AN ABSTRACT OF THE DISSERTATION OF

Bianca N.I. Eskelson for the degree of Doctor of Philosophy in Forest Resources presented on November 21, 2008.
Title: Examination of Imputation Methods to Estimate Status and Change of Forest Attributes from Paneled Inventory Data.

Abstract approved:

_____

Temesgen Hailemariam

The Forest Inventory and Analysis (FIA) program conducts an annual inventory throughout the United States. In the western United States, 10% of all plots (one panel) are measured annually, and a moving average is used for estimating current condition and change of forest attributes while alternative methods are sought in all regions of the United States.

This dissertation explored alternatives to the moving average in the Pacific Northwest using Current Vegetation Survey data collected in Oregon and Washington. Several nearest neighbor imputation methods were examined for their suitability to update plot-level forest attributes (basal area/ha, stems/ha, volume/ha, biomass/ha) to the current point in time. The results were compared to estimates obtained using a moving average and a weighted moving average. In terms of bias and accuracy, the weighted moving average performed better than the moving average. When the most recent measurements of the variables of interest were used as ancillary data, randomForest imputation outperformed both the moving average and the weighted moving average.

For estimating current basal area/ha, stems/ha, volume/ha, and biomass/ha, tree-level imputation outperformed plot-level imputation. The difference in bias and accuracy between tree- and plot-level imputation was more pronounced when the variables of interest were summarized by species groups.

Nearest neighbor imputation methods were also investigated for estimating mean annual change in selected forest attributes. The imputed mean annual change was used to update unmeasured panels to the current point in time. In terms of bias and accuracy, the resulting estimates of current basal area/ha, stems/ha, volume/ha, and biomass/ha outperformed the results obtained using plot-level imputation.

Information on hard to estimate forest attributes such as cavity tree and snag abundance are important for wildlife management plans. Using FIA data collected in Washington, Oregon, and California, nearest neighbor imputation approaches and negative binomial regression models were examined for their suitability in estimating cavity tree and snag abundance. The negative binomial models were preferred to the nearest neighbor imputation approaches.

Examination of Imputation Methods to Estimate Status and Change of Forest Attributes from Paneled Inventory Data

by
Bianca N.I. Eskelson

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented November, 21 2008
Commencement June 2009

Doctor of Philosophy dissertation of Bianca N.I. Eskelson presented on November, 21 2008.

APPROVED:

_____

Major Professor, representing Forest Resources

_____

Head of the Department of Forest Engineering, Resources and Management

_____

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

_____

Bianca N.I. Eskelson, Author

ACKNOWLEDGEMENTS

The task is almost done and I have the chance to look back on years full of new experiences, rewarding work, innumerable challenges, and many achievements. This would not be complete without appreciating the outstanding people who have accompanied me along the way.

I offer my most sincere gratitude to Dr. Temesgen Hailemariam, who did an outstanding job as my advisor. Without his careful guidance, his constant support, and his unfailing confidence in me, this work would not have been possible. I also extend my greatest thanks to my doctoral committee composed of Drs. Tara Barrett, David Hann, Daniel Schafer, and Steven Radosevich, who have answered a never ending flow of questions about FIA and CVS data, general modeling issues, and statistics. Thank you all for your help, guidance, patience, and encouragement.

Special thanks go to Carol Apple, Jim Alegria, Bob Brown, and Melinda Moeur for providing help in obtaining national forest data, to Kurt Campbell for assistance with volume and biomass equations, to Matt Gregory and Janet Ohmann for sharing their ancillary data with me, and to Greg Johnson and Dave Marshall for inviting me to Federal Way and providing me with crucial feedback on some of my work.

Many thanks go to my fellow Forest Biometrics students and fellow basement-dwellers for countless lunch and tea breaks as well as discussions about forest biometrics and life in general.

I am most grateful for the love and support of my family and friends at home who have encouraged me in all my decisions and who I could turn to whenever I needed advice or simply someone to talk to.

Finally, I want to thank Emiliano and the Corvallis judo community, who have become my Corvallis family over the past years and who have provided me with friendship and a way of life that has shaped me more than anything else. Thank you for your faith in me!

This research was supported by funding from the US Forest Service, Pacific Northwest Research Station, Forest Inventory and Analysis program. The views described here are those of the author alone and do not represent those of the US Forest Service.

CONTRIBUTION OF AUTHORS

Drs. Tara Barrett and Temesgen Hailemariam provided extensive comments, professional expertise, and financial support for chapters 2, 3, 4, and 5.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF TABLES

LIST OF TABLES (Continued)

EXAMINATION OF IMPUTATION METHODS TO ESTIMATE STATUS AND CHANGE OF FOREST ATTRIBUTES FROM PANELED INVENTORY DATA

CHAPTER 1: INTRODUCTION

Information on current forest condition and change is essential to assess and characterize resources and to support resource management and policy decisions. Since the 1930s the Forest Inventory and Analysis (FIA) program of the United States Department of Agriculture (USDA) Forest Service or its predecessor programs have conducted periodic inventories of forest land in the United States and provided data on the extent and condition of forest land, the volume of timber, timber growth, and timber removals (McRoberts 2000). These traditional periodic inventories provide data that can be used to characterize current conditions for only two or three years after each inventory and become less useful for this purpose over time (Reams et al. 1999). Immediate estimation of the effects of catastrophic events (e.g., hurricanes, fire, insect infestations) on the forest resources is usually impossible with the data provided by the periodic inventory design (McRoberts 2000), and the inventory data of bordering states may differ in age by ten or more years, which makes analyses that span multiple states difficult (Gillepsie 1999). These and other deficiencies of the periodic inventory design have led to the blue ribbon panels on FIA in 1991 and 1997 and finally to the Agricultural Research, Extension, and Education Reform Act of 1998 (PL 105-185), known as Farm Bill, which mandates the USDA Forest Service to

conduct annual forest inventories in all states to be able to provide annual updates of each state's forests (McRoberts 2000).

The FIA developed the interpenetrating panel design which is now used for the annual inventory system of the USDA Forest Service (Van Deusen 2000). FIA had seriously considered an annual inventory approach since the early 1990s. In 1992, scientists at the North Central Research Station (NCRS) began to develop the Annual Forest Inventory System (AFIS) which uses satellite data to stratify plots into classes with different probabilities of disturbance. Plots with a high probability of disturbance have a higher probability of being sampled in a given year, while other plots are updated using models (Gillespie 1999). Shortly after the implementation of AFIS, the Southern Research Station (SRS) implemented the Southern Annual Forest Inventory System (SAFIS). This inventory system is similar to AFIS, but all field plots have the same probability of being sampled in a given year. AFIS and SAFIS are forerunners of the interpenetrating panel design that is now used for the annual inventory mandated by the 1998 Farm Bill (Frayer and Furnival 1999).

The interpenetrating panel design is an annual three-phase inventory in which the FIA and Forest Health Monitoring (FHM) plots are merged (Brand et al. 2000, McRoberts 1999). The FHM program had established a nationwide lattice of hexagonal cells as a sampling framework to distribute its sampling plots regularly. The plots were measured annually in a four year cycle (Scott et al. 1993). The FIA hexagons are based on the FHM hexagons, but have an increased sampling intensity by the factor 27, which means that each FIA hexagon is 1/27 the size of an FHM

hexagon (approximately one plot per 2,400 ha). Each hexagon (2,402.7 ha) was systematically assigned to one of the five interpenetrating panels that provide systematic coverage of each state (Reams et al. 2005).

Depending on the region, 10 or 20 percent of the total FIA plots in each state are measured each year. In the eastern states, all plots located in one of the five interpenetrating panels (20%) are measured each year so that a state's inventory is completed in five years. Funding for all western states is for an inventory cycle of 10 years (Brand et al. 2000, McRoberts 1999, Roesch and Reams 1999). The west coast states (California, Oregon, and Washington) have 10% of the plots being measured each year with Alaska and Hawaii having modified inventory systems (Azuma 2000).

In Phase 1 of the inventory, remotely sensed data is used to classify land into forested and non-forested land, and spatial measurements of fragmentation, urbanization, and distance variables are made. Historically, aerial photographs have been used for this phase but the system is changing to methods based on satellite imagery. In Phase 2, data on the permanently established FIA field plots is collected. On accessible forested field plots, information on forest type, site attributes, tree species, tree size, and overall tree condition is collected. Non-forest plots are also visited in order to quantify the rates of land use change. A subset of the Phase 2 plots is visited in Phase 3 (approximately one plot every 38,400 ha) to collect an additional set of data related to forest health conditions. Phase 3 plots are visited during the growing season in order to collect data covering a full vegetation inventory, tree and

crown condition, soil data, lichen diversity, coarse woody debris, and ozone damage (FIA 2005).

The annual inventory harmonizes inventory techniques in assessing and monitoring current and future status of forest resources across temporal and spatial scales. This enables consistent comparison and reporting across states and among jurisdictions (Gillespie 1999, Reams et al. 1999, Van Deusen 2000). The annual inventory provides current information and quantifies variations that occur between the periods. With this information, it is possible to estimate annual current forest conditions and change, which are needed for effective policy and forest management decisions (Reams and Van Deusen 1999, Reams et al. 1999). The effect of catastrophic events on the forest resources may be observed sooner with an annual inventory, which is generally more adequate to observe trends and changes on a national scale than a periodic inventory. Most trends and changes would either be missed by the periodic inventory or documented years after their occurrence (Reams et al. 1999). Therefore, an annual inventory is preferred for forests with a high rate of change (Gillespie 1999). In areas where fieldwork does not depend on the seasons, the annual system allows field staff to be stationed in a certain working area. This eliminates the costs of constantly relocating staff and leads to lower training costs for replacements and greater retention of experienced staff. The fact that the fieldwork of the FIA and FHM programs has been merged increases the efficiency and effectiveness of both programs (Gillespie 1999, Van Deusen et al. 1999).

In an annual inventory, the sample size of one year is lower than in a periodic inventory and therefore the precision of the estimations in any given year is lower compared to the precision achieved with data of a periodic inventory (Van Deusen 2000). When change is small, detection may take up to 20 years in western states. The new inventory design requires the development of new software for data management and analysis (Van Deusen 2000), which produces costs. The plots in one panel are distributed systematically in a state which requires longer travel times between the plots so that travel costs increase (Gillespie 1999, Van Deusen 2000). Where data collection is seasonal, field crews must travel more throughout the season, leading to higher employee turnover and greater training costs.

The interpenetrating panel design is a rigid inventory system. This causes problems if the inventory budgets are uncertain and the measurements of one panel cannot be completed in one year. This is called "panel creep" and could be avoided by creating extra panels that allow some flexibility in measurement (Van Deusen 2000, Van Deusen 2002a). The rigid inventory design also leads to having only 5-year change intervals in the eastern states and 10-year change intervals in the western states. Change estimates for intervals other than five or ten years, respectively, depend on models and assumptions. By rotating the panel assignments, a mix of change intervals could be assured (Van Deusen 2000).

Moving to an annual inventory system implicates significant changes not only in the FIA reporting but also in the estimation of various forest attributes. If only current year data are used for estimations, the results reflect current conditions, but

due to the small sample size, the achieved precision may be unacceptable (McRoberts 2000). Estimation procedures that use data from previously measured panels can significantly improve the precision (Van Deusen 2000).

Currently, FIA uses a moving average (MA) approach, which is operationally convenient and requires a minimum of assumptions (Gartner and Reams 2001), as default estimator. Using data from the panels measured in the most recent years, a MA approach can improve the precision of the estimates, when compared to using only data from plots measured in the current year. However, it rather reflects an average of conditions over the past ten years than current forest conditions (McRoberts 2000), resulting in a bias of the current year's population parameter (Johnson et al. 2003). If a variable of interest indicates a strong trend, the moving average is likely to underestimate the current condition of this variable. If an abrupt shift in the inventory takes place, overestimation might occur (Reams et al. 1999) because the moving average reacts slowly to sudden changes in tree attribute variables (Johnson et al. 2003).

The MA estimator has been accepted as the FIA default estimator for the annual inventory system with the provision that it may be replaced if it can be improved by using updating techniques based on imputation techniques, growth models, or other methods (McRoberts 2000). Estimates based on updated data provide nearly the same precision that was achieved with periodic inventories, provided that updating procedures used are unbiased and sufficiently precise (McRoberts 2001). Different imputation and modeling approaches have been examined for updating data

of unmeasured panels to the current year (Van Deusen 1997, Gartner and Reams 2001, McRoberts 2001). So far, no research has been done in this area in the Pacific Northwest (PNW).

Size and abundance of cavity trees and snags are among the variables measured and reported by FIA. Snags are important structural components of many forest ecosystems (Harmon et al. 1986) and as cavity trees they provide nesting, foraging, and roosting habitat for many wildlife species (Carey 1983, Bate et al. 1999). Knowledge about the size and abundance of cavity trees and snags is important for selecting and modeling wildlife habitat which can support forest planning efforts, regional inventories, and evaluation of different management scenarios. The relationship between cavity tree and snag abundance and stand attributes acquired from paneled data and the suitability and predictive abilities of parametric and nearest neighbor imputation methods have not yet been examined.

The overall goal of this thesis is to explore new methods for estimating current forest condition and change that are designed for the type of data obtained in the annual inventory for the PNW region and compare their results to those from the MA approach. Specific objectives are to (1) examine plot-level nearest neighbor imputation techniques for estimating current plot-level attributes; (2) investigate tree-level nearest neighbor imputation techniques for estimating current plot-level attributes; (3) explore the suitability of nearest neighbor imputation techniques to estimate mean annual change at plot-level; and (4) analyze the suitability of nearest

neighbor imputation methods and negative binomial regression models to estimate cavity tree and snag abundance.

Objectives 1-4 of the dissertation are addressed in Chapters 2-5, respectively. Chapter 2 assesses the use of plot-level imputation techniques. Chapter 3 illustrates the estimation of current forest attributes using tree-level imputation techniques. Chapter 4 explores plot-level imputation methods for estimating mean annual change using annual inventory data. Chapter 5 demonstrates the use of imputation techniques and negative binomial regression models to estimate cavity tree and snag abundance. Finally, Chapter 6 summarizes the key findings of the study and discusses future research that is needed to improve estimation of current status and change based on annual inventory data.

CHAPTER 2: ESTIMATING CURRENT FOREST ATTRIBUTES FROM PANELED INVENTORY DATA USING PLOT-LEVEL IMPUTATION: A STUDY FROM THE PACIFIC NORTHWEST

Bianca N.I. Eskelson

Temesgen Hailemariam

Tara M. Barrett

In press: Forest Science

Abstract

Information on current forest condition is essential to assess and characterize resources and to support resource management and policy decisions. The 1998 Farm Bill mandates the US Forest Service to conduct annual inventories to provide annual updates of each state's forest. In annual inventories, the sample size of one year (panel) is only a portion of the full sample and therefore the precision of the estimations for any given year is low. To achieve higher precision, the Forest Inventory and Analysis program (FIA) uses a moving average (MA), which combines the data of multiple panels, as default estimator. The MA can result in biased estimates of current conditions and alternative methods are sought after. Alternatives to the MA have not yet been explored in the Pacific Northwest. Data from Oregon and Washington national forests were used to examine a weighted moving average (WMA) and three imputation approaches: Most similar neighbor (MSN), gradient nearest neighbor (GNN), and randomForest (RF). Using the most recent measurements of the variables of interest as ancillary variables, RF provided almost unbiased estimates that were comparable to those of the MA and WMA estimators in terms of root mean square error.

Introduction

Initiated by the Agricultural Research, Extension, and Education Reform Act of 1998 (PL 105-185) the Forest Inventory and Analysis (FIA) program of the US Forest Service has switched from periodic inventories that varied from state-to-state to

a consistent nationwide annual inventory. A portion of the inventory of the nation's forests is now conducted annually within each state. The fraction of the plots measured annually is 10% in the western United States and 20% in the eastern United States.

The precision of the estimates of current status and changes in the forest resources using only data from the panel of plots measured in the current year has been found to be unacceptable due to the small annual sample size (McRoberts and Hansen 1999). There have been efforts to combine data of multiple panels in order to achieve a higher precision. The current FIA default estimator is a moving average (MA), which is operationally convenient and requires few assumptions (Gartner and Reams 2001). The MA approach can improve the precision of the estimates by using data from the panels measured in the most recent years. However, MA reflects an average of conditions over the past ten years rather than current forest conditions, resulting in a bias of the current year's population parameter (McRoberts 2000, Johnson et al. 2003). The MA estimates can be improved with a weighted moving average (WMA) which weighs panels that were measured more recently more heavily than those measured earlier (Roesch and Reams 1999). Other approaches to combine data from all panels include: 1) updating unmeasured panel data to the current year using a) growth models (Lessard et al. 2001, McRoberts 2001); b) time series models (Johnson et al. 2003); or c) mixed estimation (Scott et al. 1999, Van Deusen 1996, 1999, 2002b); 2) filling in missing panel data using tree- and plot-level imputation techniques (McRoberts 2001, Gartner and Reams 2001, 2002); or 3) modifying the

annual inventory of interpenetrating, non-overlapping panels to an inventory system with balanced annual partial remeasurements so that estimators based on sampling with partial replacement can be used (Scott et al. 1999, Arner et al. 2004).

There is a need to develop new methods which will be included in the annual inventory system according to their performance (Reams et al. 1999). Since spatial, temporal, and forest characteristics differ within and among regions it is unclear if any single technique will work for all regions (Patterson and Reams 2005), and it is necessary to evaluate different methods in all regions. Studies comparing different alternatives to the MA approach for estimating current forest attributes in the Pacific Northwest (PNW) are still lacking whereas a variety of methods have been tested in the other regions of the United States (McRoberts 2001, Lessard et al. 2001, Van Deusen 1996, 1997, 1999, 2002b; Arner et al. 2004).

The imputation and modeling approaches examined by McRoberts (2001) asserted that model development requires a greater resource investment than development of an imputation procedure. As the difference in the estimation results was negligible, it is reasonable to focus on investigating and improving the imputation techniques. McRoberts (2001) pointed out that development of models might be facilitated as soon as the annual inventory is established for several years and provides calibration data from fixed radius FIA plots at five or ten year intervals. Unlike modeling approaches, imputation techniques require reference data at the application phase. An advantage, however, is that they update themselves when data are added or

removed from the data base (Sironen et al. 2003) and the reference data will increase every year after establishment of the annual inventory.

Depending on the intended use, tree- and plot-level imputation techniques differ in their predictive abilities and suitability (Gartner and Reams 2002). If diameter distributions by species are required, tree-level imputation will be necessary. Therefore, tree-level imputation might be more suitable for complex uneven-aged multi-species stands, where detailed information in the form of tree-lists is needed to describe the stand structure. Only tree-level imputation techniques allow determination of the distribution of individual tree growth and mortality, individual tree size change, and change by species and tree size classes. In a separate study, we are comparing the performance of tree-level and plot-level imputation.

The objectives of this study are to: 1) use paneled data from the PNW to estimate current forest attributes with the FIA default method and compare the MA results with estimates based only on the data from the current panel; and 2) examine three different plot-level imputation methods to fill-in values for the missing panels as well as a WMA and assess their performance against MA.

Methods

Data

The data used in this study consist of 618 plots from six national forests that were collected as part of the Pacific Northwest Regions' Current Vegetation Survey (CVS) of the US Forest Service. The plots were installed between 1993 and 1997 and

remeasured in 2000. The particular national forests sampled were the Colville (28), Mt. Hood (111), Ochoco (82), Rogue River (70), Wallowa-Whitman (199), and Winema (128) (Table 2.1).

Panel data is a special case of inventory data with measurements taken at different times. In order to mimic a panel system with the available data the plots were assigned to the following panels: panel 1 (*P1*) were those measured in 1993 and 1994, panel 2 (*P2*) were those measured in 1995, panel 3 (*P3*) were those measured in 1996 and 1997, and panel 4 (*P4*) were a part of those measured in 2000. All plots were measured in the year 2000 but for the simulations 25% of the plots were randomly assigned to *P4* and the remaining 75% of the plots belong to *P1*, *P2*, and *P3* based on their year of installation. This resulted in *P1*, *P2*, and *P3* having different sizes in each iteration. *P1*, *P2*, and *P3* lack data of the national forests Rogue River, Coleville, and Winema, respectively, since no data was collected in those forests in the corresponding years (Table 2.1).

The basic CVS sampling unit is one hectare (ha) in size. Five plots are installed in each sampling unit with each plot consisting of three permanent circular, nested subplots of different sizes. Which trees are measured in each of the three nested subplots depends upon their diameter at breast height (DBH in cm). Max et al. (1996) provided a detailed description of the inventory. In this study only live trees with DBH of 12.7 cm or larger were used. Missing heights (HT in m) were filled using height models developed in Barrett (2006). Volume and biomass equations from the US Forest Service were used to calculate gross cubic-meter volume and total gross oven

dry weight biomass (USDA 2000). For each plot, basal area in m^2 per ha (BA), stems per ha (SPH), volume in m^3 per ha (VOL), and biomass in tons per ha (BIOT) were calculated and summarized (Table 2.2).

A total of 33 species were present on the plots (Table 2.3). The most frequently encountered species were Douglas-fir (Pseudotsuga menziesii (Mirb.) Franco), ponderosa pine (Pinus ponderosa C. Lawson), grand fir (Abies grandis (Douglas ex D. Don) Lindl.), lodgepole pine (Pinus contorta Douglas ex Louden), white fir (Abies concolor (Gord. & Glend.) Lindl. ex Hildebr.), and western hemlock (Tsuga heterophylla (Raf.) Sarg.), in decreasing order.

Thematic Mapper (TM) images from 2000 were extracted from the national land-cover database 2001 (Homer et al. 2004) and were used as ancillary data. The raw imagery bands 1 to 5 and band 7 (TM1, TM2, TM3, TM4, TM5, TM7) as well as the Tasseled Cap (TC) transformations of the 6 axes (TC1 – TC6) were used. The normalized difference vegetation index (NDVI) and three commonly used band ratios (band 4 to band 3 (R43), band 5 to band 4 (R54), and band 5 to band 7 (R57)) were calculated. Tree canopy cover (CANOPY) was extracted from the national land-cover database 2001 (Homer et al. 2004).

