim, Y.H.; Kim, S.H.; Jeong, J.H.; Shin, M.Y. 2011. Comparison of the k-nearest neighbor technique with geographical calibration for estimating forest growing stock volumeThis article is one of a selection of papers from Extending Forest Inventory and Monitoring over Space and Time. *Canadian Journal of Forest Research* **41**(1): 73-82.

Zeide, B. 1993. Analysis of growth equations. *Forest Science* **39**(3): 594-616.

Zhang, X.; Ni-meister, W. 2014. Remote sensing of forest biomass. *Remote Sensing/Photogrammetry*. J. M. Hanes (ed.). Verlag Berlin Heidelberg, Springer.

Zhao, D.; Borders, B.; Wilson, M. 2004. Individual-tree diameter growth and mortality models for bottomland mixed-species hardwood stands in the lower Mississippi alluvial valley. *Forest Ecology and Management* **199**(2-3): 307-322.

Zhao, K.; Popescu, S.C.; Nelson, R. 2009. LiDAR remote sensing of forest biomass: A scale-invariant estimation approach using airborne lasers. *Remote Sensing of Environment* **113**(1): 182-196.

Zheng, D.; Heath, L.S.; Ducey, M.J. 2007. Forest biomass estimated from MODIS and FIA data in the Lake States: MN, WI and MI, USA. *Forestry* **80**(3): 265-278.