Climate and topography variables were used as another source of ancillary data. Elevation (EL in m) was recorded as part of the CVS inventory. Annual precipitation (ANNPRE) and mean annual temperature (ANNTMP) (Table 2.2) were extracted from DAYMET Daily Surface Weather Data and Climatological Summaries (Thornton et al. 1997, Thornton and Running 1999). Slope (%) and aspect (degrees)

were derived from a 30 m digital elevation model using Arc Workstation GRID surface functions and commands (Environmental Systems Research Institute 1991).

Plot-level imputation techniques

The available 618 plots were randomly split without replacement into 154 plots (25%) constituting *P4* and 464 plots (75%) that, based on the year of their first measurement, belong to *P1*, *P2*, and *P3*.

Using the data from *P4*, the mean values of the variables of interest (*Y*) for the year 2000 (SAMPLE25 estimator) were calculated as:

$$[2.1] \quad \overline{Y}_{SAMPLE25} = \sum_{i:Y_i \in P4} Y_{t,i} / n_4$$

where $Y_{t,i}$ is the observed *Y* value of the i^{th} plot at time *t*, which is the year 2000, and n_4 is the number of plots in *P4*.

The MA estimator, the FIA default method, was also used to calculate the current mean values for the variables of interest:

$$[2.2] \quad \overline{Y}_{MA(4)} = 0.25 * \overline{Y}_{t-3,i} + 0.25 * \overline{Y}_{t-2,i} + 0.25 * \overline{Y}_{t-1,i} + 0.25 * \overline{Y}_{t,i}$$

where $\overline{Y}_{t-3,i}$, $\overline{Y}_{t-2,i}$, $\overline{Y}_{t-1,i}$, and $\overline{Y}_{t,i}$ are the mean values of the variables of interest of *P1*, *P2*, *P3*, and *P4*, respectively. This MA(4) estimator will be referred to as MA in the following. The MA takes into account that the panels include different numbers of plots. The following WMA takes the varying number of plots per panel into account and allows allocating weights declining with time lapsed since the most recent measurement:

[2.3] $$\overline{Y}_{WMA(4)} = w_{t-3} * \overline{Y}_{t-3,i} + w_{t-2} * \overline{Y}_{t-2,i} + w_{t-1} * \overline{Y}_{t-1,i} + w_t * \overline{Y}_{t,i}$$

where $w_{t-3}, w_{t-2}, w_{t-1}$, and w_t are the weights of *P1*, *P2*, *P3*, and *P4*, respectively. Larger weights were chosen for *P3* and *P4* ($w_{t-1} = w_t = 0.3$) than for *P1* and *P2* ($w_{t-3} = w_{t-2} = 0.2$). WMA(4) will be referred to as WMA.

Nearest neighbor (NN) imputation methods are donor-based methods where the imputed value is either a value that was actually observed for another plot or the average of values for more than one plot. Forest attributes that are measured on all plots are referred to as ancillary variables. Variables of interest are those forest attributes that are only measured on a subset of plots. Plots with measured ancillary variables and variables of interest are called reference plots and target plots are those that only have ancillary variables measured. In this study, the target plots were assumed to be non-sampled plots lacking inventory data (panels 1-3). The reference plots constituted the pool of potential plots with ground and ancillary data (*P4*), which could be selected to impute the inventory data for the target plots.

The most similar neighbor (MSN) method (Moeur and Stage 1995) has been shown to provide reasonable imputation results for forest attributes (Moeur and Stage 1995, LeMay and Temesgen 2005). The gradient nearest neighbor (GNN) method (Ohmann and Gregory 2002) has successfully been used to map forest composition and structure (Ohmann and Gregory 2002, Ohmann et al. 2007). The randomForest (RF) method (Crookston and Finley 2008) has been found to provide a flexible and robust alternative to traditional NN imputation methods such as MSN and GNN for

estimating forest attributes such as BA and SPH (Hudak et al. 2008). MSN, GNN, and RF were examined using the yaImpute R package version 1.0-6 (Crookston and Finley 2008). For MSN and GNN, the similarity between reference and target plots is defined using a weighted Euclidean distance:

[2.4] $$D_{ij}^{\,2} = (X_i - X_j) W (X_i - X_j)'$$

where W is the weight matrix, X_i is a vector of standardized values of the ancillary variables for the i^{th} target plot; and X_j is a vector of standardized values of ancillary variables for the j^{th} reference plot. The ancillary variables for both target and reference plots were standardized using the mean and variance of the ancillary variables of the reference plots.

For MSN, the weight used is $W = \Gamma \Lambda^2 \Gamma'$, where Γ is the matrix of standardized canonical coefficients for the ancillary variables and Λ^2 is the diagonal matrix of squared canonical correlations between ancillary attributes and ground variables (Moeur and Stage 1995). The "most similar" reference plot is hence selected based on similarity of the ancillary data, weighted by the correlations to the ground data. The ground data of the reference plot with the smallest distance is then imputed to the target plot.

The gradient nearest neighbor method (GNN) employs a projected ordination of the ancillary data based on canonical correspondence analysis (CCA) to assign the weights (Ohmann and Gregory 2002).

The RF method is a classification and regression tree (CART) method (Breiman 2001). The data and variables are randomly and iteratively sampled to generate a large group, or forest, of classification and regression trees. For RF two plots are considered similar if they tend to end up in the same terminal nodes in a forest of classification and regression trees. The distance measure is one minus the proportion of trees where a target plot is in the same terminal node as a reference plot (Crookston and Finley 2008, Hudak et al. 2008).

Instead of filling in the missing values for panels 1 to 3 with their previous measurements, as was done in the MA calculation, MSN, GNN, and RF were explored to impute the missing values, and then estimate the overall mean of the variables of interest for the year 2000:

$$[2.5] \quad \overline{Y}_{IMP} = \left[\sum_{i:Y_i \in P1} Y_{imp,i} + \sum_{i:Y_i \in P2} Y_{imp,i} + \sum_{i:Y_i \in P3} Y_{imp,i} + \sum_{i:Y_i \in P4} Y_{t,i} \right] / n$$

where IMP refers to the NN imputation method used and $Y_{imp,i}$ is the imputed Y value for the i^{th} plot.

BA, SPH, VOL, and BIOT were used as variables of interest and SAMPLE25, MA, WMA, and the three imputation methods were compared based on the overall means of the variables of interest in the year 2000 (see Equations 2.1-2.3 and 2.5).

Two sets of ancillary variables were tested for the imputation methods: The first set included climate, topography, and satellite data and the second set consisted of the previous measurements of the variables of interest that were taken at measurement occasion 1 in the years 1993 to 1997 (BAocc1, SPHocc1, VOLocc1, BIOTocc1).

The methods were compared by randomly splitting the available data of 618 plots into 154 reference and 464 target plots, applying each method, determining mean estimates for the variables of interest in the year 2000 (see Equations 2.1 – 2.3 and 2.5), and comparing the estimates to the observed mean values of the variables of interest in the year 2000:

[2.6] $$\overline{Y}_{OBS} = \sum_{i=1}^{n} Y_{t,i} / n$$

where $Y_{t,i}$ is the observed Y value of the i^{th} plot at time t, which is the year 2000.

The basis of evaluation was accuracy, as expressed by the root mean square error (RMSE), and bias, calculated as the mean difference between the estimates and the observed mean values (Equation 2.6) from 500 iterations of randomly splitting the data. Five hundred iterations were considered sufficient because other studies have found RMSE and bias to stabilize at around 200 iterations (e.g., Arner et al. 2004). Both RMSE and bias were expressed as percent of the observed mean for each variable of interest:

[2.7] $$Bias\ \% = \frac{\sum_{i=1}^{n} \frac{\left(est_i - obs_i\right)}{m}}{\frac{\sum_{i=1}^{n} obs_i}{m}} * 100$$

[2.8] $$RMSE\ \% = \frac{\sqrt{\sum_{i=1}^{n} \frac{\left(est_i - obs_i\right)}{m}}}{\frac{\sum_{i=1}^{n} obs_i}{m}} * 100$$

where m = 500.

Results

For BA and SPH, the RMSE values of MA were about half the size of those observed for SAMPLE25. For VOL and BIOT the RMSE values for MA were about a third of those observed for SAMPLE25. SAMPLE25 results were virtually unbiased with absolute values of 0.13% and less. Bias for the MA results ranged from -2.63% for SPH to -1.98% for BIOT. MA estimates were very precise and the bias contributed most to the RMSE. The opposite was true for the virtually unbiased SAMPLE25 estimates, where the variance contributed most to the RMSE (Table 2.4). WMA reduced the bias and with that the RMSE for SPH even further. For BA, VOL, and BIOT the bias became positive and the RMSE values increased for VOL and BIOT compared to those of the MA (Table 2.4).

When climate, topography, and satellite data were used as ancillary variables, MSN provided better results than SAMPLE25 in terms of RMSE for BA, VOL, and BIOT but worse results than MA and WMA. MSN imputation resulted in negligible bias with absolute values less than 0.3%, hence, outperforming the MA and WMA results in terms of bias. The variance contributed most to the RMSE values of the MSN estimates (Table 2.4). Using climate, topography, and satellite data as ancillary variables, RF provided slightly better results than MSN in terms of RMSE for all four variables of interest. With values ranging from 0.26% to 0.89% bias was slightly larger than for MSN but still negligible. As for SAMPLE25 and MSN, the variance contributed most to the RMSE (Table 2.4). GNN imputation results were by far the

worst when climate, topography, and satellite data were used as ancillary variables with RMSE values around 15% and positive bias around 10% (Table 2.4).

When BAocc1, SPHocc1, VOLocc1, and BIOTocc1 were used as ancillary variables, the MSN results had a negative bias ranging between -2.90% and -4.56%. The bias contributed most to the RMSE values, which were still slightly better than those of SAMPLE25. However, MA and WMA now outperformed MSN both in terms of bias and RMSE (Table 2.5). RF results improved both in terms of bias and RMSE when previous measurements were used as ancillary variables and outperformed MA in terms of bias and RMSE. RF also provided better results than WMA in terms of bias for all four variables of interest and for VOL and BIOT in terms of RMSE (Table 2.5). GNN estimates were even worse with the second set of ancillary variables, resulting in large positive bias exceeding 29% and large RMSE values exceeding 36% (Table 2.5).

Discussion

The SAMPLE25 estimator should provide unbiased estimates. In this study the bias was not equal to zero but reached values up to 0.13%. If all possible subsamples of size 154 were taken, SAMPLE25 should result in a bias of zero. Since not all possible subsamples were taken, the negligible bias observed for the method in this study was probably due to the number of iterations that was performed.

As found in other studies (Van Deusen 2002b, Arner et al. 2004), MA, the FIA default estimator, resulted in improvements in terms of RMSE compared to using only

the current panel as the basis of estimating current forest attributes. However, MA resulted in negatively biased estimates. This bias is commonly referred to as lag bias, which arises because the MA estimator tends to underestimate current forest conditions. In the given example, the four year gap between *P3* and *P4* increased the lag bias and it is expected that the lag bias would have been smaller for a regular four panel inventory where panels are only a year apart. Most studies on the MA performance have been done in other regions where the inventory cycle is five years, and the lag bias of the MA has been found to be more than compensated by a reduction in variance for a five year inventory cycle by 'borrowing' strength in terms of sample size from previous years (Johnson et al. 2003).

MA provides unbiased estimates for the midpoint of the period and is hence not valid as end of period estimator. When used as end of period estimator as done by FIA and in this study, the MA has the tendency to mask temporal trends (Roesch and Reams 1999) and provide biased estimates for the end of the period. One approach to solve this problem is to apply weights that give more weight to the most recently measured panels. This was done for the WMA, which provided improved estimates in terms of bias and RMSE for BA and SPH but increased bias and RMSE values for VOL and BIOT compared to MA. The selection of the weights poses a problem that is not yet solved. Choosing appropriate weights requires the knowledge of the trend inherent in the data which is hardly ever known. Breidt (1999) presented models that can be used for selecting the weights somewhat objectively. Arner et al. (2004) found an increase in RMSE for mean volume and mean annual volume change with

increasing larger weights for recent years, and Johnson et al. (2003) have shown that equal weights lead to the lowest RMSE in most situations.

P1, *P2*, and *P3* lack data of the national forests Rogue River, Coleville, and Winema, respectively (Table 2.1), which suggests that the panels might not have accurately characterized the population of interest. This feature could have been exacerbated by the random assignment of plots to *P4*. MA assumes that each yearly sample covers the population of interest (Johnson and Williams 2004). Hence, the MA results in this study might have been compromised by this data feature. FIA plots are assigned to the panels in a systematic manner, so that each FIA panel covers the population of interest systematically which ensures that the annual sample maintains its spatial properties. Hence, the performance of the MA estimator using actual FIA data is expected to be better than in the given example.

Longer inventory cycles will have negative effects on the performance of the MA and WMA estimators in terms of bias (Johnson et al. 2003). Hence, it is questionable whether the MA estimator is optimal for the PNW region where the inventory cycle length is 10 years. However, if the lag bias could be corrected, the MA and WMA estimators could provide RMSE values substantially lower than those of the SAMPLE25 estimator.

Three plot-level imputation techniques were examined which performed differently in terms of bias and RMSE compared to the SAMPLE25, MA, and WMA estimators. Although MSN imputation using climate, topography, and satellite data improved the results compared to the SAMPLE25 estimates in terms of RMSE for

BA, VOL, and BIOT, the improvements in RMSE seemed minor considering the computational expenses of applying imputation techniques. Employing imputation techniques is questionable if the improvements are not substantial. MA and WMA estimators outperformed MSN imputation in terms of RMSE when climate, topography, and satellite data were used as ancillary data and in both bias and RMSE when previous measurements were used as ancillary data. Hence, the results of this study did not indicate any advantage of MSN imputation over the MA and WMA estimators.

GNN results were not close to those obtained by SAMPLE25, MA, WMA, MSN or RF, which might be due to the fact that CCA requires the use of environmental factors for the ordination. GNN has been developed for pixel imputation (Ohmann and Gregory 2002) and it is possible that gradients in the environmental factors are not picked up when plot-level data is being used in combination with the available climate, topography, and satellite data. GNN should not be used with previous measurements as ancillary data since those do not provide any environmental factors that are necessary for the CCA step in the GNN analysis. This explains the bad results achieved by GNN with previous measurements as ancillary variables (see Table 2.5).

The results of this study support the findings of Hudak et al. (2008) that RF represents a robust alternative to traditional imputation methods. In this study, RF was the only imputation method that provided results that could compete with the results of the MA and WMA estimators. When RF was used with previous measurements as

ancillary variables it did not only outperform the WMA estimates in terms of bias but also in terms of RMSE for two of the four variables of interest. This suggests further exploration of this method with different data sets.

In a 10 panel inventory system, using previous measurements as ancillary variables is expected to result in overpredictions of the variables of interest. The current panel is used as reference data and its previous measurements are 10 years old. The previous measurements of the remaining nine panels constituting the target data are one to nine years old. Matching on previous measurements will result in overpredicting growth. Using an updated MA as introduced by Gartner and Reams (2002), where only the panels that have the most outdated measurements are being updated, might avoid the problem of overprediction when previous measurements are being used as ancillary variables. In a 10 panel system, the first five panels would be updated with imputation methods based on previous measurements as ancillary variables for estimating the status of the variables of interest in year 10. Then a MA would be calculated based on the updated values of panels one through five and the measurements obtained for panels six through 10.

The efficiency of the imputation methods depends on the strengths of relationships between the variables of interest and the ancillary data. The data in this study showed only weak association between forest inventory attributes and ancillary variables from TM images, climate, and topography data. The findings of this study do not provide any incentive to prefer the use of NN imputation methods that employ climate, topography, and satellite data as ancillary variables over the use of MA and

WMA estimators. Data of higher quality than that derived from TM images could have the potential to improve the NN imputation techniques. Variables derived from Light Detection and Ranging (LiDAR) data are an example (e.g., Hudak et al. 2008).

Throughout all estimation methods, RMSE was larger for VOL and BIOT than for BA. The poorer results for VOL and BIOT might be due to the fact that these two variables are transformations of both tree DBH and HT and, therefore, they are three dimensional variables on the landscape. BA, on the other hand, is a two dimensional variable because it is based only on the DBH measurements. Many of the ancillary variables available in this study for imputation are themselves only two dimensional variables. Again, three dimensional LiDAR data have the potential to improve the imputation techniques for VOL and BIOT.

The results of the imputation methods might have been impaired by a combination of the number of plots used as reference stands (*P4*, 154 plots) and the large number of species and forest types in the six national forests that were used in this study. The diversity in the data and the small number of plots suggest that it was probably not easy to find good matches in some of the cases. Since imputation methods do not extrapolate and only interpolate when $k > 1$ (Crookston et al. 2002), it is important that the reference data spans the full range of the population in the space of the ancillary variables without any large gaps. If this is not given, the availability of similar reference observations may be reduced and imputation error increases (Stage and Crookston 2007). The random assignment of plots to *P4* might have resulted in plot combinations for *P4* that did not represent the population well which would have

negatively influenced the performance of the imputation methods. FIA annual inventory data assures a systematic coverage of the population of interest for each panel so that it seems more likely to find good matches and an improvement of imputation results could be expected.

Conclusions

Compared to the SAMPLE25 estimator the MA estimator improved the estimates in terms of RMSE and worsened the estimates in terms of bias. The WMA estimator improved the results for two of the variables of interest compared to the MA. The performance of the MA and WMA estimators should be explored using an actual 10-year inventory system in order to examine the increase in lag bias for a long inventory cycle. Different weighting schemes in a 10-year inventory system need to be explored for the WMA estimator.

With the available ancillary data, MSN and GNN could not compete with any of the other estimation methods. RF results were best when previous measurements of the variables of interest were used as ancillary variables and outperformed the MA and WMA estimators in terms of bias and were comparable in terms of RMSE. Using RF imputation with previous measurements as ancillary variables might provide an approximately unbiased alternative to the biased MA and WMA estimators in the PNW. Because overprediction of the variables of interest might occur, more research on the behavior of this method in a 10 panel system is warranted.

For the MA and WMA estimates, the variance was very small and bias contributed most to the RMSE values. If the lag bias could be corrected, the RMSE values would be reduced substantially and the MA and WMA results might outperform all other methods. Methods for correcting the MA and WMA lag bias should be sought. If the lag bias is not corrected for, users should be aware that they are estimating a midpoint value rather than an end of period value when they use the MA estimator.

Due to the data structure and the random assignment of plots to *P4*, the panels did not always represent the population well. This had impacts on the MA and WMA estimates as well as on the NN imputation results. All methods are expected to show improved results when actual FIA data is used since FIA panels provide complete coverage of the population with equal number of plots for each year.

Acknowledgements

We thank Janet Ohmann and Matt Gregory for sharing their ancillary data from the NLCD and DAYMET with us and for their insights on the data. We appreciate the help that Jim Alegria, Carol Apple, Bob Brown, and Melinda Moeur provided in obtaining national forest data, and thank Kurt Campbell for assistance with volume and biomass equations. We also thank David Hann, three anonymous reviewers, and the assistant editor for their helpful review comments.

Table 2.1: Number of plots measured by year of installation and national forest and corresponding panel assignment. All plots listed were remeasured in 2000.

Year of Installation	Colville	Mt. Hood	Ochoco	Rogue River	Wallowa-Whitman	Winema	Total	Assigned Panel
1993	7	0	0	0	0	0	**7**	**1**
1994	4	9	23	0	99	94	**229**	**1**
1995	0	51	41	20	77	34	**223**	**2**
1996	16	51	18	50	23	0	**158**	**3**
1997	1	0	0	0	0	0	**1**	**3**
Total	**28**	**111**	**82**	**70**	**199**	**128**	**618**	

Table 2.2: Summary of plot-level variables.

Variable	Minimum	Mean	Maximum	Std
Basal area (m^2/ha)	0.24	24.32	105.35	19.00
SPH (stems/ha)	1	305	1517	221
Volume (m^3/ha)	0.66	224.82	1444.74	221.04
Total gross oven dry weight biomass (tons/ha)	0.58	134.09	800.64	132.64
Canopy cover (%)	0	54	97	29
Slope (%)	0	23	83	17
Elevation (m)	274	1389	2377	321
Annual precipitation (ln cm) (scaled * 100)	577	683	817	48
Mean annual Temperature (ºC) (scaled * 100)	60	579	1067	166

Table 2.3: Tree species found in this study.

Scientific Name	Common Name	Frequency
Abies amabilis (Douglas ex Louden) Douglas ex Forbes	Pacific silver fir	913
Abies concolor (Gord. & Glend.) Lindl. ex Hildebr.	White fir	2,325
Abies grandis (Douglas ex D. Don) Lindl.	Grand fir	2,935
Abies lasiocarpa (Hook.) Nutt.	Subalpine fir	645
Abies ×*shastensis* (Lemmon) Lemmon [*magnifica* × *procera*]	Shasta red fir	854
Abies procera Rehder	Noble fir	245
Acer macrophyllum Pursh	Bigleaf maple	93
Alnus rubra Bong.	Red alder	81
Arbutus menziesii Pursh	Pacific madrone	126
Betula papyrifera Marsh. var. *commutata* (Regel) Fernald	Paper birch	6
Castanopsis chrysophylla (Douglas ex Hook.) A. DC.	Golden chinquapin	71
Calocedrus decurrens (Torr.) Florin	Incense cedar	140
Cornus nuttallii Audubon ex Torr. & A. Gray	Pacific dogwood	3
Juniperus occidentalis Hook.	Western juniper	517
Larix occidentalis Nutt.	Western larch	801
Pinus albicaulis Engelm.	Whitebark pine	96
Pinus attenuata Lemmon	Knobcone pine	14
Pinus contorta Douglas ex Louden	Lodgepole pine	2,758
Picea engelmannii Parry ex Engelm.	Engelmann spruce	651
Pinus lambertiana Douglas	Sugar pine	177
Pinus monticola Douglas ex D. Don	Western white pine	78
Pinus ponderosa C. Lawson	Ponderosa pine	5,040
Populus balsamifera L. ssp. *trichocarpa* (Torr. & A. Gray ex Hook.) Brayshaw	Black cottonwood	5
Populus tremuloides Michx.	Quaking aspen	33
Prunus emarginata (Douglas ex Hook.) D. Dietr.	Bitter cherry	2
Pseudotsuga menziesii (Mirb.) Franco	Douglas-fir	8,202
Quercus chrysolepis Liebm.	Canyon live oak	184
Quercus garryana Douglas ex Hook.	Oregon white oak	11
Quercus kelloggii Newberry	California black oak	8
Taxus brevifolia Nutt.	Pacific yew	35
Thuja plicata Donn ex D. Don	Western redcedar	577
Tsuga heterophylla (Raf.) Sarg.	Western hemlock	2,123
Tsuga mertensiana (Bong.) Carrière	Mountain hemlock	960

Table 2.4: Imputation results for the set of ancillary variables that included climate, topography, and satellite data. The weights for the WMA are given in parentheses as follows (w_{t-3}, w_{t-2}, w_{t-1}, w_t).

	BA		SPH		VOL		BIOT	
Method	**% bias**	**% RMSE**	**% bias**	**% RMSE**	**% bias**	**% RMSE**	**% bias**	**% RMSE**
SAMPLE25	-0.02	5.29	0.13	5.05	-0.08	6.59	-0.06	6.67
MA	-2.54	2.60	-2.63	2.68	-1.92	2.06	-1.98	2.12
WMA (0.2, 0.2, 0.3, 0.3)	0.58	0.98	-1.58	1.74	2.51	2.71	2.58	2.78
MSN	0.05	3.73	0.29	5.13	-0.19	5.01	-0.15	4.97
GNN	10.14	15.10	10.11	14.89	8.97	15.72	9.67	16.35
RF	0.44	3.60	0.89	4.99	0.37	4.96	0.26	4.89

Table 2.5: Imputation results for using occasion 1 measurements of the variables of interest (BAocc1, SPHocc1, VOLocc1, and BIOTocc1) as ancillary data. The weights for the WMA are given in parentheses as follows (w_{t-3}, w_{t-2}, w_{t-1}, w_t).

	BA		SPH		VOL		BIOT	
Method	**% bias**	**% RMSE**	**% bias**	**% RMSE**	**% bias**	**% RMSE**	**% bias**	**% RMSE**
SAMPLE25	-0.02	5.29	0.13	5.05	-0.08	6.59	-0.06	6.67
MA	-2.54	2.60	-2.63	2.68	-1.92	2.06	-1.98	2.12
WMA (0.2, 0.2, 0.3, 0.3)	0.58	0.98	-1.58	1.74	2.51	2.71	2.58	2.78
MSN	-3.91	4.35	-2.90	3.68	-4.41	4.97	-4.56	5.09
GNN	30.67	36.12	43.92	51.48	29.55	36.57	34.53	41.09
RF	-0.30	1.58	-0.85	2.78	-0.06	1.90	-0.09	1.79

CHAPTER 3: TREE-LEVEL IMPUTATION TECHNIQUES TO ESTIMATE CURRENT PLOT-LEVEL ATTRIBUTES IN THE PACIFIC NORTHWEST USING PANELED INVENTORY DATA

Bianca N.I. Eskelson

Temesgen Hailemariam

Tara M. Barrett

Submitted to: Proceedings of the 9th Annual Forest Inventory and Analysis Symposium

Abstract

The Forest Inventory and Analysis program (FIA) of the US Forest Service conducts a nationwide annual inventory. One panel (20% or 10% of all plots in the eastern and western United States, respectively) is measured each year. The precision of the estimates for any given year from one panel is low, and the moving average (MA), which is considered to be the default estimator, can result in biased estimates of current conditions. An alternative to the MA is sought after, and studies comparing different alternatives to the MA approach for estimating current forest attributes in the Pacific Northwest are lacking. Paneled data from national forests in Oregon and Washington were used to explore nearest neighbor (NN) imputation methods to project all panels to a common point in time. When using the most recent ground measurements of the panels measured in prior years as ancillary data, tree-level NN imputation outperformed the MA estimator in estimating basal area/ha, stems/ha, volume/ha, and biomass/ha in terms of bias and root mean square error (RMSE) and plot-level NN imputation in terms of RMSE. When basal area/ha, stems/ha, volume/ha, and biomass/ha were summarized by three species groups, tree-level NN imputation outperformed plot-level NN imputation in terms of both bias and RMSE. Tree-level NN imputation outperformed the MA in terms of bias and RMSE for estimating basal area/ha, stems/ha, volume/ha, and biomass/ha for species group ‘pine’ and provided comparable results in terms of bias and RMSE for species groups ‘Douglas-fir’ and ‘other.’

Introduction

Information on current forest condition is essential to assess and characterize resources and to support management and policy decisions. The 1998 Farm Bill mandates the US Forest Service to conduct annual inventories to provide annual updates of each state's forest. Only 10% or 20% of all plots in the western and eastern United States, respectively, are measured annually. Because only a portion of the full sample is measured annually, the precision of the estimates for any given year is low. To achieve higher precision, the Forest Inventory and Analysis program (FIA) uses a moving average (MA) as default estimator which combines the data of multiple panels. In the presence of trend, biased estimates will result, if the MA is applied to the end of the period to estimate current conditions. Other approaches to combine data from all panels include: 1) updating unmeasured panel data to the current year with growth models (Lessard et al. 2001); 2) using time series models (Johnson et al. 2003); 3) mixed estimation (Van Deusen 1996); or 4) filling in missing panel data using tree- and plot-level imputation techniques (Gartner and Reams 2001, 2002, McRoberts 2001). Since spatial, temporal, and forest characteristics differ within and among regions it is unclear if any single technique will provide satisfactory results for all regions (Patterson and Reams 2005). It may be necessary to evaluate different methods for a variety of issues and regions. Studies comparing different alternatives to the MA approach for estimating current forest attributes in the Pacific Northwest (PNW) are lacking.

Nearest Neighbor (NN) imputation methods are donor-based, which means that the imputed value was either observed for another unit or was calculated as the average of values from more than one unit. NN imputation can be performed on different levels. Eskelson et al. (in press; Chapter 2) have shown that plot-level imputation, that is plot-level attributes (e.g., basal area/ha) are imputed, can provide more accurate results than the MA approach. They found the randomForest (RF) imputation method (Crookston and Finley 2008), which is an extension of classification and regression tree (CART) methods (Breiman 2001), to outperform other NN imputation methods. Imputation can also be performed at the tree-level, that is tree-level attributes (e.g., diameter at breast height (DBH in cm)) are imputed, and the results of the tree-level imputation are then summarized for each plot (e.g., imputed DBH is used to calculate basal area/ha).

Depending on the intended use, tree- and plot-level imputation techniques differ in their predictive abilities and suitability (Gartner and Reams 2002). Plot-level and tree-level NN imputation techniques might have a similar relationship to each other as whole stand growth models, which might not apply in heterogeneous conditions (Curtis and Hyink 1985), have with single-tree growth models, which can provide more detailed information about stand dynamics and structure (Burkhart 1992). Tree-level nearest neighbor (NN) imputation techniques have been successfully used to estimate tree volumes and heights (Korhonen and Kangas 1997), single-tree biomass (Fehrmann et al. 2008) as well as 5-year diameter growth and bark thickness (Sironen et al. 2001, 2003, 2008).

The objectives of this study are to: 1) use paneled data from the PNW to estimate current forest attributes (see Table 3.1) using tree-level imputation methods and compare their performance against the MA and the estimates based only on the data from the current panel; 2) examine the performance of tree-level imputation methods to estimate current forest attributes by species groups; and 3) compare tree-level and plot-level imputation results.

Methods

Data

The data used in this study consist of 618 plots from six national forests that were collected as part of the Pacific Northwest Region's Current Vegetation Survey (CVS) of the US Forest Service. The plots were installed between 1993 and 1997 and remeasured in 2000. The particular national forests sampled were the Colville (28), Mt. Hood (111), Ochoco (82), Rogue River (70), Wallowa-Whitman (199), and Winema (128).

Five plots are installed in each basic CVS sampling unit, which is one hectare (ha) in size. Each plot consists of three permanent circular, nested subplots of different sizes in which trees are measured depending upon their DBH. For a detailed description of the CVS inventory see Max et al. (1996). Tree height (HT in m) is only subsampled and missing HTs were filled using height models developed in Barrett (2006) for live trees with DBH of 12.7 cm or larger. Volume and biomass equations from the US Forest Service were used to calculate gross cubic-meter volume and total

gross oven dry weight biomass (USDA 2000). For each plot, basal area in m^2 per ha (BA), stems per ha (SPH), volume in m^3 per ha (VOL), and biomass in tons per ha (BIOT) were calculated and summarized (Table 3.1). BA, SPH, VOL, and BIOT were also calculated for each of the following three species groups: 1) 'Douglas-fir'; 2) 'pine' including all occurring pine species; and 3) 'other' including other conifers and hardwoods. Basal area in larger trees (BAL in m^2) was calculated for each tree. Ingrowth for each plot was determined by calculating BA, SPH, VOL, and BIOT for all trees that were present in 2000 but not present at the first measurement occasion. BA and SPH were also calculated for small trees with DBH larger than 2.54 cm and smaller than 12.7 cm.

The data set comprises 30,709 trees in 33 species. The most common species in decreasing order are Douglas-fir (*Pseudotsuga menziesii* (Mirb.) Franco), ponderosa pine (*Pinus ponderosa* C. Lawson), grand fir (*Abies grandis* (Douglas ex D. Don) Lindl.), lodgepole pine (*Pinus contorta* Douglas ex Louden), white fir (*Abies concolor* (Gord. & Glend.) Lindl. ex Hildebr.), and western hemlock (*Tsuga heterophylla* (Raf.) Sarg.) (see Chapter 2 (Eskelson et al. in press) for details).

In NN imputation methods, ancillary variables are those variables that are measured on all units. Thematic Mapper (TM) images from 2000 were extracted from the national land-cover database 2001 [NLCD 2001 (Homer et al. 2004)] and used as ancillary data. The raw imagery bands 1 to 5 and band 7 (TM1, TM2, TM3, TM4, TM5, TM7) as well as the Tasseled Cap (TC) transformations (TC1 – TC6) (Kauth and Thomas 1976) were used. The normalized difference vegetation index (NDVI)

and three commonly used band ratios (band 4 to band 3 (R43), band 5 to band 4 (R54), and band 5 to band 7 (R57)) were calculated. Percent canopy cover was extracted from the NLCD 2001 (Homer et al. 2004).

The following climate and topography variables for plot locations were additionally used as ancillary data: Annual precipitation and mean annual temperature (Table 3.1) [Data source: DAYMET Daily Surface Weather Data and Climatological Summaries (Thornton et al. 1997, Thornton and Running 1999)], elevation (EL in m) and transformations (EL^2, ln(EL)) [Data source: CVS inventory], and slope (%) and aspect (degrees) and transformations (cosine(aspect), sine(aspect), cosine(aspect)*slope, and sine(aspect)*slope) [Data source: 30 m digital elevation model using Arc Workstation GRID surface functions and commands (Environmental Systems Research Institute 1991)]. These climate, topography, and satellite variables have been successfully used as ancillary data for NN imputation methods in previous studies (e.g., Chapter 2 (Eskelson et al. in press), Ohmann and Gregory 2002).

Imputation techniques

Panel data is a special case of inventory data with measurements taken at different times. All plots were remeasured in 2000. In order to mimic a panel system with the available data 25% of the plots (154) were randomly assigned to *P4* and the remaining 75% of the plots (464) were assigned to *P1*, *P2*, and *P3* based on their year of installation. This resulted in *P1*, *P2*, and *P3* having different sizes for each iteration (Table 3.2).

The variables of interest (*Y*) in this study were BA, SPH, VOL, and BIOT. Their observed mean value in the year 2000 was calculated as:

$$[3.1] \quad \overline{Y}_{OBS} = \sum_{i=1}^{n} Y_i / n$$

where Y_i is the observed *Y* value of the i^{th} plot in 2000 and $n = 618$. The observed mean value was used as best available estimate of the true mean.

For each plot in *P4*, BA, SPH, VOL, and BIOT were calculated using the tree data from *P4*. The mean values of *Y* for the year 2000 (SAMPLE25 estimator) were calculated as:

$$[3.2] \quad \overline{Y}_{SAMPLE25} = \sum_{i:Y_i \in P4} Y_i / n_4$$

where Y_i is the observed *Y* value of the i^{th} plot, and n_4 is the number of plots in *P4*.

The MA estimator, the FIA default method, is:

$$[3.3] \quad \overline{Y}_{MA(4)} = 0.25 * \overline{Y}_{t-3,i} + 0.25 * \overline{Y}_{t-2,i} + 0.25 * \overline{Y}_{t-1,i} + 0.25 * \overline{Y}_{t,i}$$

where $\overline{Y}_{t-3,i}$, $\overline{Y}_{t-2,i}$, $\overline{Y}_{t-1,i}$, and $\overline{Y}_{t,i}$ are the mean values of the variables of interest of *P1*, *P2*, *P3*, and *P4*, respectively. The MA takes into account that the panels include different numbers of plots. Instead of equal weighting of the panels a weighted version of [3.3] is proposed:

$$[3.4] \quad \overline{Y}_{WMA(4)} = w_{t-3} * \overline{Y}_{t-3,i} + w_{t-2} * \overline{Y}_{t-2,i} + w_{t-1} * \overline{Y}_{t-1,i} + w_t * \overline{Y}_{t,i}$$

where w_{t-3}, w_{t-2}, w_{t-1}, and w_t are the weights of *P1*, *P2*, *P3*, and *P4*, respectively. Larger weights were chosen for *P3* and *P4* ($w_{t-1} = w_t = 0.3$) than for *P1* and *P2*

($w_{t-3} = w_{t-2} = 0.2$). MA(4) and WMA(4) will be referred to as MA and WMA, respectively.

Instead of using the previous measurements to fill in the *Y* values for *P1*, *P2*, and *P3*, as is done with MA and WMA, the current *Y* values of *P1*, *P2*, and *P3* were imputed using tree-level RF and plot-level RF imputation. Target data are units that have ancillary variables measured only (e.g., trees or plots in *P1*, *P2*, and *P3*). Reference data are units where both variables of interest and ancillary variables were measured (e.g., trees or plots in *P4*). RF imputation was employed using the yaImpute R package (Crookston and Finley 2008). Details on RF imputation can, for example, be found in Hudak et al. (2008).

For tree-level RF, the target trees were assumed to be non-sampled trees lacking inventory data in 2000. DBH, HT, and mortality for each target tree were imputed using DBH, HT, and BAL at the previous measurement (DBHocc1, HTocc1, and BALocc1) as ancillary data. The reference trees constituted the pool of potential trees with inventory and ancillary data (*P4*), which could be selected to impute the DBH, HT, and mortality for the target trees. Ingrowth of BA, SPH, VOL, and BIOT was imputed at the plot-level using BA and SPH of small trees at the previous measurement as well as the available climate, topography, and satellite data as ancillary data. BA, SPH, VOL, and BIOT were calculated for each plot based on the imputed tree data and the imputed ingrowth.

For plot-level RF the previous measurements of the four variables of interest (BAocc1, SPHocc1, VOLocc1, BIOTocc1) were used as ancillary data since this was

found to provide better imputation results than using climate, topography, and satellite data in a previous study. For more details see Chapter 2 (Eskelson et al. in press).

For both tree-level and plot-level RF the overall mean of the variables of interest for the year 2000 was estimated as:

$$[3.5] \quad \bar{Y}_{IMP} = \left[\sum_{i:Y_i \in P1} Y_{imp,i} + \sum_{i:Y_i \in P2} Y_{imp,i} + \sum_{i:Y_i \in P3} Y_{imp,i} + \sum_{i:Y_i \in P4} Y_{t,i} \right] / n$$

where $Y_{imp,i}$ is the imputed Y value for the i^{th} plot and 'IMP' refers to either tree-level RF or plot-level RF.

SAMPLE25, MA, WMA, and the tree-level and plot-level RF imputation methods were compared based on the overall means of the variables of interest in 2000 (see Equations 3.2 – 3.5). The five estimation methods were also compared based on their performance of estimating the four variables of interest by species groups 'Douglas-fir', 'pine', and 'other.'

The basis of evaluation was accuracy, as expressed by the root mean square error (RMSE), and bias, calculated as the mean difference between the estimates (Equations 3.2 – 3.5) and the observed mean values (Equation 3.1) from $m = 200$ iterations of randomly splitting the data. Two hundred iterations were considered sufficient because other studies have found RMSE and bias to stabilize at around 200 iterations (e.g., Arner et al. 2004). Both RMSE and bias were expressed as percent of the observed mean for each variable of interest:

[3.6] $$Bias\ \% = \frac{\sum_{i=1}^{n} \frac{(est_i - obs_i)}{m}}{\frac{\sum_{i=1}^{n} obs_i}{m}} * 100$$

[3.7] $$RMSE\ \% = \frac{\sqrt{\sum_{i=1}^{n} \frac{(est_i - obs_i)}{m}}}{\frac{\sum_{i=1}^{n} obs_i}{m}} * 100$$

Results

The SAMPLE25 estimator provided virtually unbiased estimates for all four variables of interest. Its RMSE values ranged from 4.89% for SPH to 6.58% for BIOT. The MA estimates had a negative bias with values from -1.93% for VOL to -2.58% for SPH. The MA estimator provided very precise estimates with the bias contributing most to the RMSE. The WMA estimator reduced both bias and RMSE for BA and SPH. For WMA the bias of VOL and BIOT estimates was positive and the RMSE was larger than those for MA (Table 3.3).

Plot-level RF imputation resulted in small negative bias and smaller RMSE values than those of the MA estimator. In terms of RMSE, plot-level RF imputation only outperformed the WMA for VOL and BIOT (Table 3.3).

Tree-level RF imputation produced a small positive bias in BA, VOL, and BIOT but a small negative bias in SPH. Its RMSE values were smaller than those of the MA and the plot-level RF imputation. Tree-level RF imputation outperformed the

WMA estimates in terms of bias and RMSE for SPH, VOL, and BIOT. The variance contributed most to the RMSE for both tree- and plot-level imputation (Table 3.3).

By species group the SAMPLE25 estimator provided virtually unbiased results (0.62% or less). RMSE values ranged from 8.52% for 'pine' BA to 11.17% for 'other' BIOT (Table 3.4).

The MA estimator resulted in a larger negative bias for the four variables of interest for species group 'pine' which contributed most to the RMSE values of more than 9%. For the species group 'Douglas-fir' and 'other,' MA resulted in small bias with absolute values ranging from 0.30% to 1.17% and RMSE values ranging from 1.00% to 1.68% (Table 3.4).

WMA estimates were biased for all three species groups with the bias being largest for 'pine.' The bias contributed most to the RMSE values, which exceeded the RMSE values of the MA estimates and the RMSE values for 'pine' for the SAMPLE25 estimates (Table 3.4).

Plot-level RF imputations resulted in a smaller bias than with WMA for all species groups. However, RMSE values for RF exceeded those of WMA for all but 'pine' (Table 3.4).

Tree-level RF imputation outperformed SAMPLE25, WMA, and plot-level RF imputation in terms of RMSE. Compared to MA, tree-level RF imputation provided smaller RMSE values for 'pine' and slightly larger RMSE values for 'Douglas-fir' and 'other' (Table 3.4).

Discussion

The performance of the MA estimator in terms of the variance-bias trade-off was as expected. As in most other studies (e.g., Arner et al. 2004, Johnson et al. 2003, Van Deusen 2002), the large bias was found to be more than compensated for by the high precision. Hence, MA provided better estimates in terms of accuracy than SAMPLE25. MA is a temporal 'midpoint' estimator yielding biased estimates at the end of a time-series in the presence of trend (Roesch and Reams 1999). Giving more weight to the more recently measured panels resulted in the WMA estimator, which improved the estimates for BA and SPH in terms of bias and hence, also in terms of RMSE.

MA by species groups outperformed WMA in terms of both bias and RMSE. The larger weights applied for *P3* and *P4* for WMA increased the negative bias for species group 'pine' and resulted in large positive bias for all variables of interest for species groups 'Douglas-fir' and 'other.' This indicates that weights applied to the WMA which improve the MA for estimating BA, SPH, VOL, and BIOT do not necessarily improve the MA when the variables of interest are summarized by species group. Choosing appropriate weights for the WMA requires the knowledge of the trend inherent in the data. If the trend inherent in BA, SPH, VOL, and BIOT differs from the trend of the variables of interest summarized by species group, different weights need to be chosen for the WMA. Objective ways for choosing appropriate weights are still lacking. Panels that do not change much should receive larger weights than panels that exhibit a lot of change. Knowledge about change could possibly be

acquired from remotely sensed data, growth models, or other information on, for example, fire or insect outbreaks.

Tree-level RF imputation outperformed MA in terms of bias and RMSE for estimating BA, SPH, VOL, and BIOT. This is due to the lag bias inherent in the MA estimator. Tree-level imputation attempts to update the tree data, which results in a smaller bias than that observed for MA. Compared to the WMA, which tries to adjust the lag bias of the MA estimator, the improvement of tree-level RF imputation is less pronounced. If the lag bias of the MA could be adjusted, MA might outperform RF tree-level imputation in terms of both bias and RMSE since the MA estimates are more precise than those of the tree-level RF imputation.

When the variables of interest were summarized by species groups, the MA slightly outperformed tree-level RF imputation in terms of RMSE for species groups ‘Douglas-fir’ and ‘other’ because the MA resulted in low bias for those variables. For species group ‘pine’ MA resulted in large bias and therefore tree-level RF imputation provided much better results for the variables of interest for species group ‘pine’ in terms of bias and RMSE.

Tree-level RF imputation outperformed plot-level imputation for estimating BA, SPH, VOL, and BIOT as well as for estimating the variables of interest summarized by species groups. The results of this study suggest that tree-level RF imputation should be preferred over plot-level RF imputation for estimating total BA, SPH, VOL, and BIOT or for estimating BA, SPH, VOL, and BIOT by species group. The same considerations for choosing single-tree growth models over whole-stand

growth models probably apply for choosing tree-level NN imputation over plot-level NN imputation and depend mainly on the demands of the user.

In this study, tree-level variables were imputed using reference trees irrespective of whether the tree species of reference and target trees matched. Imputing only within tree species or species group might improve the results for tree species such as Douglas-fir, ponderosa pine, grand fir, lodgepole pine, white fir, and western hemlock which occur frequently in the data set. However, results for rare tree species would definitely degrade with decreasing number of possible reference trees. Overall results could possibly be improved by imputing tree-variables for frequent tree species within tree species but using the complete reference data set for rare tree species.

Conclusions

This study has shown that tree-level RF imputation has the potential to provide better results in terms of bias and accuracy for estimating plot-level attributes such as BA, SPH, VOL, and BIOT than can be achieved with the SAMPLE25, MA, and WMA estimators, or plot-level RF imputation.

Giving more weight to most recently measured panels by using a WMA improved the estimates for BA, SPH, VOL, and BIOT compared to the MA estimates. When the variables of interest were summarized by species group, MA outperformed WMA in terms of bias and accuracy. More research is warranted for finding objective methods for choosing appropriate weights.

Tree-level RF imputation outperformed MA and WMA in terms of bias and accuracy when BA, SPH, and VOL were estimated. When the variables of interest were summarized by species group, MA provided slightly better estimates in terms of accuracy than tree-level RF imputation.

Tree-level imputation outperformed plot-level imputation. This might be due to the fact that tree-level NN imputation requires more information and is based on a more detailed representation of the stand than plot-level imputation.

The results of the tree-level NN imputation methods tested in this study provide a good argument to further develop the application of tree-level NN imputation techniques for estimating current forest attributes from paneled inventory data.

Acknowledgements

We thank Janet Ohmann and Matt Gregory for sharing their ancillary data from the NLCD and DAYMET with us and for their insights on the data. We appreciate the help that Jim Alegria, Carol Apple, Bob Brown, and Melinda Moeur provided in obtaining national forest data, and thank Kurt Campbell for assistance with volume and biomass equations. We also thank Frank Roesch, Steen Magnussen, and Emilie Grossmann for their helpful review comments.

Table 3.1: Summary of plot-level variables in 2000.

Variable	Minimum	Mean	Maximum	Std
Basal area (m^2/ha)	0.24	24.32	105.35	19.00
SPH (stems/ha)	1	305	1517	221
Volume (m^3/ha)	0.66	224.82	1444.74	221.04
Total gross oven dry weight biomass (tons/ha)	0.58	134.09	800.64	132.64
Canopy cover (%)	0	54	97	29
Slope (%)	0	23	83	17
Elevation (m)	274	1389	2377	321
Annual precipitation (ln cm) (scaled * 100)	577	683	817	48
Mean annual Temperature (ºC) (scaled * 100)	60	579	1067	166

Table 3.2: Number of plots measured by year of installation and corresponding panel assignment. All plots listed were remeasured in 2000.

Year of Installation	# of Plots	Assigned Panel
1993	7	1
1994	229	1
1995	223	2
1996	158	3
1997	1	3

Table 3.3: Imputation results.

Method	BA		SPH		VOL		BIOT	
	% bias	% RMSE	% bias	% RMSE	% bias	% RMSE	% bias	% RMSE
SAMPLE25	0.05	5.29	-0.20	4.89	0.20	6.53	0.26	6.58
MA	-2.53	2.60	-2.58	2.63	-1.93	2.08	-1.97	2.12
WMA	0.59	1.03	-1.54	1.72	2.52	2.74	2.62	2.83
plot-level RF	-0.44	1.50	-0.73	2.52	-0.26	1.78	-0.22	1.66
tree-level RF	0.44	1.09	-0.60	1.31	0.43	1.36	0.42	1.35

Table 3.4: Tree- and plot-level imputation results by species group.

	BA 'Douglas-fir'		BA 'pine'		BA 'other'		SPH 'Douglas-fir'		SPH 'pine'		SPH 'other'	
Method	**% bias**	**% RMSE**	**% bias**	**% RMSE**	**% bias**	**% RMSE**	**% bias**	**% RMSE**	**% bias**	**% RMSE**	**% bias**	**% RMSE**
SAMPLE25	0.41	9.94	0.29	8.52	-0.33	10.10	-0.39	10.21	0.49	10.04	-0.56	9.99
MA	0.40	1.14	-9.85	9.89	-0.70	1.18	-0.30	1.00	-9.28	9.35	0.61	1.15
WMA	6.63	6.79	-13.49	13.57	3.86	4.12	2.21	2.64	-14.51	14.61	4.95	5.14
plot-level RF	2.34	9.65	2.19	9.75	-4.01	8.25	2.05	11.01	2.01	10.36	-4.13	9.01
tree-level RF	-0.54	1.63	0.98	1.68	-0.39	1.21	-0.57	2.25	-1.54	2.80	-0.93	1.74

	VOL 'Douglas-fir'		VOL 'pine'		VOL 'other'		BIOT 'Douglas-fir'		BIOT 'pine'		BIOT 'other'	
Method	**% bias**	**% RMSE**	**% bias**	**% RMSE**	**% bias**	**% RMSE**	**% bias**	**% RMSE**	**% bias**	**% RMSE**	**% bias**	**% RMSE**
SAMPLE25	0.62	10.99	0.23	8.84	-0.15	11.05	0.56	10.68	0.23	8.62	0.02	11.17
MA	-0.77	1.50	-9.27	9.31	0.48	1.19	-1.17	1.68	-9.36	9.40	0.48	1.21
WMA	6.88	7.09	-11.89	11.98	5.59	5.82	6.06	6.28	-12.15	12.23	5.95	6.17
plot-level RF	2.00	9.94	1.94	10.36	-3.17	8.57	1.81	9.63	2.18	10.24	-3.04	8.39
tree-level RF	-0.84	2.05	3.74	4.02	-1.25	1.97	-3.37	3.76	8.36	8.54	-0.91	1.81

CHAPTER 4: IMPUTING MEAN ANNUAL CHANGE AND ESTIMATING CURRENT FOREST ATTRIBUTES

Bianca N.I. Eskelson

Tara M. Barrett

Temesgen Hailemariam

Submitted to: Silva Fennica

Abstract

When a temporal trend in forest conditions is present, standard estimates from paneled forest inventories can be biased. Thus methods that use more recent remote sensing data to improve estimates are desired. Paneled inventory data from national forests in Oregon and Washington, U.S.A., were used to explore three nearest neighbor (NN) imputation methods to estimate mean annual change of four forest attributes (basal area/ha, stems/ha, volume/ha, biomass/ha). The randomForest imputation method outperformed the other imputation approaches in terms of root mean square error. The imputed mean annual change was used to project all panels to a common point in time by multiplying the mean annual change with the length of the growth period between measurements and adding the change estimate to the previously observed measurements of the four forest attributes. The resulting estimates of the mean of the forest attributes at the current point in time outperformed the estimates obtained from the standard national estimator, a moving average approach.

Introduction

Keeping national inventories of forests updated to reflect current conditions poses substantial logistical and accuracy issues (Gillis and Leckie 1996). In theory, a paneled inventory system can provide current information when only the most recent panel is used (Reams and Van Deusen 1999), but in practice, most forestry applications are at a spatial scale that require combining field plots from multiple years to achieve sufficient information (McRoberts 2001, Tomppo et al. 2008).

However, combining field plots from multiple years to estimate current conditions can cause a lag bias when forest conditions are changing over time.

In the United States (US), the national inventory of forests is collected with a panel system by the Forest Inventory and Analysis (FIA) program of the US Forest Service. The FIA default estimator is a moving average (MA) approach (Bechtold and Patterson 2005) which is known to result in biased estimates when trend is present (Van Deusen 2002b). In the western US, the problem of lag bias is exacerbated by a relatively long remeasurement interval (10 years), shifts in forest management in response to altered economic and social conditions, changing climate, and a high but variable disturbance rate from wildfire, disease, and insects. Thus there is interest in using remote sensing information to reduce lag bias in estimates of current forest condition. Combining remote sensing and other ancillary data with field plots has become common for improving forest inventory information (Tomppo et al. 2008)

Imputation to combine field and remote sensing data is often used for mapping or small area estimation (e.g., Katila and Tomppo 2002, Ohmann and Gregory 2002, Finley and McRoberts 2008), and these methods typically impute point-in-time plot attributes. In contrast, when using imputation to update paneled inventory data, it is possible to impute mean annual change (MAC) for a plot. For example Arner et al. (2004) estimated mean annual net volume change using MA approaches and sampling with partial replacement approaches and McRoberts (2001) imputed the difference in basal area between two measurements to plots with missing measurements to update basal area for plots measured in previous years to a current point in time. There are

few studies for the western US that examine alternatives to the MA (e.g., Chapter 2 (Eskelson et al. in press)) and none that we are aware of that impute MAC rather than point-in-time measurements.

The objectives of this study are to: 1) use paneled data from a study area in the western United States to estimate mean annual change (MAC) of forest attributes using three nearest neighbor imputation methods; and 2) to estimate current forest attributes from paneled inventory data by updating the most recent measurement with imputed MAC. The results are compared with the estimates obtained from the MA estimator and the data from the current panel.

Methods

Data

In this study, 618 plots from six national forests that were collected as part of the Pacific Northwest Region's Current Vegetation Survey (CVS) of the US Forest Service were used. The particular national forests, sampled between 1993 and 1997 and remeasured in 2000, were the Colville (28), Mt. Hood (111), Ochoco (82), Rogue River (70), Wallowa-Whitman (199), and Winema (128) (Table 4.1).

Five plots were installed in each basic CVS sampling unit (1 ha in size) with each plot consisting of three permanent circular, nested subplots of different sizes. For a detailed description of the CVS inventory see Max et al. (1996). Live trees with diameter at breast height (DBH in cm) of 12.7 or larger were used in this study and height models developed in Barrett (2006) were employed to fill in missing heights

(HT in m). Gross cubic-meter volume and total gross oven dry weight biomass were calculated with volume and biomass equations from the US Forest Service (USDA 2000). For each plot, basal area in m^2 per ha (BA), stems per ha (SPH), volume in m^3 per ha (VOL), and biomass in tons per ha (BIOT) were calculated and summarized (Table 4.2). MAC for BA, SPH, VOL, and BIOT were calculated by dividing the difference of the observed values in 2000 and the observed value at the occasion one measurement by the growth period length (GPL) between the two measurements.

Thematic Mapper (TM) images from 2000 were used as ancillary data. The raw imagery bands 1 to 5 and band 7 (TM1, TM2, TM3, TM4, TM5, TM7), the Tasseled Cap (TC) transformations of the 6 axes (Kauth and Thomas 1976) as well as percent canopy cover were extracted from the national land-cover database 2001 (Homer et al. 2004). Three commonly used band ratios (band 4 to band 3, band 5 to band 4, and band 5 to band 7) and the normalized difference vegetation index (NDVI) were calculated.

In addition to the satellite data, the following climate and topography variables were used as ancillary data (Table 4.2): Annual precipitation and mean annual temperature [Data source: DAYMET Daily Surface Weather Data and Climatological Summaries (Thornton et al. 1997, Thornton and Running 1999)], Elevation (EL in m) and its transformations EL^2 and ln(EL) [Data source: CVS inventory], and slope (%) and aspect (degrees) and transformations (cosine(aspect), sine(aspect), cosine(aspect)*slope, sine(aspect)*slope) [Data source: 30 m digital elevation model

using Arc Workstation GRID surface functions and commands (Environmental Systems Research Institute 1991)].

Nearest Neighbor Imputation

Nearest neighbor (NN) imputation methods are donor-based methods. Variables of interest are those forest attributes that are only measured on a subset of plots (e.g., MAC of BA, SPH, VOL, and BIOT). Ancillary variables are the attributes that are measured on all plots. In this study, satellite, climate, and topography data as well as the most recent measurements of BA, SPH, VOL, and BIOT that were taken at measurement occasion 1 in the years 1993 to 1997 constitute the available ancillary data. Reference data are the plots for which both variables of interest and ancillary variables are available. Target data are the plots for which only the ancillary variables are available. The reference plots constitute the pool of potential plots which could be selected to impute the MAC data for the target plots.

The most similar neighbor (MSN) method (Moeur and Stage 1995), the gradient nearest neighbor (GNN) method (Ohmann and Gregory 2002), and the randomForest (RF) method (Crookston and Finley 2008) have been shown to provide reasonable imputation results for forest attributes (Moeur and Stage 1995, LeMay and Temesgen 2005, Hudak et al. 2008) and for mapping forest composition and structure (Ohmann and Gregory 2002, Ohmann et al. 2007). MSN, GNN, and RF were conducted using the yaImpute R package version 1.0-6 (Crookston and Finley 2008).

The similarity between reference and target plots is defined using a weighted Euclidean distance for MSN and GNN:

[4.1] $$D_{ij}^{\ 2} = (X_i - X_j) W (X_i - X_j)'$$

where W is the weight matrix, X_i is a vector of standardized values of the ancillary variables for the i^{th} target plot; and X_j is a vector of standardized values of ancillary variables for the j^{th} reference plot. The ancillary variables for both target and reference plots were standardized using the mean and variance of the ancillary variables of the reference plots. For MSN, the weight used is $W = \Gamma \Lambda^2 \Gamma'$, where Γ is the matrix of standardized canonical coefficients for the ancillary variables and Λ^2 is the diagonal matrix of squared canonical correlations between ancillary attributes and variables of interest (Moeur and Stage 1995).

For the gradient nearest neighbor method (GNN) the weights are assigned by employing a projected ordination of the ancillary data based on canonical correspondence analysis (CCA) (Ohmann and Gregory 2002).

The RF method is an extension of classification and regression tree (CART) methods (Breiman 2001). The data and variables are randomly and iteratively sampled to generate a forest of classification and regression trees. If two plots tend to end up in the same terminal nodes in a forest of classification and regression trees, they are considered to be similar. The RF distance measure is one minus the proportion of trees where a target plot is in the same terminal node as a reference plot (Crookston and Finley 2008, Hudak et al. 2008).

Estimation procedures

Panel data is a special case of inventory data with measurements taken at different times. A panel system was imitated with the available data by randomly assigning 25% of the plots to *P4* (154 plots) and the remaining 75% of the plots (464 plots) to *P1*, *P2*, and *P3* based on their year of installation (Table 4.1).

BA, SPH, VOL, and BIOT were the variables of interest (*Y*). The observed mean values of *Y* in 2000 were used as the best available estimate of the true mean:

$$[4.2] \quad \overline{Y}_{OBS} = \sum_{i=1}^{n} Y_i / n$$

where Y_i is the observed *Y* value of the ith plot in 2000 and $n = 618$.

Using the data from *P4* only, the mean values of *Y* for the year 2000 (SAMPLE25 estimator) were calculated as:

$$[4.3] \quad \overline{Y}_{SAMPLE25} = \sum_{i:Y_i \in P4} Y_i / n_4$$

where Y_i is the observed *Y* value of the i^{th} plot, and n_4 is the number of plots in *P4*.

The MA estimator, the FIA default method, is:

$$[4.4] \quad \overline{Y}_{MA(4)} = 0.25 * \overline{Y}_{t-3,i} + 0.25 * \overline{Y}_{t-2,i} + 0.25 * \overline{Y}_{t-1,i} + 0.25 * \overline{Y}_{t,i}$$

where $\overline{Y}_{t-3,i}$, $\overline{Y}_{t-2,i}$, $\overline{Y}_{t-1,i}$, and $\overline{Y}_{t,i}$ are the mean values of the variables of interest of *P1*, *P2*, *P3*, and *P4*, respectively. This MA(4) estimator will be referred to as MA in the following.

Instead of filling in the missing values for panels 1 to 3 with their previous measurements, as was done in the MA calculation, MSN, GNN, and RF were explored

to impute the MAC of the variables of interest. The imputed MAC was then used to update the variables of interest for *P1*, *P2*, and *P3* to the year 2000 as follows:

[4.5] $$Y_{imp,i} = Y_{t,i} + MAC_{imp,i} * GPL_i$$

where $Y_{imp,i}$ is the imputed Y value for the ith plot in 2000 and $Y_{t,i}$ is the observed Y value for the i^{th} plot at time t, which is 1993/1994, 1995, and 1996/1997 for *P1*, *P2*, and *P3*, respectively. $MAC_{imp,i}$ is the imputed MAC for the i^{th} plot with imp referring to the NN imputation method used. GPL_i is the growth length period between the initial measurement (1993-1997) and 2000.

The overall mean of the variables of interest for the year 2000 is then estimated as follows:

[4.6] $$\overline{Y}_{IMP} = \left[\sum_{i:Y_i \in P1} Y_{imp,i} + \sum_{i:Y_i \in P2} Y_{imp,i} + \sum_{i:Y_i \in P3} Y_{imp,i} + \sum_{i:Y_i \in P4} Y_{t,i} \right] / n$$

where IMP refers to the NN imputation method used and $Y_{imp,i}$ is the imputed Y value for the i^{th} plot as described in Equation 4.5.

Two sets of ancillary variables were tested for the imputation methods: The first set included the available climate, topography, and satellite data (Data set A) and the second set consisted of the previous measurements of the variables of interest that were taken at measurement occasion 1 from 1993 to 1997 (Data set B: BAocc1, SPHocc1, VOLocc1, BIOTocc1).

SAMPLE25, MA, and the three imputation methods were compared based on the overall means of the variables of interest in 2000 (see Equations 4.3, 4.4, and 4.6). The basis of evaluation was accuracy, as expressed by the root mean square error

(RMSE), and bias, calculated as the mean difference between the estimates and the observed mean values (Equation 4.2) from m = 500 iterations of randomly splitting the data. Both RMSE and bias were expressed as percent of the observed mean for each variable of interest:

[4.7] $$Bias\ \% = \frac{\sum_{i=1}^{n} \frac{(est_i - obs_i)}{m}}{\frac{\sum_{i=1}^{n} obs_i}{m}} * 100$$

[4.8] $$RMSE\ \% = \frac{\sqrt{\sum_{i=1}^{n} \frac{(est_i - obs_i)}{m}}}{\frac{\sum_{i=1}^{n} obs_i}{m}} * 100$$

The MAC estimates based on MSN, GNN, and RF imputation were also compared with RMSE and bias calculated from m = 500 iterations of randomly splitting the data.

Results

MSN imputation provided similar results for MAC estimates for both sets of ancillary variables with the variance contributing most to the RMSE. When BAocc1, SPHocc1, VOLocc1, and BIOTocc1 were used as ancillary variables, the estimates were slightly biased but had smaller RMSE values than the virtually unbiased estimates based on climate, topography, and satellite data (Table 4.3). GNN estimates of MAC had large bias (> 66%), which contributed most to the RMSE. Using climate,

topography, and satellite data as ancillary variables resulted in smaller bias and hence, smaller RMSE values than using BAocc1, SPHocc1, VOLocc1, and BIOTocc1 as ancillary variables (Table 4.3). In terms of RMSE, RF using climate, topography, and satellite data as ancillary variables provided the best estimates of MAC. RF using BAocc1, SPHocc1, VOLocc1, and BIOTocc1 as ancillary data provided virtually unbiased estimates for three of the four variables of interest and smaller RMSE values than those achieved by either MSN or GNN imputation (Table 4.3).

For estimating current BA and SPH the RMSE values of MA were about half the size of those observed for SAMPLE25. For estimating current VOL and BIOT the RMSE values for MA were about a third of those observed for SAMPLE25. SAMPLE25 results were virtually unbiased. MA results were precise but biased and the bias contributed most to the RMSE. The opposite was true for the virtually unbiased SAMPLE25 estimates, where the variance contributed most to the RMSE (Table 4.4).

MSN resulted in estimates of current BA, SPH, VOL, and BIOT that showed negligible bias for both sets of ancillary variables, with the bias being smaller when climate, topography, and satellite data were used as ancillary variables. The RMSE values were smaller when BAocc1, SPHocc1, VOLocc1, and BIOTocc1 were used as ancillary variables. For both sets of ancillary variables, the MSN estimates outperformed the MA estimates in terms of both bias and RMSE (Table 4.4).

For the GNN estimates of current forest attributes, bias contributed most to RMSE. When BAocc1, SPHocc1, VOLocc1, and BIOTocc1 were used as ancillary

variables, the bias and hence the RMSE was larger than for the ancillary variable set including climate, topography, and satellite data. For both sets of ancillary variables, bias and RMSE were larger than for any of the other estimators (Table 4.4).

For both sets of ancillary variables, the RF estimates exhibited negligible bias, with the bias being smaller for BA and SPH when climate, topography, and satellite data were used as ancillary variables and the bias for VOL and BIOT being smaller when BAocc1, SPHocc1, VOLocc1, and BIOTocc1 were used as ancillary variables. RMSE values were smallest when climate, topography and satellite data were used as ancillary variables. For both sets of ancillary variables, RF imputation outperformed the MA estimates both in terms of bias and RMSE (Table 4.4).

RF using climate, topography, and satellite data as ancillary variables provided the best results overall in terms of bias and RMSE for estimating current BA, SPH, VOL, and BIOT, followed by MSN and RF using BAocc1, SPHocc1, VOLocc1, and BIOTocc1 as ancillary variables (Table 4.4).

In a previous study, RF imputation using BAocc1, SPHocc1, VOLocc1, and BIOTocc1 as ancillary data for directly imputing current BA, SPH, VOL, and BIOT provided the most accurate estimates compared to the SAMPLE25 and MA estimators and MSN and GNN imputation. The results of this plot-level RF imputation are provided in Table 4. See Chapter 2 (Eskelson et al. in press) for details. The MSN and RF results for estimating current BA, SPH, VOL, and BIOT in this study outperform the plot-level RF imputation from the previous study in terms of both bias and RMSE (Table 4.4).

Discussion

After the start of the second inventory cycle, MAC can be estimated using remeasured plots (Arner et al. 2004). Considering the year 2000 as the start of the second inventory cycle, MAC was estimated in this study. The results of the three NN imputation methods that were explored to impute MAC of forest attributes showed the same pattern that was observed in an earlier study where the same three NN imputation methods were used to impute BA, SPH, VOL, and BIOT (see Chapter 2 (Eskelson et al. in press)). RF imputation provided the best results in terms of RMSE followed by MSN. The results of this study suggest that GNN plot-level imputation is not adequate to impute MAC of forest attributes. This might be due to the fact that the CCA in the GNN procedure requires the use of environmental factors for the ordination (Ohmann and Gregory 2002) which might not be picked up in the ancillary data that was used for the imputation.

The performance of using MAC of forest attributes of one inventory cycle to predict MAC for the next inventory cycle could not be tested in this study since only one remeasurement of the plots was available for the CVS data. This is an area of research that should be pursued as soon as multiple remeasurements of the FIA annual inventory are available in the western US.

When BA, SPH, VOL, and BIOT were updated to the year 2000 using MAC estimates obtained from MSN and RF imputation, the estimates of the mean BA, SPH, VOL, and BIOT in 2000 outperformed those of the SAMPLE25 and MA estimators in terms of accuracy. These results indicate that updating the variables of interest for

unmeasured plots to the current point in time using estimated MAC from MSN or RF imputation should be preferred over using the SAMPLE25 or MA estimators for estimating the current forest attributes.

The estimates of mean BA, SPH, VOL, and BIOT in 2000 using MAC estimates obtained from RF and MSN imputation also outperformed the estimates from RF plot-level imputation that was used to directly impute BA, SPH, VOL, and BIOT in 2000 (see Chapter 2 (Eskelson et al. in press)). This is due to the fact that the approach using the imputed MAC estimates makes use of the previously observed measurements. Adding a multiple of estimated MAC to the previously observed measurement will result in current estimates that will be close to the actual values even if the estimated MAC values were not perfect. If the current BA, SPH, VOL, and BIOT are imputed directly as was done in Chapter 2 (Eskelson et al. in press), the imputed values can be either close to the actual values or they can be completely different.

Conclusions

MSN and RF imputation provided adequate estimates of MAC of forest attributes, whereas GNN imputation should not be used to impute MAC of forest attributes with the ancillary data that was available in this study.

Updating previously observed measurements of forest attributes with imputed MAC estimates resulted in estimates of mean BA, SPH, VOL, and BIOT for the year

2000 that outperformed the estimates of SAMPLE25 and MA estimators in terms of accuracy.

Updating previously observed measurements of forest attributes with imputed MAC estimates also outperformed imputing BA, SPH, VOL, and BIOT for the year 2000 directly using RF imputation and should therefore be preferred.

Table 4.1: Number of plots measured by year of installation and corresponding panel assignment. All plots listed were remeasured in 2000.

Year of Installation	# of Plots	Assigned Panel
1993	7	1
1994	229	1
1995	223	2
1996	158	3
1997	1	3

Table 4.2: Summary of plot-level variables in 2000.

Variable	Minimum	Mean	Maximum	Std
Basal area (m^2/ha)	0.24	24.32	105.35	19.00
SPH (stems/ha)	1	305	1517	221
Volume (m^3/ha)	0.66	224.82	1444.74	221.04
Biomass (tons/ha)	0.58	134.09	800.64	132.64
Canopy cover (%)	0	54	97	29
Slope (%)	0	23	83	17
Elevation (m)	274	1389	2377	321
Annual precipitation (ln cm) (scaled * 100)	577	683	817	48
Annual temperature (ºC) (scaled * 100)	60	579	1067	166

Table 4.3: Bias and RMSE of mean annual change of the variables of interest BA (basal area/ha), SPH (stems/ha), VOL (volume/ha), and BIOT (biomass/ha). Data set A comprised climate, topography, and satellite data. Data set B comprised occasion 1 measurements of the variables of interest.

		Mean annual change in BA		Mean annual change in SPH		Mean annual change in VOL		Mean annual change in BIOT	
Method	**Data**	**Bias%**	**RMSE%**	**Bias%**	**RMSE%**	**Bias%**	**RMSE%**	**Bias%**	**RMSE%**
MSN	**A**	0.7	21.66	-3.09	41.91	-0.03	25.58	0.15	23.93
GNN	**A**	69.18	80.35	156.68	172.18	72.75	85.25	66.74	79.67
RF	**A**	-1.78	15.98	-2.18	35.18	-4.68	19.5	-4.74	17.87
MSN	**B**	1.6	21.37	-0.49	38.76	4.18	24.48	3.04	23.02
GNN	**B**	98.34	136.74	297.26	475.71	88.54	121.62	95.48	131.76
RF	**B**	-0.33	20.67	-2.5	38.58	0.28	23.92	-0.65	22.36

Table 4.4: Bias and RMSE of mean BA (basal area/ha), SPH (stems/ha), VOL (volume/ha), and BIOT (biomass/ha) in year 2000. Data set A comprised climate, topography, and satellite data. Data set B comprised occasion 1 measurements of the variables of interest.

		BA		SPH		VOL		BIOT	
Method	**Data**	**Bias%**	**RMSE%**	**Bias%**	**RMSE%**	**Bias%**	**RMSE%**	**Bias%**	**RMSE%**
SAMPLE25		-0.02	5.29	0.13	5.05	-0.08	6.59	-0.06	6.67
MA		-2.54	2.60	-2.63	2.68	-1.92	2.06	-1.98	2.12
MSN	A	0.17	1.56	0.02	2.01	0.05	1.94	0.11	1.90
GNN	A	5.45	6.25	7.90	8.67	6.03	7.02	5.82	6.88
RF	A	-0.04	1.12	0.01	1.70	-0.34	1.42	-0.33	1.35
MSN	B	0.33	1.47	0.25	1.87	0.48	1.77	0.44	1.73
GNN	B	7.66	10.38	15.02	23.77	7.16	9.76	8.16	11.10
RF	B	0.29	1.47	0.27	1.88	0.29	1.77	0.25	1.72
Plot-level RF	B	-0.30	1.58	-0.85	2.78	-0.06	1.90	-0.09	1.79

CHAPTER 5: ESTIMATING CAVITY TREE AND SNAG ABUNDANCE USING NEGATIVE BINOMIAL REGRESSION MODELS AND NEAREST NEIGHBOR IMPUTATION METHODS

Bianca N.I. Eskelson

Tara M. Barrett

Temesgen Hailemariam

Submitted to: Canadian Journal of Forest Research

Abstract

Cavity tree and snag abundance data are highly variable and contain many zero observations. This study predicts cavity tree and snag abundance from variables that are readily available from forest cover maps or remotely sensed data using negative binomial (NB), zero-inflated NB (ZINB), and zero-altered (ZANB) regression models as well as nearest neighbor (NN) imputation methods. The models are developed and fit to data collected by the Forest Inventory and Analysis (FIA) program of the US Forest Service in Washington, Oregon, and California. All three NB regression models provided reasonable results and outperformed the NN imputation methods.

Introduction

In the past decades, traditional timber-oriented forest management has broadened to commodity production while managing forest resources in an ecologically sustainable manner. Derived benefits include managing forest for wildlife, enhancing biodiversity, and protecting water quality.

Snags (standing dead trees) are a significant structural component of many forest ecosystems (Harmon et al. 1986). They create nesting, foraging, and roosting habitat for a variety of wildlife species that depend on snags and large trees for survival and reproduction (Bate et al. 1999, Russell et al. 2006, Wisdom and Bate 2008). Snags are important for the maintenance of biodiversity (Shorohova and Tetioukhin 2004, Aakala et al. 2008) as many dead wood dependent organisms are confined to snags during their life cycle (Nilsson et al. 2002). Snags also contribute to

ecological processes and decay dynamics (Ganey and Vojta 2005). Episodic events (e.g., insect outbreaks, fire, snow- and wind-caused stem breakage) create large quantities of snags. Small-scale mortality caused, for example, by competition or suppression, continuously creates smaller quantities of snags (Aakala et al. 2008).

Cavity trees contribute to diverse forest structure and wildlife habitat (Temesgen et al. 2008). Cavity trees are live trees or snags that contain a hole that provides wildlife species with shelter from the elements and protection from disturbance by predators and competitors (Carey 1983). Cavity trees provide many birds, mammals, reptiles, and amphibians with habitat for nesting, roosting, loafing, hibernating, and eating. They also provide escape cover and food storage locations (Carey 1983, Jensen et al. 2002).

Cavity-nesting birds and other wildlife species depend on an adequate and continuous supply of cavity trees and snags (Fan et al. 2003a). Timber harvest and human access can have substantial effects on snag density (Wisdom and Bate 2008). Because of the removal of cavity trees and snags under intensive timber management, the availability of cavity trees is a concern in resource management and conservation (Fan et al. 2004). The maintenance of snags in suitable abundance and stages of decay is critical to the preservation of biodiversity and the sustained functioning of forest ecosystems (DeLong et al. 2008).

Managers need to understand the nature of the cavity resource and the patterns of abundance of cavity trees in order to effectively manage forest resources for ensuring viable populations of cavity-using wildlife (Carey 1983). Regional

summaries of current amounts of dead wood, including snags and down woody debris, are needed for broad-scale assessment of wildlife habitat (Ohmann and Waddell 2002). Snag abundance is frequently used to incorporate habitat requirements of cavity-nesting wildlife into management plans. Snag abundance, however, does not take into account live cavity trees (Allen and Corn 1990). Although the proportion of stems with cavities is often at least twice as high for snags as for live trees (e.g., Goodburn and Lorimer 1998, Fan et al. 2003b, Temesgen et al. 2008), cavities tend to be more common in live trees because live trees are more abundant than snags (Goodburn and Lorimer 1998). Hence, cavity tree abundance, which considers both live trees and snags, should be used as an indicator in wildlife habitat models or to formulate wildlife tree retention in management plans.

Cavity tree abundance is highly variable even among forest stands that are similar in many other respects (Fan et al. 2004). This is due to the fact that cavity development is a relatively rare event governed by stochastic processes such as fire, insect attack, disease, and mechanical or chemical injury that can lead to tree death or injury (Carey 1983). Tree characteristics and stand attributes such as size, decay class, and species only play a partial role in cavity tree development (Fan et al. 2003a).

Recent studies have estimated cavity tree abundance at the stand level. For example, Temesgen et al. (2008) used nearest neighbor imputation and classification and regression tree (CART) methods to estimate cavity tree abundance. Fan et al. (2003a) estimated cavity tree abundance by stand age and basal area using CART, and described the cavity tree density distribution within a cluster using the Weibull

probability density function. They found that the proportion of stands with cavity trees increases with increasing stand age and increasing basal area. Fan et al. (2005) quantified the frequency and size distribution of cavity trees in seedling/sapling, pole, sawtimber, and old-growth stands based on plot data. Fan et al. (2004) simulated cavity tree dynamics under alternative harvest regimes. Most other studies concentrated on the distribution of cavity trees or snags on the individual tree or species level. Fan et al. (2003b) explored factors associated with cavity tree abundance and developed models that can be used to predict relative frequency of cavity trees based on tree size, species, and decay class. Carey (1983) found tree diameter measured at 1.37 m above the ground (DBH) and site index to be good indicators for cavity tree abundance.

In the states of California, Oregon, and Washington, information on cavity tree and snag occurrence along with other information such as species, DBH, and height of individual trees is collected as part of the national inventory of public and private forests (the Forest Inventory and Analysis [FIA] Inventory). The FIA inventory uses an interpenetrating panel design with 10 panels in the western states, where all plots located in one of the 10 interpenetrating panels (10%) are measured each year (Brand et al. 2000). In order to estimate current cavity tree and snag abundance from paneled inventory data, the information on cavity tree and snag abundance in the current year needs to be updated for all unmeasured panels. It is of interest to be able to do this with variables that are readily available from forest cover maps or remotely sensed data (e.g., aerial photographs, satellite data, LiDAR). These variables will collectively

be referred to as map label variables in this study. If cavity tree and snag abundance can be predicted from map label variables, accurate, spatially comprehensive, current, and very detailed information on cavity tree and snag abundance could also be provided to managers and planners interested in assessing wildlife habitat of their forests (Temesgen et al. 2008).

Count distributions are useful to describe non-negative integer values such as the number of snags or cavity trees per plot. Poisson regression is the basic count model upon which a variety of other count models are based. The Poisson distribution assumes equidispersion which means that the mean and variance are equal. The development of more general count models such as the negative binomial (NB) distribution, which do not assume equidispersion, has been driven by the fact that equality of mean and variance is rarely found in natural resource data. A variety of NB regression models has been developed to accommodate additional violations of distributional assumptions, such as no zeros or excess zeros in the data, which often occur in natural resource data (Hilbe 2007, p. 8-10).

The objectives of this study are 1) to compare the suitability and predictive abilities of negative binomial regression models to estimate snag and cavity tree abundance using map label variables; and 2) to use distribution-free nearest neighbor (NN) imputation methods to impute snag and cavity tree abundance and compare the imputation results with the results of NB regression models.

Negative binomial regression models

Count regression models are a subset of discrete response regression models and aim to explain the number of occurrences or counts of an event. Poisson regression is the basic count model and the Poisson distribution is characterized as

$$[5.1] \quad P(Y = y) = \frac{\mu^y e^{-\mu}}{y!}$$

where $y = 0,1,2,\ldots$, $\mu > 0$, the random variable y is the count response, and the parameter μ is the mean. A Poisson regression model is obtained by relating the mean μ to a vector of explanatory variables, $\boldsymbol{x}$, by $\mu = e^{x^T\beta}$, where $\boldsymbol{\beta}$ is a vector of regression coefficients to be estimated.

A consequence of the Poisson probability mass function (Equation 5.1) is that the mean and variance are equal, that is $Var[Y \mid x] = E[Y \mid x] = \mu$. When data do not fit the Poisson distribution, it is typically because of overdispersion, meaning the variance of the model exceeds the value of the mean. The NB distribution, which can be derived as a gamma mixture of Poisson distributions, employs an extra parameter α that directly addresses the overdispersion in the Poisson models. The NB distribution is characterized as

$$[5.2] \quad P(Y = y) = \frac{\Gamma(y + 1/\alpha)}{\Gamma(y+1)\Gamma(1/\alpha)}\left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y}$$

where $y = 0,1,2,\ldots$ and $\mu > 0$. α represents the degree of overdispersion. The mean is μ, the same as the Poisson, but the variance is $\mu + \alpha\mu^2$ thus allowing the mean to exceed

μ (Hilbe 2007, p. 78, 80). NB regression models are obtained in the same way as Poisson regression models by relating the mean μ to a vector of explanatory variables, $\boldsymbol{x}$, by $\mu = e^{x^T\beta}$.

Although the NB model is more flexible than the Poisson, there is no guarantee that it provides an adequate model for the count data. Excess zeros or no zeros in the data violate the distributional assumptions that apply equally to the Poisson and NB distributions. Other possible violations of the distributional assumptions occur when the data contain censored or truncated observations, or when the data are structured as panels (e.g., clustered and longitudinal data) (Hilbe 2007, p. 11-13).

Zero-inflated Poisson (ZIP) and zero-inflated NB (ZINB) regression models have been developed to account for data with a high percentage of zero counts (Lambert 1992, Welsh et al. 1996). Zero-inflated models are mixture models combining a count distribution with a point mass at zero. In zero-inflated models there are two sources of zeros: zeros come from either the count distribution or from the point mass (Lambert 1992, Hall 2000). The ZINB model is defined as follows:

[5.3] $$P(Y=y)=\begin{cases} \pi+(1-\pi)*f_{count}(0;x,\beta) & if\ y=0 \\ (1-\pi)*f_{count}(y;x,\beta) & if\ y=1,2,.... \end{cases}$$

where $\pi = f_{zero}(0;z,\gamma)$ is the probability of belonging to the point mass component and $(1-\pi)$ is the probability of belonging to the count distribution. $\boldsymbol{z}$ is a vector of explanatory variables used in the logistic model and $\boldsymbol{\gamma}$ is a vector of regression coefficients to be estimated. $f_{count}(y;x,\beta)$ corresponds to the NB probability function

given in Equation 5.2 and $f_{count}(0;x,\beta) = \left(1-(1+\alpha\mu)^{-(1/\alpha)}\right)$ where α represents the degree of overdispersion and μ is related to a vector of explanatory variables, $\boldsymbol{x}$, by $\mu = e^{x^T\beta}$, where $\boldsymbol{\beta}$ is a vector of regression coefficients to be estimated.

Another approach for dealing with excess zeros in the data is to model the response as having two states: a state in which no cavity trees or snags occur and a state in which cavity trees or snags occur with varying levels of abundance (Welsh et al. 1996). The first state, a binary process which generates positive versus zero counts, is modeled applying logistic regression. Given that cavity trees are observed, the number of cavity trees (the second state) can be modeled by a zero-truncated Poisson (ZTP) or zero-truncated NB (ZTNB) distribution. The process generating positive counts only commences after crossing a zero barrier or hurdle. The combined models are known as conditional models (Welsh et al. 1996) or are referred to as Poisson and NB hurdle models or as zero-altered Poisson (ZAP) and zero-altered NB (ZANB) models (Hilbe 2007). In the NB case, the combined regression model is defined as follows:

[5.4] $$P(Y=y) = \begin{cases} f_{zero}(0;z,\gamma) & if\ y=0 \\ (1-f_{zero}(0;z,\gamma)) * f_{zt}(y;x,\beta) & if\ y=1,2,.... \end{cases}$$

where $f_{zero}(0;z,\gamma)$ is the probability of a zero count and $(1-f_{zero}(0;z,\gamma))$ is the probability of overcoming the hurdle. $\boldsymbol{z}$ is a vector of explanatory variables used in the logistic model and $\boldsymbol{\gamma}$ is a vector of regression coefficients to be estimated.

$f_{zt}(y;x,\beta) = \frac{f_{count}(y;x,\beta)}{1 - f_{count}(0;x,\beta)}$ is a ZTNB model, with $f_{count}(y;x,\beta)$ and $f_{count}(0;x,\beta)$ defined as above. All observations are used to fit $f_{zero}(0;z,\gamma)$, treating positive counts as 1's in the logistic regression framework. The data are separated into two subsets, using only data with positive counts to fit $f_{zt}(y;x,\beta)$. For more details on ZAP and ZANB models see Cameron and Trivedi (1998, p.123-128; 2005, p. 544-546, p. 680, 681).

Nearest neighbor imputation methods

NN imputation approaches are donor-based methods where the imputed value is either a value that was actually observed for another item or unit or the average of values from more than one item or unit. These donors can be determined in a variety of ways. Forest attributes that are measured on all units are referred to as *X*-variables. *Y*-variables are those forest attributes that are only measured on a subset of units—in this case cavity tree and snag abundance. Units with measured *X*- and *Y*-variables are called reference data and target data are those units that only have *X*-variables measured. The similarity between target and reference data is determined with a distance metric defined in the feature space of the *X*-variables (LeMay and Temesgen 2005).

Methods

Data

Data for this study were obtained from the FIA inventory from Washington, Oregon, and California collected in the years 2001 to 2007. For details about the inventory see Bechtold and Patterson (2005). Each field plot is composed of a cluster of four points, with each point being composed of two nested fixed-radius plots (subplot and microplot) used to sample trees of different size (Bechtold and Scott 2005). Unique polygons (also called condition-classes) on the FIA plot are distinguished by structure, management history, or forest type. Only data collected on the subplots were used for this study and summarized by condition-classes. The data set contained 10,607 stands (or condition-classes) that covered a wide range of ground and map label variables (Tables 5.1 and 5.2).

Cavity presence was collected in the field by classifying each live tree or snag taller than 1.5 m and greater than 12.5 cm DBH into one of three categories: 1) no cavity present, 2) cavity greater than 15.2 cm diameter present, and 3) cavity less than 15.2 cm diameter but no larger cavities present. Cavity presence was only recorded for trees with cavities that could, in the field crew's judgment, be used by wildlife such as birds or mammals. Cavity tree abundance is assumed to be additive from individual trees in a stand and is quantified as the number of cavity trees (both live trees and snags) per stand without apportioning it by species or species groups. Cavity tree abundance can be assumed to be under-recorded, as field crews are more likely to miss cavities than record cavities that do not exist. Snag abundance is the number of

standing dead trees per stand, and snags only included dead trees that leaned less than 45 degrees from vertical.

While 66% of the cavities were observed in snags, 34% of the cavities were found in live trees. Live trees (88% of standing trees) are more abundant than snags (12% of standing trees), but only 0.86% of all live trees had cavities compared to 12.55% of the snags. Of the 10,607 stands, 2,796 and 6,293 contain cavity trees and snags, respectively, resulting in large numbers of zero counts for both cavity tree and snag abundance (Figure 5.1).

Average stand age was used to represent stand development stage. The midpoint of five height classes was used (Table 5.2). Slope, aspect, elevation and transformations of these three variables (Salas et al. 2008) as well as the midpoint of seven site classes (Table 5.2) represented site conditions. Percent of conifer basal area and very broad forest type groups described general species composition. Four forest type groups were used: 1) Douglas-fir (2549), 2) fir/spruce/mountain hemlock (1180), 3) other conifers (4324), and 4) hardwoods (2554). Four owner groups were distinguished: 1) Forest Service (4975), 2) other federal (1039), 3) state and local government (653), and 4) private (3940). The group of private forest owners includes corporations, non-governmental conservation and natural resources organizations, unincorporated local partnerships, associations and clubs, Native Americans, and individuals. The group of other federal forest owners includes the National Park Service, the Bureau of Land Management, the Fish and Wildlife Service, the Department of Defense/Energy, and other federal owners.

Negative binomial (NB) regression models

The data set was randomly split into modeling (75%, 7,955 stands) and validation (25%, 2,652 stands) data sets. The modeling data set was used to fit NB, ZINB, and ZANB models in R using quasi-Newton optimization methods (R Development Core Team 2008, Zeileis et al. 2008). Snag abundance and cavity tree abundance were used as response variables, respectively. Map label variables related to site, ownership, forest development stage, and general species composition were used as explanatory variables.

Explanatory variables that did not contribute significantly in explaining variation were dropped from the NB, ZINB, and ZANB models. Nested and non-nested models were compared using Akaike's information criterion (AIC; Akaike 1973, 1974) and Schwarz's Bayesian information criterion (BIC, Schwarz 1978):

[5.5] $$AIC = -2 * \ln(L) + 2 * p$$

[5.6] $$BIC = -2 * \ln(L) + p * \ln(n)$$

where p is the number of parameters that were estimated in the model, n is the number of observations in the modeling data set, and $\ln(L)$ is the natural logarithm of the log-likelihood of the model. The parameter estimates as well as the AIC and BIC values for the cavity tree and snag models are shown in Appendix B.1 and B.2, respectively.

The parameter estimates of the NB, ZINB, and ZANB models were used to predict cavity tree and snag abundance for the validation data set. In order to assess the adequacy of the models for predicting the overall counts of cavity trees and snags,

the chi-square goodness-of-fit test was calculated using observed and predicted counts from the validation data set of cavity trees and snags, respectively.

$$[5.7] \quad \chi^2 = \sum_{k=1}^{m} \frac{\left[\#\left(y_i = k\right) - \sum_i \Pr\left(y_i = k\right)\right]^2}{\sum_i \Pr\left(y_i = k\right)}$$

where # denotes the frequency of observations y_i in count class *k* across the data set and $\Pr\left(y_i = k\right)$ is the predicted probability of an observation to belong to count class *k*. This statistic is χ^2-distributed with (*m*-1) degrees of freedom since no model parameters were estimated from the validation data set. The number of count classes, *m*, was 14 and 78 for cavity tree and snag counts, respectively. For cavity trees the largest count class *m* = 14 included counts of 13 and up (13+). For snags the largest count class *m* = 78 included counts of 77 and up (77+). It is questionable whether the χ^2-statistic is reliable since many observed frequencies were either zero or smaller than 5. Hence, diagnostic plots, which plot the differences between predicted and observed probabilities against the count classes *k*, were used to detect any predictive bias and assess goodness-of-fit (Lambert 1992, Fortin and DeBlois 2007). The difference d_k between predicted probabilities and observed frequencies is computed as:

$$[5.8] \quad d_k = \sum_{i=1}^{n} \left(\frac{\Pr\left(y_i = k\right)}{n} \right) - \left(\frac{\#\left(y_i = k\right)}{n} \right)$$

where *n* is the number of observations in the validation data set and the rest is defined as above. As suggested by Fortin and DeBlois (2007), the sum of the absolute

differences d_k was defined as w and used as an index of the goodness-of-fit of the different NB regression models and the imputation methods.

Nearest Neighbor (NN) imputation

Two imputation methods were employed in this study using the yaImpute package in R (Crookston and Finley 2008):

1) MSN — Most Similar Neighbor (Moeur and Stage 1995)

2) RF — randomForest (Breiman 2001, Crookston and Finley 2008)

In the MSN procedure (Moeur and Stage 1995), the distance metric is of quadratic form:

[5.9] $$d_{ij}^2 = \left(X_i - X_j\right) W \left(X_i - X_j\right)'$$

where X_i is the (1 x p) vector of X-variables for the i^{th} observation unit, X_j is the (1 x p) vector of X-variables for the j^{th} reference observation unit, and W is a (p x p) symmetric matrix of weights. W is derived from canonical correlation analysis which makes use of the relationships between X- and Y-variables so that stronger correlations result in higher weights for a particular X (LeMay and Temesgen 2005).

The RF method is a classification and regression tree (CART) method (Breiman 2001). The data and variables are randomly and iteratively sampled to generate a large group, or forest, of classification and regression trees. For RF two units are considered similar if they tend to end up in the same terminal nodes in a forest of classification and regression trees. The distance measure is one minus the

proportion of trees where a target unit is in the same terminal node as a reference unit (Crookston and Finley 2008, Hudak et al. 2008).

Cavity tree and snag abundance were used as *Y*-variables, as well as square root (sqrt(*Y*)), inverse (1/(*Y*+1)), and logarithmic (ln(*Y*+1)) transformations of these variables. The X-variables used for the NN imputation are: average stand age, percent conifers, height class midpoints, site class midpoints, elevation (EL), EL^2, slope, slope*cosine(aspect), and slope*sine(aspect).

The 7,955 stands that were used as modeling data sets for the NB, ZINB, and ZANB models were randomly split into target (25%) and reference (75%) stands 200 times and nearest neighbors were imputed for each of the 200 target data sets using MSN and RF for each of the four *Y*-variable sets (original units and three transformations of cavity tree and snag abundance). The average difference between imputed and observed values (often called bias) and the root mean squared error (square root of the average squared difference; RMSE) were calculated as fit statistics to evaluate the results for each simulation:

$$[5.10] \quad bias = \frac{\sum_{i=1}^{n} \left(imputed_i - observed_i \right)}{n}$$

$$[5.11] \quad RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(imputed_i - observed_i \right)^2}{n}}$$

where *n* is the number of target stands.

Bias and RMSE were used to select three imputation approaches that were then used to compare NB regression models with imputation methods.

Comparison of NB regression models and NN imputation methods

To compare NB regression models with NN imputation methods on a per stand basis, the validation and modeling data sets used for the NB regression models are used for the NN imputation as target and reference data sets, respectively. The results of the NB regression models and the NN imputation are then visually compared by plotting histograms of the prediction errors (PE) that are computed as follows:

[5.12] $PE = predicted_i - observed_i$

where $observed_i$ are the observed cavity tree and snag counts for the i^{th} stand and $predicted_i$ are the predicted counts for the i^{th} stand from the NB regression models and the NN imputation methods. Positive and negative PE values indicate overestimation and underestimation, respectively.

Results

The bias (average difference) for cavity tree and snag abundance averaged over the 200 sampling replications was zero or close to zero for the original units and the three transformations when the nearest neighbor was imputed using MSN. RF imputation resulted in a negative mean bias of the 200 sampling replications for estimating cavity tree and snag abundance using the original units or any of the three transformations. The range of the bias of the 200 replications was slightly smaller for

the RF imputation than for the MSN imputation when the original units or either of the transformations was used (Table 5.3).

For MSN and RF, the RMSE for snag abundance averaged over the 200 sampling replications (mean RMSE) was smallest for the original units and the range of the RMSE was smallest for the logarithmic transformation for MSN imputation. Mean RMSE for cavity tree and snag abundance was smaller for RF than for MSN when the original units or either of the transformations was used as *Y*-variables. The range of the RMSE of the 200 replications was also smaller for the RF imputation for all four sets of *Y*-variables (Table 5.3).

MSN and RF imputation with the original units as *Y*-variables as well as MSN imputation with the logarithmic transformation as *Y*-variables were chosen to be compared to the NB regression models.

For estimating cavity tree abundance, the ZINB model had the largest χ^2-goodness-of-fit statistic with a value of 11.57 compared to values of 5.63 and 6.89 for the NB and ZANB models, respectively. The critical χ^2-statistic was 22.36 for the probability of a greater value = 0.05 and 13 degrees of freedom (no parameters were estimated from the validation data set). Since the χ^2-goodness-of-fit statistic of none of the three models exceeded the critical χ^2-statistic, there is no evidence that any of the three models are inadequate for estimating cavity tree abundance.

In addition, the diagnostic plots (Figure 5.2) were used to compare the models and to identify potential model misspecifications. Differences d_k were calculated as predicted probabilities minus observed proportions (Equation 5.7) so that positive

values indicate overestimations and negative values indicate underestimation. All three regression models underestimated zero counts and showed a good fit for counts of five and larger (Figure 5.2). The NB model had the lowest $w = 0.0279$.

Diagnostic plots for the MSN and RF imputation with the original units as *Y*-variables as well as MSN imputation with the logarithmic transformation as *Y*-variables showed that the RF imputation highly overestimated the zero cavity tree counts. MSN imputation with the original units and the logarithmic transformation resulted in an underestimation of zero counts. The *w* values indicated that the differences were closest to the reference line $d_k = 0$ for the MSN model with the original units as *Y*-variables, resulting in a *w* value of 0.0279 (Figure 5.2).

For estimating snag abundance, the NB model outperformed the ZINB and ZANB models according to the χ^2-goodness-of-fit statistic. The critical χ^2-statistic was 98.48 for the probability of a greater value = 0.05 and 77 degrees of freedom (no parameters were estimated from the validation data set). The χ^2-goodness-of-fit statistic of the ZANB model was slightly smaller than the critical χ^2-statistic which suggests that the model fit was still acceptable. The χ^2-goodness-of-fit statistic of the ZINB model greatly exceeded the critical χ^2-statistic which indicated that the model did not adequately characterize snag abundance.

The diagnostic plots (Figure 5.3) show that the NB and ZANB models underestimated the zero snag counts. The ZINB model estimated the zero counts well, however, it greatly underestimated $k = 1$ which caused the large *w* value. The NB model had the smallest *w* value indicating generally small differences d_k.

The diagnostic plots for the MSN and RF imputation with the original units as *Y*-variables as well as MSN imputation with the logarithmic transformation as *Y*-variables show that the differences d_k for small k values were farther away from the reference line $d_k = 0$ than those for the three NB regression models. RF imputation resulted in the largest d_k values highly overestimating the zero counts. MSN imputation with the logarithmic transformation had the smallest differences d_k resulting in the smallest w value of 0.0852 which still exceeded the w values of all three regression models (Figure 5.3).

The NN imputation methods had larger numbers of prediction errors fall between -0.5 and 0.5 than the NB regression models for the cavity tree counts. The RF imputation using the original units as *Y*-variables had the largest number of prediction errors (1611) and the ZINB model had the smallest number of prediction errors (1461) fall within this range. The prediction errors of the NN imputation methods covered the whole range of possible values between -13 and 13 which resulted in mean square prediction error (MSPE) values that were about twice as large as the MSPE observed for the NB regression models (Figure 5.4). The ZINB model had the smallest MSPE value of 1.25. None of the three NB regression models resulted in overpredictions larger than 3 counts. However, each model had one underprediction that exceeded 12 counts (Figure 5.4).

As for the cavity tree abundance on a per stand level, the NB regression models resulted in smaller MSPE values for the snag abundance with the ZINB model having the smallest MSPE value (19.88) (Figure 5.5). None of the NB regression

models overpredicted snag counts by more than 13 counts but for all three models the largest underprediction was around 54 counts. Most of the prediction errors fell into the range between 1 and 3 for the NB regression models, whereas most of the prediction errors of the NN imputation methods were between -1 and 1. RF imputation performed best among the NN imputation methods in terms of MSPE (Figure 5.5). The prediction errors for the RF imputation ranged between -55 and 77. For the MSN imputation using the original units as *Y*-variables the range was (-58, 70) and for the logarithmic MSN imputation the range was (-58, 60).

Discussion

The results of the simulation of the NN imputation methods indicated that RF generally performed better than MSN in terms of the range of bias and RMSE and the mean RMSE for predicting cavity tree and snag abundance. The RF method was employed in this study because it has been found to produce results that were generally superior to other NN imputation methods for predicting basal area and tree density by species (Hudak et al. 2008). The results of this study confirm the conclusion of Hudak et al. (2008) that RF imputation represents an alternative to traditional NN imputation methods.

Transformations on the *Y*-variables had been tested with the hope to improve the relationships between the *X*- and *Y*-variable sets. An improved relationship between *X*- and *Y*-variables could have had a positive impact on the canonical correlation analysis that is used in MSN imputation to determine the weight matrix *W*.

Temesgen et al. (2008) imputed cavity trees per hectare using MSN imputation and found that using the square root transformation of cavity trees per hectare as *Y*-variable improved the imputation results. It was assumed that the assumption about linear correlations was better met with the square root transformation (Temesgen et al. 2008). The ranges of cavity tree abundance (0-13) and snag abundance (0-77) were very small in this study due to using actual tree counts rather than the expanded tree per hectare values as *Y*-variable. It is suspected that none of the transformations have substantially improved the results of the MSN and RF imputation in terms of bias and RMSE due to the small observed ranges of the *Y*-variables.

Poisson, ZIP and ZAP models were fit to the data but did not provide adequate results due to the overdispersion that is present in the data. The expected values for zero as well as large cavity tree and snag counts were much too low for the Poisson model and were only somewhat improved by the ZIP and ZAP models (results not shown). Hence, this study focused on the application of NB regression models which allow for overdispersion. For predicting overall cavity tree and snag abundance the NB, ZINB, and ZANB models all fit reasonably well. However, the NB model resulted in the lowest χ^2-goodness-of-fit statistic and *w*. Since some of the counts have less than 5 observations, it was questionable if the χ^2-goodness-of-fit statistic provided reliable results, but its results were confirmed by the diagnostic plots and goodness-of-fit index *w*. For predicting cavity tree and snag abundance on a per stand basis, the ZINB model slightly outperformed the ZANB and NB models in terms of the MSPE value. However, the differences were only minor so that the use of the simpler NB

model seems preferable for predicting cavity tree and snag abundance. In addition, the difficulty of assigning biological meaning to the components of the ZINB and ZANB models raises the issue of overfitting (Affleck 2006) and interpretation of these models. If mechanisms could be identified that separate the conditions associated with zero cavity trees and snags from conditions associated with positive counts of cavity trees and snags, ZANB models would provide the advantage of modeling these two aspects separately (Welsh et al. 1996). However, in the given case, this property of the ZANB model is considered an unnecessary complication to the application. Another disadvantage of ZINB and ZANB models is that they provide a composite predictor that does not benefit from the path invariance property so that models like the NB model that allow for unobserved heterogeneity seem preferable for assigning further structure to the mean (Affleck 2006).

Cavity trees and snags are rare in forest ecosystems and thus more difficult to predict than many other forest attributes (Temesgen et al. 2008). The prediction of cavity tree and snag abundance is complicated by the fact that generally little association is found with environmental factors and stand-level attributes such as forest type, slope, aspect, and site index (Fan et al. 2003b). This is due to the fact that random processes such as fire, wind, and insect outbreaks play a major role in creating snags and cavities resulting in a large variability in cavity tree and snag abundance (Carey 1983). Hence, it will probably always be difficult to predict cavity tree and snag abundance from stand-level variables that are readily available from forest cover maps or remotely sensed data. However, the inclusion of variables which were not

available in this study could have potentially improved the results of the NB regression models.

Snag abundance generally increases with successional development (Ohmann and Waddell 2002). Fan et al. (2005) found mean cavity tree abundance to increase with increasing stand-size class expressed as seedling/sapling, pole, sawtimber, old-growth. In this study, average stand age was the only explanatory variable that was used to represent stand development stage. Including other variables such as stand-size class in the set of explanatory variables might improve the prediction of cavity tree and snag abundance. Timber harvest and human access can have substantial effects decreasing snag abundance in areas of intensive timber harvest and increased human access (Wisdom and Bate 2008). Explanatory variables that represent harvest history and degree of human access could potentially improve the prediction of cavity tree and snag abundance.

Formal tests that allow comparing parametric models such as the NB regression models with NN imputation methods do not exist. The diagnostic plots proposed by Lambert (1992) to detect model misspecifications in zero-inflated models as well as the goodness-of-fit statistic *w* introduced by Fortin and DeBlois (2007) provided efficient and convenient ways to show differences between observed and expected counts which could not only be used to detect model misspecification in the NB, ZINB, and ZANB models but also to compare the results of these models with those provided by the NN imputation methods. All NB regression models and NN imputation methods resulted in large differences between observed and expected

counts for counts of 5 and smaller. However, the differences between observed and expected counts decreased faster with increasing counts for the NB regression models than for the NN imputation models. This suggests that the NB regression models, in particular the NB model, should be preferred over the NN imputation methods to predict overall cavity tree and snag abundance.

Frequency histograms of the prediction and prediction errors were used to visualize and compare the predictions of cavity tree and snag counts per stand of the NB regression models and the NN imputation methods. The NB regression models resulted in a few large underpredictions but no large overpredictions of cavity tree and snag counts. Hence, the NB regression models tend to be more conservative in their predictions of cavity tree and snag abundance than NN imputation methods which result both in large over- and underpredictions. For management applications that take into account wildlife habitat it is better to base actions on models that are conservative with respect to overpredictions of cavity tree and snag abundance.

All cavity trees and snags were considered equally valuable in this study. Neither cavity size and cavity location on the tree nor decay stage and size of snags, which are important criteria in evaluating habitat quality for certain wildlife species, was taken into account. Hence, no inferences can be made about the quality or potential use of cavity trees and snags for wildlife species, even though the results of this study provided reasonable estimates of cavity tree and snag abundance. In order to be useful for management purposes, it will be necessary to use methods that allow the estimation of cavity tree and snag abundance while simultaneously providing

information on the size and location of cavities as well as the decay stage and size of snags. Taking advantage of the multivariate nature of the NN imputation methods, they could be used to simultaneously impute the abundance of cavity trees and snags as well as their quality attributes. This could be a major advantage of the NN imputation methods over the NB regression models. Other quality attributes that are important for certain wildlife species are tree species, percent bark cover, and presence of a broken top (Spiering and Knight 2005). Information on snag dynamics, such as longevity and the rates at which their quality changes, is also required to be able to fully take snags into account in forest management (Aakala et al. 2008).

Conclusions

NB, ZINB, and ZANB models provided reasonable results for predicting overall cavity tree and snag abundance as well as for predicting cavity tree and snag abundance per stand. The NB model should be preferred to the ZINB and ZANB models due to its easier application and simpler interpretation.

NB regression models performed better than NN imputation methods. For predicting cavity tree and snag abundance per stand, NB regression models should be preferred to NN imputation methods since they do not result in large overpredictions of the cavity tree and snag counts and hence provide more conservative results.

NN imputation methods might provide a tool for predicting cavity tree and snag abundance as well as their qualities in one step. The knowledge of cavity tree and snag quality will be necessary for evaluating wildlife habitat.

Acknowledgements

We gratefully acknowledge the support provided by the Forest Inventory and Analysis program, Pacific Northwest Research Station, United States Forest Service. We also thank Achim Zeileis for his support on the pscl R package and Vicente Monleon for his assistance in creating graphs in R.

Table 5.1: Descriptive statistics for stands, n=10,607.

Variable	Minimum	Mean	Median	Maximum	Std
Cavity tree counts	0	0.50	0	13	21.88
Snag counts	0	2.63	1	77	4.80
Percent Conifer	0	0.78	1	1	0.36
Average Stand Age (years)	0	92	75	1009	85
Elevation (m)	0	1031	1006	3366	682
Aspect (degrees)	0	158	155	360	115
Slope (%/100)	0	0.31	0.27	1.51	0.24

Table 5.2: Site class and height class descriptions and number of stands in each class.

Site class	**# stands**	**Height class**	**# stands**
15.7+ m^3/ha/year	221	0 – 9.99 m	229
11.6-15.6 m^3/ha/year	872	10 – 19.99 m	3963
8.4-11.5 m^3/ha/year	2046	20 – 29.99 m	2626
5.9-8.3 m^3/ha/year	1722	30 – 39.99 m	1112
3.5-5.8 m^3/ha/year	2293	40 – 49.99 m	432
1.4-3.4 m^3/ha/year	1865	50+ m	145
0-1.3 m^3/ha/year	1588		

Table 5.3: Minimum, mean, and maximum bias and RMSE for the Y-variables (cavity tree abundance and snag abundance) and the square root (sqrt), inverse, and logarithmic (ln) transformations of the Y-variables over 200 sampling replications. MSN and RF stand for the most similar neighbor and randomForest imputation methods, respectively.

		Cavity trees						Snags					
		Bias			RMSE			Bias			RMSE		
Method	**Response**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
MSN	**Y**	-0.10	0.00	0.10	1.41	1.55	1.68	-0.41	-0.01	0.46	5.81	6.64	7.62
MSN	**sqrt(Y)**	-0.12	0.00	0.10	1.38	1.55	1.70	-0.49	0.00	0.47	5.91	6.69	7.60
MSN	**1/(1+Y)**	-0.11	0.00	0.10	1.43	1.55	1.70	-0.41	-0.01	0.40	5.76	6.69	7.65
MSN	**ln(Y+1)**	-0.09	0.00	0.10	1.39	1.55	1.68	-0.39	0.01	0.49	5.96	6.66	7.56
RF	**Y**	-0.20	-0.13	-0.04	1.29	1.43	1.56	-0.70	-0.29	0.08	5.53	6.33	7.21
RF	**sqrt(Y)**	-0.21	-0.13	-0.05	1.29	1.43	1.57	-0.68	-0.27	0.06	5.61	6.39	7.13
RF	**1/(1+Y)**	-0.21	-0.12	-0.05	1.31	1.44	1.54	-0.66	-0.23	0.12	5.61	6.46	7.31
RF	**ln(Y+1)**	-0.21	-0.13	-0.04	1.30	1.43	1.55	-0.64	-0.26	0.09	5.63	6.42	7.13

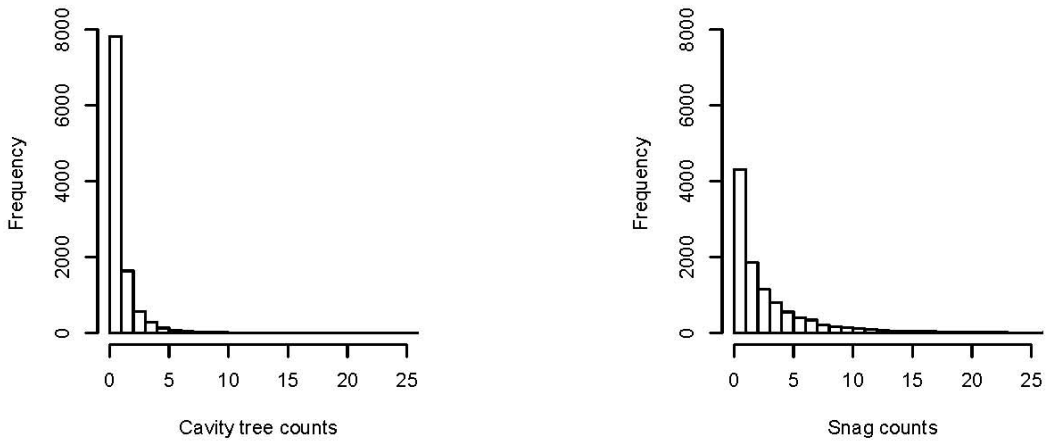

Figure 5.1: Frequency distribution of stands with up to 25 counts of cavity trees (left) and snags (right).

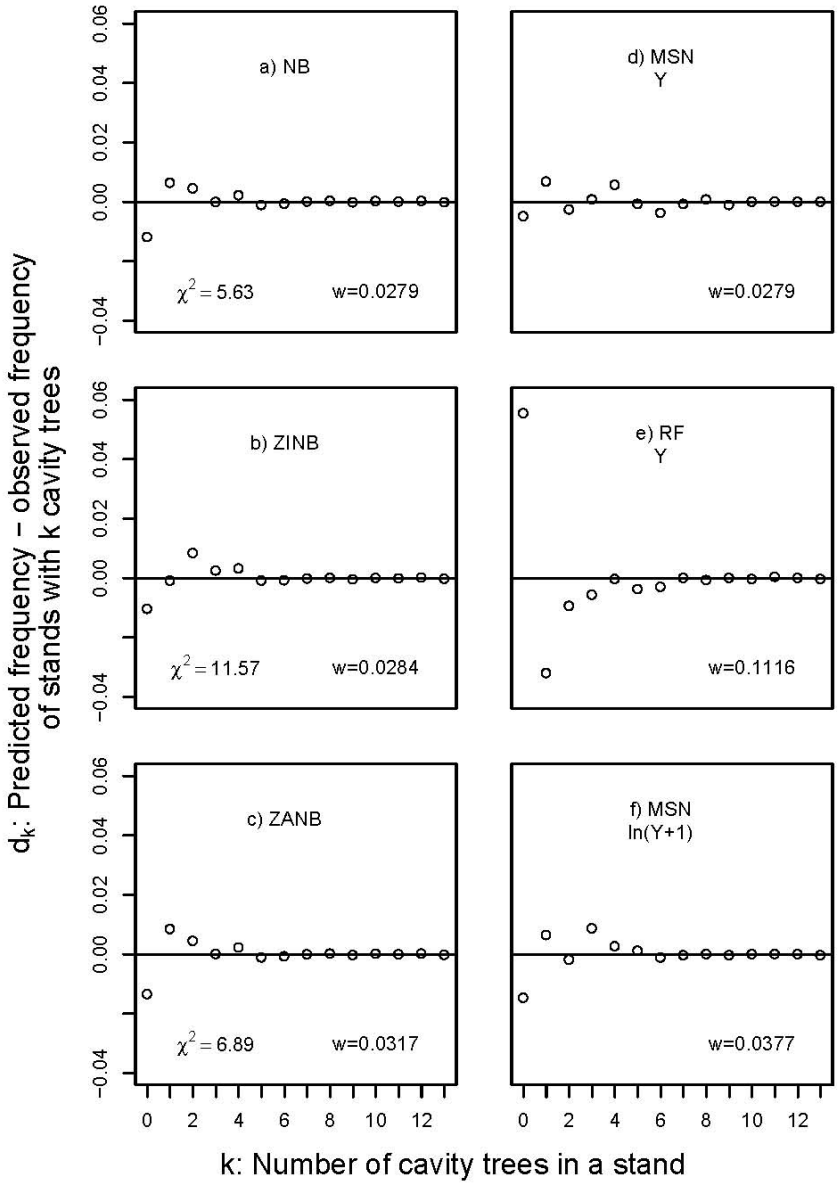

Figure 5.2: Diagnostic plots for cavity tree abundance for the negative binomial (NB), zero-inflated NB (ZINB), zero-altered NB (ZANB), and three NN imputation methods. χ^2 is the χ^2-statistic for the NB, ZINB, and ZANB models. *W* is the sum of the absolute values of d_k.

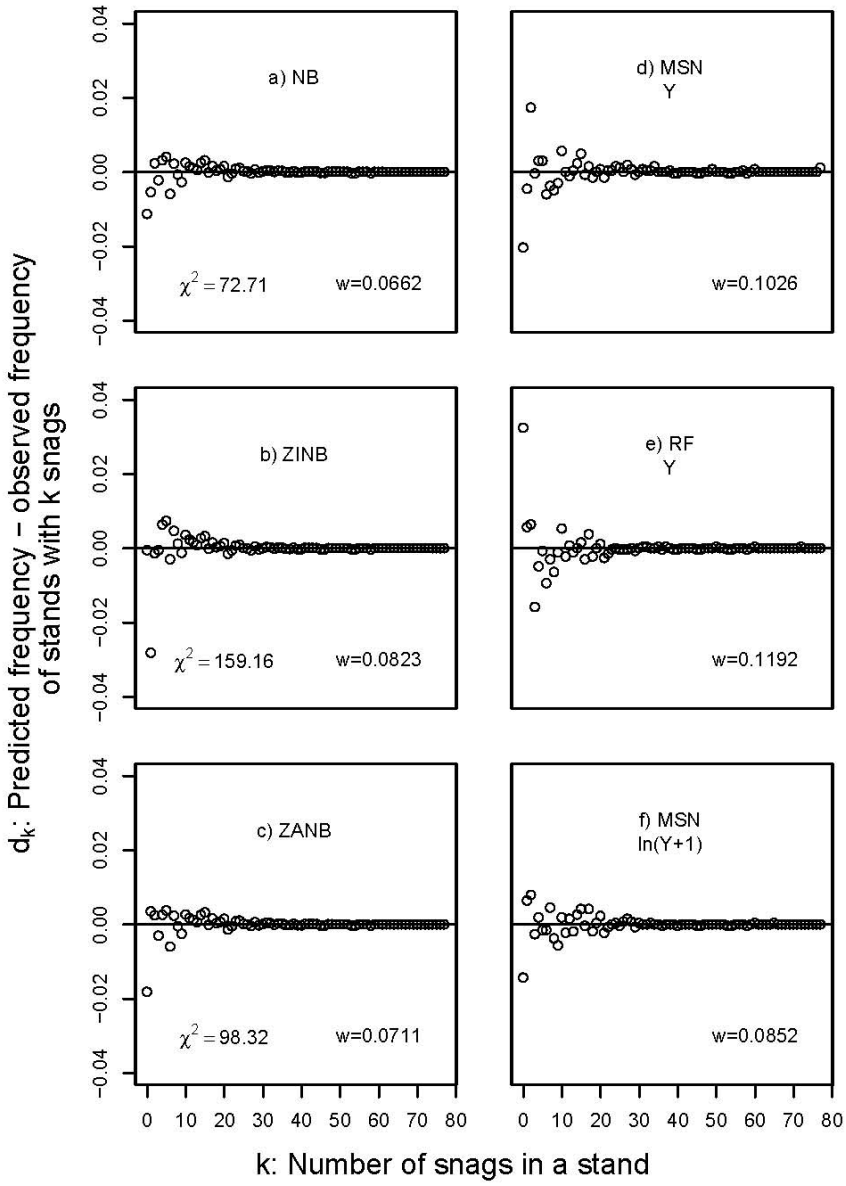

Figure 5.3: Diagnostic plots for cavity tree abundance for the negative binomial (NB), zero-inflated NB (ZINB), zero-altered NB (ZANB), and three NN imputation methods. χ^2 is the χ^2-statistic for the NB, ZINB, and ZANB models. W is the sum of the absolute values of d_k.

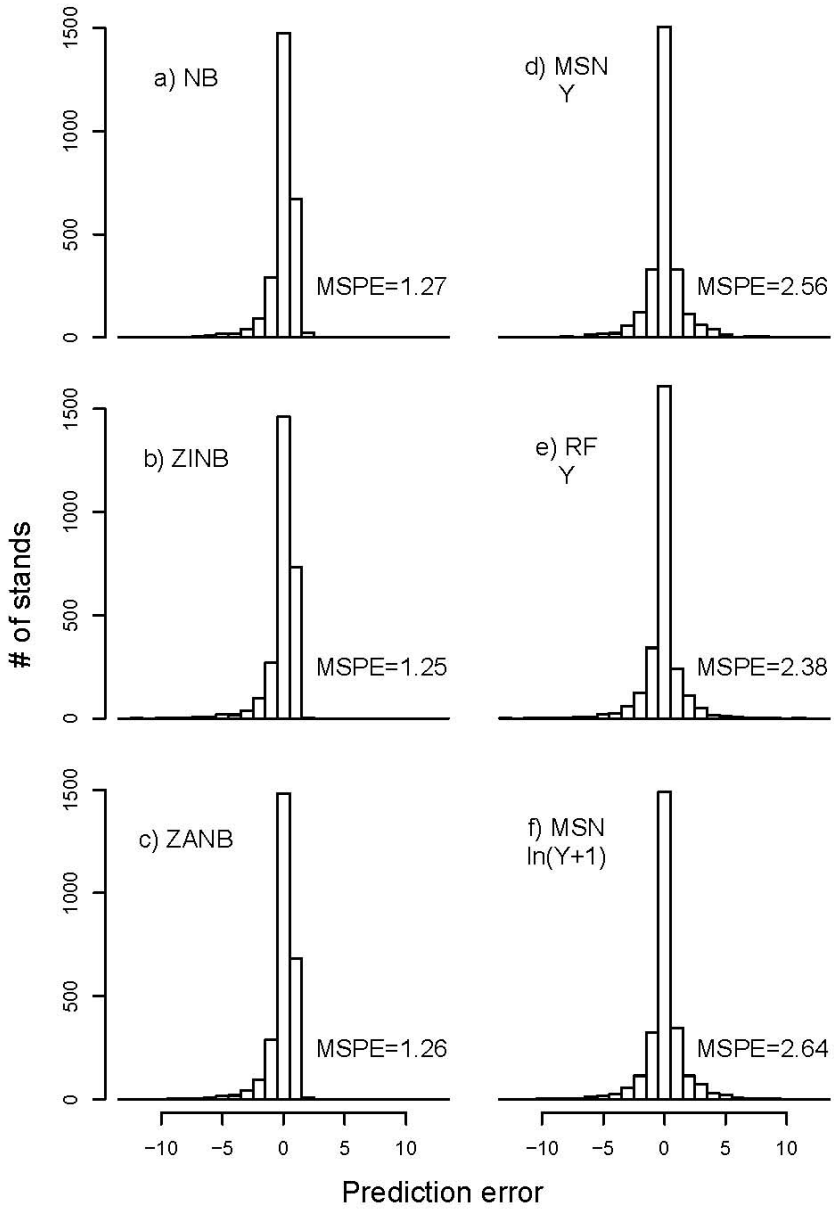

Figure 5.4: Frequency plots of prediction error of cavity tree abundance for the negative binomial (NB), zero-inflated NB (ZINB), zero-altered NB (ZANB), and three NN imputation methods. MSPE is the mean square prediction error.

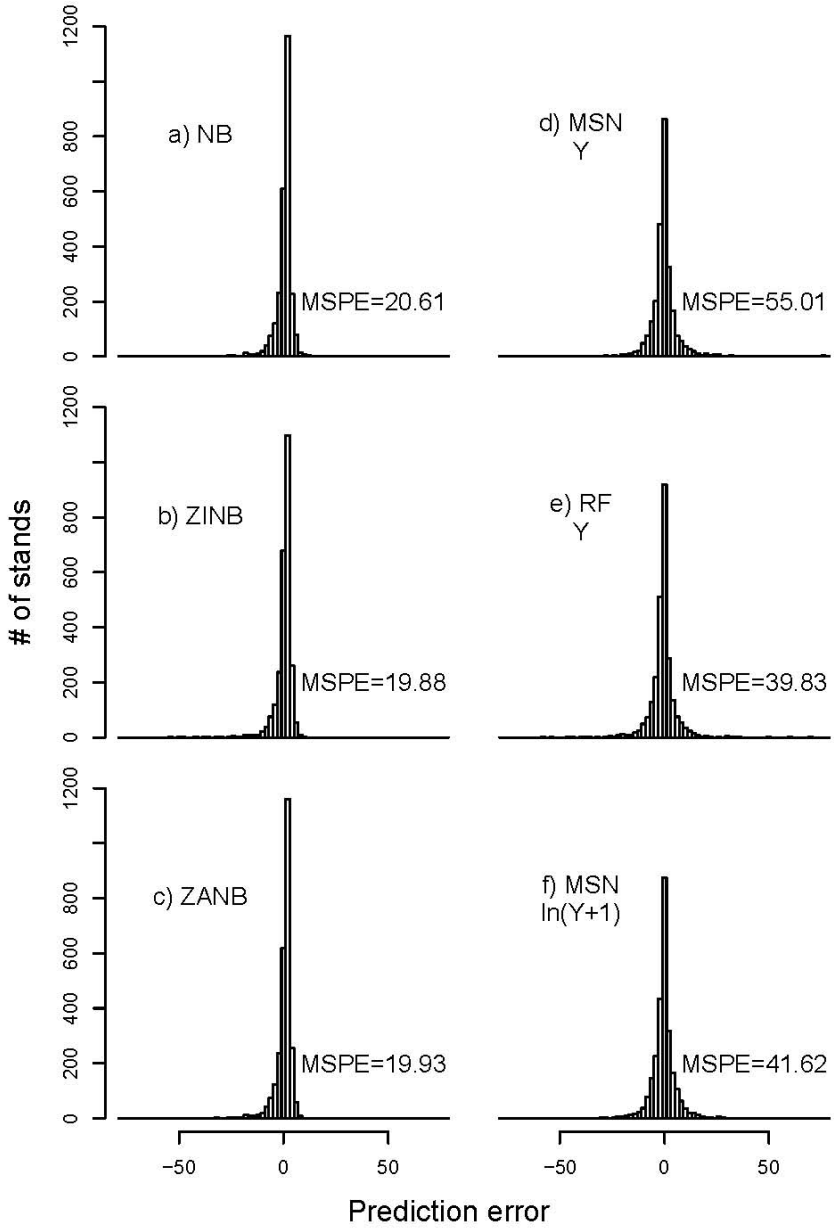

Figure 5.5: Frequency plots of prediction error of snag abundance for the negative binomial (NB), zero-inflated NB (ZINB), zero-altered NB (ZANB), and three NN imputation methods. MSPE is the mean square prediction error.

CHAPTER 6: CONCLUSION

The goal of this dissertation was to explore alternatives to the moving average (MA), the Forest Inventory and Analysis (FIA) default estimator, for estimating current forest condition and change from paneled inventory data in the Pacific Northwest (PNW). In the process, a variety of data sets were examined for their potential use for this research. Data from the Current Vegetation Survey (CVS) and annual FIA inventory data were used to complete the study. Specific objectives were to 1) examine plot-level nearest neighbor imputation techniques for estimating current plot-level attributes; 2) investigate tree-level nearest neighbor imputation techniques for estimating current plot-level attributes; 3) explore the suitability of imputation techniques to estimate mean annual change at plot-level; and 4) analyze the suitability of nearest neighbor imputation methods and negative binomial regression models to estimate cavity tree and snag abundance.

In Chapter 2, a weighted moving average (WMA) and three nearest neighbor (NN) plot-level imputation techniques (MSN, GNN, and RF) were examined as alternatives to the MA. Using the most recent measurements of the variables of interest as ancillary variables, RF provided almost unbiased estimates that were comparable to those of the MA and WMA estimators. MSN and GNN could not compete with any of the other methods with the available ancillary variables. For the MA and WMA estimates, the variance was very small and bias contributed most to the root mean square error (RMSE). For the imputation methods, the variance contributed

most to the RMSE. If the lag bias of the MA could be corrected, MA might outperform all other methods.

In Chapter 3, tree-level NN imputation techniques were assessed and compared with MA and WMA estimators as well as plot-level NN imputation. RF imputation was used for both tree-level and plot-level imputation. Tree-level imputation outperformed plot-level imputation as well as the MA and WMA estimators in estimating current forest attributes such as BA, SPH, VOL, and BIOT. When the variables of interest were summarized by three species groups, tree-level imputation also outperformed plot-level imputation. The more detailed information that is required for the tree-level imputation provides the potential for more detailed output. For species groups 'Douglas-fir' and 'other,' MA and tree-level imputation provided comparable results, whereas tree-level imputation outperformed MA for species group 'pine.'

In Chapter 4, mean annual change (MAC) estimation of forest attributes using MSN, GNN, and RF imputation was attempted. The imputed MAC was used to project all panels to a common point in time. The resulting mean estimates of the forest attributes outperformed the estimates based on SAMPLE25, MA, and WMA estimators in terms of accuracy. Updating previously observed measurements of forest attributes with imputed MAC estimates also outperformed imputing BA, SPH, VOL, and BIOT for the year 2000 directly using RF imputation as was done in Chapter 2.

In Chapter 5, the use of negative binomial regression models and NN imputation methods to predict cavity tree and snag abundance was explored. Negative

binomial (NB), zero-inflated NB (ZINB), and zero-altered NB (ZANB) models were found to provide reasonable results for estimating cavity tree and snag abundance. Due to its simpler application and interpretation the NB model is preferred over the ZINB and ZANB models. NN imputation models resulted in both large under- and overpredictions of cavity tree and snag abundance, whereas NB models only resulted in large underpredictions, therefore providing more conservative results.

Future directions

Several future research areas were identified: 1) Verifying results with 10 panel data set; 2) Correcting lag bias for moving average; 3) Examining alternative sources of ancillary data for nearest neighbor imputation; 4) Comparing the performance of tree-level nearest neighbor imputation vs. tree-level growth models; 5) Estimating change; and 6) Improving cavity tree and snag abundance estimation. Specific needs for each of the research topics are discussed below.

Verifying results with 10 panel data set

For the exploration of tree-level and plot-level imputation, CVS data was used in this study because it is the only data in the PNW that is comparable to the FIA inventory data and has remeasurements available. FIA data from the PNW could not be used because no remeasurements exist yet. The CVS data set could only be used to imitate a four panel inventory and could not be used to examine how the selected methods would perform in a 10 panel inventory. It is strongly recommended that this research be replicated with another data set. This could be done with a simulated test

population such as those constructed by McRoberts (2001), Johnson et al. (2003), Arner et al. (2004), and Roesch (2007a).

In this study the sampling intensity equaled 25%. For the 10-panel inventory in Oregon the sampling intensity is only 10%. Hence, the sampling intensity decreases for the FIA annual inventory in the PNW compared to this study. A sampling intensity of 20% has been found to be sufficient for estimating stand level variables using NN imputation methods (Moeur 2000, LeMay and Temesgen 2005). Further research is warranted to test whether a sampling intensity of 10% provides enough reference plots for NN imputation methods.

The CVS data used in this study only covered national forests. Therefore, several aspects that are pertinent to the analysis of FIA data could not be considered. For example, ownership related attributes could not be taken into account. Ownership can be considered an important predictor variable for estimating current forest attributes from FIA data and should be taken into account for FIA data analyses. This could possibly be achieved by imputing within strata, for example, impute within ownership group. Eskelson et al. (2008) did not find stratified MSN approaches to improve the imputation results but ascribed this to the small number of plots used in the study. The FIA inventory data from the PNW should provide enough plots for performing imputation separately within ownership group. The CVS data was measured in 1994 and the following years. Hence, no or few management operations were carried out between measurements due to the Northwest Forest Plan that came into effect in 1994. When the FIA data is analyzed, silvicultural treatments and other

management operations need to be taken into account in the models used to update the panels that are not measured in the current year.

Correcting lag bias for moving average

The results of this study showed that the MA estimator provides very precise but biased estimates whereas the NN imputation methods provide less biased results with large variance. If the lag bias of the MA estimator could be corrected, the MA would possibly outperform all other methods in terms of both bias and RMSE due to its high precision. Using weighted moving averages which apply weights that give more weight to the most recently measured panels is an approach to adjust for the lag bias of the MA. However, the selection of the weights is not yet solved. Research for finding methods that allow an objective selection of the weights is warranted. The trend that is inherent in the data needs to be known for choosing appropriate weights for the WMA, which poses another problem for the weight selection. Furthermore, the application of the MA and WMA estimators in a 10 panel inventory needs to be tested in order to explore their behavior for a long inventory cycle. It can be expected that the lag bias of the MA increases with increasing inventory cycle length (Johnson et al. 2003).

Examining alternative sources of ancillary data for nearest neighbor imputation

The climate, topography, and satellite data available in this study were not found to be very useful for NN imputation. Instead of using Landsat TM data as ancillary data, research should focus on using higher resolution data such as LiDAR.

The potential of LiDAR data for predicting forest attributes such as plot-level basal area, tree density, and volume has been demonstrated using multiple linear regression (Hudak et al. 2006), ratio estimation (Corona and Fattorini 2008), and NN imputation approaches (Maltamo et al. 2006, Hudak et al. 2008). Combining ancillary data derived from LiDAR and aerial photographs has been shown to improve the estimation of species specific stand attributes in terms of accuracy (Maltamo et al. 2006, Packalén and Maltamo 2007).

Basal area/ha is a two dimensional variable whereas volume/ha and biomass/ha are three dimensional variables derived from DBH and height. In this study, many of the ancillary variables used for imputation were two dimensional variables. Using three dimensional ancillary data derived from LiDAR data could possibly improve the imputation results for three-dimensional variables such as volume/ha and biomass/ha.

In this study, initial values of the variables of interest were used as ancillary data, which provided very good results for the RF imputation. Predicted values of the variables of interest from growth and yield models could be used as ancillary variables (e.g., Gartner and Reams 2002). Yet another option could be to predict growth rates and mortality rates with growth and yield models and employ them as ancillary data for imputation methods. This approach combines growth and yield models with imputation approaches and might lead to more precise estimates than using only either imputation methods or growth and yield models. Research in this area is warranted.

Comparing the performance of tree-level nearest neighbor imputation vs. tree-level growth models

In this study, individual-tree growth models were not used to update height and diameter of *P1*, *P2*, and *P3* data to the year 2000. Tree-level NN imputation has been shown to provide comparable or better results than traditional regression models for Norway spruce (*Picea abies* Karst.) and Scots pine (*Pinus sylvestris* L.) stands with low variability (Sironen et al. 2001, 2003, Fehrmann et al. 2008). The data used in this study included 33 species and a wide variety of stand size and density conditions. It is recommended that the performance of tree-level NN imputation be compared against regression models using data obtained from the PNW. This will help to ascertain whether tree-level NN imputation can compete with regression models under more variable conditions than those shown in Sironen et al. (2001, 2003) and Fehrmann et al. (2008).

An advantage of regression techniques used in traditional growth and yield models is that, once the model parameters have been estimated, the growth model equations are easy to apply and to communicate. Growth models can be employed without having access to a database of raw data collected from a diverse set of site or stand conditions (Fehrmann et al. 2008). While growth models need to be refitted to be updated and keep their validity, NN imputation approaches update themselves when data are added or removed from the database (Sironen et al. 2001, 2003). Therefore, NN imputation approaches are considered an alternative to regression models, once a larger single-tree data base is available (Fehrmann et al. 2008).

A frequently mentioned advantage of NN imputation over growth models is that the NN imputation techniques are multivariate and able to estimate multiple variables (e.g., HT, DBH, crown ratio) simultaneously (e.g., Katila and Tomppo 2002) whereas models have to be fit for each variable separately when modeling approaches are used. NN imputation methods are able to preserve the covariance structure of the data if only the nearest neighbor is used in the imputation process, which is important if the data is used for further modeling (Fehrmann et al. 2008). This is crucial for management applications that require information on multiple forest attributes and where estimates of these attributes must be compatible (McRoberts 2008). However, compatible estimates can also be achieved with traditional growth and yield models when systems of equations are used (e.g., Borders 1989, Hasenauer et al. 1998).

Different growth intervals are required for updating paneled inventory data. Available growth models such as ORGANON (Hann 2006) and FVS (Stage 1973) provide five and 10 year growth intervals, respectively, which can be interpolated to provide the required growth intervals. Another option is to use annualized growth equations (e.g., Cao 2000, Weiskittel et al. 2007). When NN imputation is done by matching on initial values, the different growth intervals could cause a problem. The FIA annual inventory data in the PNW where each plot is remeasured every 10 years provides data on ten year growth periods only. In order to impute tree data at time *t* for the missing nine panels that were measured 1 to 9 years before, reference data is needed that can be matched to 1 to 9 year old data. The FIA inventory, however, only provides initial values from 10 years ago in the reference data. When tree-level

imputation is used with the FIA data from the PNW, the results might become biased due to the fact that the reference data only provides 10 year old data to match on. If the annual FIA inventory was a balanced annual partial remeasurement design as suggested by Arner et al. (2004), the reference data would include remeasurements that provide all possible growth periods. However, a balanced annual partial remeasurement design would not be very practical. Performing NN imputation by matching on predicted values from growth and yield models could evade the problem of different growth intervals. Further investigation of this is warranted.

A major disadvantage of the NN imputation methods is that extrapolation beyond the distribution of available reference data is impossible (Moeur and Stage 1995), which makes the predictions highly dependent on the number of available reference data. When only the nearest neighbor is used in the imputation, interpolation is impossible (Crookston et al. 2002). Hence, the reference data needs to cover the whole range of ancillary data without any large gaps (Stage and Crookston 2007). Another disadvantage of the NN imputation methods is that error estimation techniques are still under development (e.g., Kim and Tomppo 2006, McRoberts et al. 2007).

Estimating change

Change in forest attributes is at least as important to most users of FIA data as current status of forest attributes (Van Deusen 2002b). Change estimation could not be explored with the available CVS data in this study, since the data did not provide two

remeasurements of each plot. Compatible estimators of the components of change for paneled inventory data have been explored (Roesch 2007a, 2007b).

Using imputed mean annual change to update unmeasured panels to the current time has shown great potential in this study and suggests that mean annual change imputation could be used for change estimation.

Traditionally, change estimation is performed with growth and yield models. Change estimation could also be approached using tree-level nearest neighbor imputation. This study has shown the potential of tree-level randomForest imputation for updating tree diameter, height, and mortality. More research is warranted for improving tree-level nearest neighbor imputation for imputing diameter growth, height growth, change in height to crown base, and mortality.

Improving cavity tree and Snag abundance

Estimation of cavity tree and snag abundance was explored using data from the annual FIA inventory. No remeasurements were available at the time of this study. As soon as remeasurements are available, the NN imputation methods and NB models should be tested using previous information on cavity tree and snag abundance as explanatory variables. This could possibly improve the results, when the information of unmeasured FIA plots is updated to the current year.

For providing updates to paneled data or for providing data to unsampled stands or polygons, the evaluation of the cavity tree and snag abundance models

would need to include looking at whether the results are also relatively unbiased and accurate by ownership, forest-type, and other strata.

Summary

Methods for estimating current condition and change of forest attributes from paneled inventory data that could replace the moving average, which is the current Forest Inventory and Analysis default estimator, are examined. This dissertation explored a variety of tree-level and plot-level nearest neighbor imputation methods for updating unmeasured panels to the current point in time. In the process, nearest neighbor imputation methods were also compared to negative binomial regression models for their predictive abilities of estimating cavity tree and snag abundance. Tree-level as well as plot-level imputation using the randomForest method showed great potential for updating forest attributes of unmeasured panels, but further research on employing the moving average, weighted moving average, and randomForest imputation to a 10 panel inventory is warranted. Further research on comparing tree-level nearest neighbor imputation with individual-tree growth models is also highly warranted. The nearest neighbor imputation methods could potentially be improved by using growth model predictions as ancillary data or by using ancillary data with higher resolution such as LiDAR data.

BIBLIOGRAPHY

Aakala, T., T. Kuuluvainen, S. Gauthier, and L. De Grandpré. 2008. Standing dead trees and their decay-class dynamics in the northeastern boreal old-growth forests of Quebec. Forest Ecology and Management 255:410-420.

Affleck, D.L.R. 2006. Poisson mixture models for regression analysis of stand-level mortality. Canadian Journal Forest Research 36:2994-3006.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. P. 267-281 in *Second International Symposium on Information Theory*, Petrov, B.N., and F. Csaki. Akademiai Kaido, Budapest.

Akaike, H. 1974. A new look at the statistical identification model. IEEE Transactions on Automatic Control 19:716-723.

Allen, A.W. and J.G. Corn. 1990. Relationships between live tree diameter and cavity abundance in a Missouri oak-hickory forest. Northern Journal of Applied Forestry 7:179-183.

Arner, S.L., J.A. Westfall, and C.T. Scott. 2004. Comparison of annual inventory designs using forest inventory and analysis data. Forest Science 50(2): 88-203.

Azuma, D. 2000. Moving to annual inventory in the Pacific Northwest. P. 5-7 in *Proceedings of the First Annual Forest Inventory and Analysis Symposium*, McRoberts, R.E., G.A. Reams, P.C. Van Deusen (eds.). USDA For. Serv. Gen. Tech. Rep. NC-213. St. Paul, MN.

Barrett, T.M. 2006. Optimizing efficiency of height modeling for extensive forest inventories. Canadian Journal of Forest Research 36:2259-2269.

Bate, L.J., E.O. Garton, and M.J. Wisdom. 1999. Estimating snag and large tree densities and distributions on a landscape for wildlife management. USDA For. Serv. Gen. Tech. Rep. PNW-GTR-425. 76 p.

Bechtold, W.A. and C.T. Scott. 2005. The forest inventory and analysis plot design. Gen. Tech. Rep. SRS-80. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station, p. 37-52.

Bechtold, W.A., and P.L. Patterson [Editors] 2005. The enhanced forest inventory and analysis program – national sampling design and estimation procedures. Gen. Tech. Rep. SRS-80. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 85 p.

Borders, B. 1989. Systems of equations in forest stand modeling. Forest Science 35(2):548-556.

Brand, G.J., M.D. Nelson, D.G. Wendt, and K.K. Nimerfro. 2000. The hexagon/panel system for selecting FIA plots under an annual inventory. P. 8-10 in *Proceedings of the First Annual Forest Inventory and Analysis Symposium*, McRoberts, R.E., G.A. Reams, P.C. Van Deusen (eds.). USDA For. Serv. Gen. Tech. Rep. NC-213. St. Paul, MN.

Breidt, F.J. 1999. Estimation systems for rolling samples. Forest inventory and monitoring biometrics workshop, New Orleans, LA. Available online at http://www.stat.colostate.edu/~jbreidt/Lectures/rolling.pdf; last accessed Sept. 10, 2008.

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

Burkhart, H.E. 1992. Tree and stand models in forest inventory. Proceedings of Ilvessalo Symposium on National Forest Inventories. IUFRO S4.02. Finnish Forest Research Institute Research Papers 444:164–170.

Cameron, A.C. and P.K. Trivedi. 1998. Regression analysis of count data. Cambridge University Press, Cambridge. 411 p.

Cameron, A.C. and P.K. Trivedi. 2005. Microeconometrics: Methods and applications. Cambridge University Press, Cambridge. 1034 p.

Cao, Q.V. 2000. Prediction of annual diameter growth and survival for individual trees from periodic measurements. Forest Science 46(1):127-131.

Carey, A.B. 1983. Cavities in trees in hardwood forests. P. 167-184 in *Snag habitat management symposium.* USDA For. Serv. Proc. RMRS-P-25. 208 p.

Corona, P. and L. Fattorini. 2008. Area-based lidar-assisted estimation of forest standing volume. Canadian Journal of Forest Research 38:2911-2916.

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

Crookston, N.L., M. Moeur, and D. Renner. 2002. Users guide to the most similar neighbour imputation program version 2. USDA For. Serv. Gen. Tech. Rep. RMRS-GTR-96. 35 p.

Curtis, R.O. and D.M. Hyink. 1985. Data for growth and yield models. P. 1-5 in Proceedings – Growth and yield and other mensurational tricks: A regional technical

conference, Van Hooser, D.D. and N. Van Pelt (eds.). USDA For. Serv. Gen. Tech. Rep. INT-GTR-193. Ogden, UT.

DeLong, S.C., G.D. Sutherland, L.D. Daniels, B.H. Heemskerk, and K.O. Storaunet. 2008. Temporal dynamics of snags and development of snag habitats in wet spruce-fir stands in east-central British Columbia. Forest Ecology and Management 255:3613-3620.

Environmental Systems Research Institute [ESRI]. 1991. Cell-based modeling with GRID. ESRI Inc., US, Redlands, California.

Eskelson, B.N.I., H. Temesgen, and T.M. Barrett. 2008. Comparison of stratified and non-stratified most similar neighbor imputation for estimating stand tables. Forestry 81(2):125-134.

Eskelson, B.N.I., H. Temesgen, and T.M. Barrett. In press. Estimating current forest attributes from paneled inventory data using plot-level imputation: a study from the Pacific Northwest. Forest Science.

Fan, Z., D.R. Larsen, S.R. Shifley, and F.R. Thompson. 2003a. Estimating cavity tree abundance by stand age and basal area, Missouri, USA. Forest Ecology and Management 179:231-242.

Fan, Z., S.R. Shifley, M.A. Spetich, F.R. Thompson, and D.R. Larsen. 2003b. Distribution of cavity trees in Midwestern old-growth and second-growth forests. Canadian Journal of Forest Research 33:1481-1494.

Fan, Z., S.R. Shifley, F.R. Thompson, and D.R. Larsen. 2004. Simulated cavity tree dynamics under alternative timber harvest regimes. Forest Ecology and Management 193:399-412.

Fan, Z., S.R. Shifley, M.A. Spetich, F.R. Thompson, and D.R. Larsen. 2005. Abundance and size distribution of cavity trees in second-growth and old-growth central hardwood forests. Northern Journal of Applied Forestry 22(3):162-169.

Fehrmann, L., A. Lehtonen, C. Kleinn, and E. Tomppo. 2008. Comparison of linear and mixed-effect regression models and a k-nearest neighbour approach for estimation of single-tree biomass. Canadian Journal of Forest Research 38:1-9.

FIA 2005. Forest inventory and analysis, phase 2 and phase 3: Ground measurements. FIA Fact Sheet Series. 2 p. Available online at http://www.fia.fs.fed.us/library/fact-sheets; last accessed Sept. 15, 2008.

Finley, A.O. and R.E. McRoberts. 2008. Efficient k-nearest neighbor searches for multi-source forest attribute mapping. Remote Sensing of Environment 112(5):2203-2211.

Fortin, M., and J. DeBlois. 2007. Modeling tree recruitment with zero-inflated models: the example of hardwood stands in southern Québec, Canada. Forest Science 53(4):529-539.

Frayer, W.E. and G.M. Furnival. 1999. Forest survey sampling designs. A history. Journal of Forestry 97(12):4-10.

Ganey, J.L. and S.C. Vojta. 2005. Changes in snag populations in northern Arizona mixed-conifer and Ponderosa pine forests, 1997-2002. Forest Science 51(5):396-405.

Gartner, D. and G.A. Reams. 2001. A comparison of Several Techniques for estimating the average volume per acre for multipanel data with missing panels. P. 76-81 in *Proceedings of the second annual Forest Inventory and Analysis Symposium*, Reams, G.A., R.E. McRoberts, P.C. Van Deusen (eds.). USDA For. Serv. Gen. Tech. Rep. SRS-47. Asheville, NC.

Gartner, D. and G.A. Reams 2002. A comparison of several techniques for imputing tree level data. P. 103-108 in *Proceedings of the Third Annual Forest Inventory and Analysis Symposium*, McRoberts, R.E., G.A. Reams, P.C. Van Deusen, and J.W. Moser (eds.). USDA For. Serv. Gen. Tech. Rep. NC-GTR-230, St. Paul, MN.

Gillepsie, A.J. 1999. Rationale for a national Annual Forest Inventory program. Journal of Forestry 97(12): 16-20.

Gillis, M.D. and D.G. Leckie. 1996. Forest inventory update in Canada. The Forestry Chronicle 72(2):138-156.

Goodburn, J.M. and C.G. Lorimer. 1998. Cavity trees and coarse woody debris in old-growth and managed northern hardwood forests in Wisconsin and Michigan. Canadian Journal of Forest Research 28:427-438.

Hall, D.B. 2000. Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics 56: 1030-1039.

Hann, D.W. 2006. ORGANON User's manual. Edition 8.2. Oregon State University, Department of Forest Resources, Corvallis, Oregon. 129 p.

Harmon, M.E., J.F. Franklin, F.J. Swanson, P. Sollins, S.V. Gregory, J.D. Lattin, N.H. Anderson, S.P. Cline, N.G. Aumen, J.R. Sedell, G.W. Lienkaemper, K. Cromack, and

K.W. Cummins. 1986. Ecology of coarse woody debris in temperate ecosystems. Advances in Ecological Research 15:133-302.

Hasenauer, H., R. Monserud, and T.G. Gregoire. 1998. Using simultaneous regression techniques with individual-tree growth models. Forest Science 44(1):87-95.

Hilbe, J.M. 2007. Negative binomial regression. Cambridge University Press, Cambridge. 251p.

Homer, C., C. Huang, L. Yang, B. Wylie and M. Coan. 2004. Development of a 2001 National Landcover Database for the United States. Photogrammetric Engineering and Remote Sensing, 70(7):829-840.

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

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

Jensen, R.G., J.M. Kabrick, and E.K. Zenner. 2002. Tree cavity estimation and verification in the Missouri Ozarks. P. 114-129 in *Proceedings of the second Missouri Ozark Forest Ecosystem Project symposium: Post-treatment results of the landscape experiment*, Shifley, S.R., and J.M. Kabrick (eds.). USDA For. Serv. Gen. Tech. Rep. NC-227.

Johnson, D.S. and M.S. Williams. 2004. Some theory for the application of the moving average estimator in forest surveys. Forest Science 50(5):672-681.

Johnson, D.S., M.S. Williams, and R.L. Czaplewski 2003. Comparison of estimators for rolling samples using forest inventory and analysis data. Forest Science 49(1):50-63.

Katila, M. and E. Tomppo. 2002. Stratification by ancillary data in multisource forest inventories employing k-nearest neighbor estimation. Canadian Journal of Forest Research 32:1548-1561.

Kauth, R.J. and G.S. Thomas. 1976. The tasseled cap—a graphic description of the spectral-temporal development of agricultural crops as seen by Landsat. P. 4B41-4B51 in *Proceedings of the symposium on machine processing of remotely sensed data*, Purdue University. West Lafayette, IN.

Kim, H.-J. and Tomppo, E. 2006. Model-based prediction error uncertainty estimation for k-nn method. Remote sensing of Environment 104:257-263.

Korhonen, K.K. and A. Kangas. 1997. Application of nearest-neighbour regression for generalizing sample tree information. Scandinavian Journal of Forest Research 12:97-101.

Lambert, D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34(1): 1-14.

LeMay, V. and H. Temesgen. 2005. Comparison of nearest neighbor methods for estimating basal area and stems per hectare using aerial auxiliary variables. Forest Science 51(2):109-199.

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

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

Max, T.A., H.T. Schreuder, J.W. Hazard, J. Teply, and J. Alegria. 1996. The Region 6 vegetation inventory and monitoring system. USDA For. Serv. Gen. Tech. Rep. PNW-RP-493. 22 p.

McRoberts, R.E. 1999. Joint Annual Forest Inventory and Monitoring System. The North Central Perspective. Journal of Forestry 97(12): 27-31.

McRoberts, R.E. 2000. Background for AFIS, the Annual Forest Inventory System. P. 1-4 in *Proceedings of the First Annual Forest Inventory and Analysis Symposium*, McRoberts, R.E., G.A. Reams, P.C. Van Deusen (eds.). USDA For. Serv. Gen. Tech. Rep. NC-213. St. Paul, MN.

McRoberts, R.E. 2001. Imputation on model-based updating techniques for annual forest inventories. Forest Science 47(3):322-330.

McRoberts, R.E. 2008. Using satellite imagery and the k-nearest neighbors technique as a bridge between strategic and management forest inventories. Remote Sensing of Environment 112:2212-2221.

McRoberts, R.E. and M.H. Hansen. 1999. Annual forest inventories for the north central region of the United States. Journal Agricultural, Biological, and Environmental Statistics 4(4):361-371.

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

Moeur, M. 2000. Extending stand exam data with most similar neighbor inference. P. 99-107 in *Proceedings of the Society of American Foresters National Convention*. Sept. 11-15, 1999, Portland, Oregon.

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

Nilsson, S.G., M. Niklasson, J. Hedin, G. Aronsson, J.M. Gutowski, P. Linder, H. Ljungberg, G. Mikusinski, and T. Ranius. 2002. Densities of large living and dead trees in old-growth temperate and boreal forests. Forest Ecology and Management 161:189-204.

Ohmann, J.L. and M.J. Gregory. 2002. Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, U.S.A. Canadian Journal of Forest Research 32:725-741.

Ohmann, J.L. and K.L. Waddell. 2002. Regional patterns of dead wood in forested habitats of Oregon and Washington. USDA Forest Service Gen. Tech. Report PSW-GTR-181, pp. 535–560.

Ohmann, J.L., M.J. Gregory, and T.A. Spies. 2007. Influence of environment, disturbance, and ownership on forest vegetation of coastal Oregon. Ecological Applications 17(1):18-33.

Packalén, P. and M. Maltamo. The k-MSN method for prediction of species-specific stand attributes using airborne laser scanning and aerial photographs. Remote Sensing of Environment 109:328-341.

Patterson, P.L. and G.A. Reams. 2005. Combining panels for forest inventory and analysis estimation. P. 69-74 in *The enhanced forest inventory and analysis program—National sampling and estimation procedures*, Bechtold, W.A. and P.L. Patterson (eds.). USDA For. Serv. Gen. Tech. Rep. SRS-GTR-80. Asheville, NC.

R Development Core Team. 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL http://www.R-project.org/.

Reams, G.A. and P.C. Van Deusen. 1999. The Southern annual forest inventory system. Journal of Agricultural, Biological, and Environmental Statistics. 4(4):345-59.

Reams, G.A., F.A. Roesch, and N.D. Cost. 1999. Annual Forest Inventory. Cornerstone of sustainability in the South. Journal of Forestry. 97(12):21-26.

Reams, A.R., W.D. Smith, M.H. Hansen, W.A. Bechtold, F.A. Roesch, and G.G. Moisen. 2005. The forest inventory and analysis sampling frame. P. 11-26 in *The enhanced forest inventory and analysis program—National sampling and estimation procedures*, Bechtold, W.A. and P.L. Patterson (eds.). USDA For. Serv. Gen. Tech. Rep. SRS-80. Asheville, NC.

Roesch, F.A. 2007a. Compatible estimators of components of change for a rotating panel forest inventory design. Forest Science 53(1):50-61.

Roesch, F.A. 2007b. The components of change for an annual forest inventory design. Forest Science 53(3):406-413.

Roesch, F.A. and G.A. Reams. 1999. Analytical alternatives for an annual inventory system. Journal of Forestry 97(12):33-37.

Russell, R.E., V.A. Saab, J.G. Dudley, and J.J. Rotella. 2006. Snag longevity in relation to wildfire and postfire salvage logging. Forest Ecology and Management 232:179-187.

Salas, C., A.R. Stage, and A.P. Robinson. 2008. Modeling effects of overstory density and competing vegetation on tree height growth. Forest Science 54(1):107-122.

Schwarz, G. 1978. Estimating the dimension of a model. Annals of Statistics 6:461-464.

Scott, C.T., D.L. Cassell, and J.W. Hazard. 1993. Sampling design of the U.S. National forest Health Monitoring Program. P. 150-157 in *Proceedings of the Ilvessalo symposium of national forest inventories*, Nyyssönen, A., S. Poso, and J. Rautala (eds.). Finish Forest Research Institute Research Paper 444.

Scott, C.T., M. Köhl, and H.J. Schnellbächer. 1999. A comparison of periodic and annual forest surveys. Forest Science 45(3):433-451.

Shorohova, E. and S. Tetioukhin. 2004. Natural disturbances and the amount of large trees, deciduous trees and coarse woody debris in the forests of Novgorod region, Russia. Ecological Bulletins 51:137-147.

Sironen, S., A, Kangas, M. Maltamo, and J. Kangas. 2001. Estimating individual tree growth with k-nearest neighbour and k-most similar neighbour methods. Silva Fennica 35(4):453-467.

Sironen, S., A. Kangas, M. Maltamo, and J. Kangas. 2003. Estimating individual tree growth with nonparametric methods. Canadian Journal Forest Research 33:444-449.

Sironen, S., A. Kangas, M. Maltamo, and J. Kangas. 2008. Localization of growth estimates using non-parametric imputation methods. Forest Ecology and Management 256:674-684.

Spiering, D.J. and R.L. Knight. 2005. Snag density and use by cavity-nesting birds in managed stands of the Black Hills National Forest. Forest Ecology and Management 214:40-52.

Stage, A.R. 1973. Prognosis model for stand development. USDA For. Serv. Gen. Tech. Rep. INT-137. 32 p.

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

Temesgen, H., T.M. Barrett, and. G. Latta. 2008. Estimating cavity tree abundance using nearest neighbor imputation methods for western Oregon and Washington forests. Silva Fennica 42(3): 337-354.

Thornton, P.E. and S.W. Running. 1999. An improved algorithm for estimating incident daily solar radiation from measurements of temperature, humidity, and precipitation. Agricultural and Forest Meteorology 93:211-228.

Thornton, P.E., S.W. Running, and M.A. White. 1997. Generating surfaces of daily meteorological variables over large regions of complex terrain. Journal of Hydrology 190:214-251.

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

USDA 2000. Western Oregon 1995-1997, August 15th, 2000. A CD available online from USDA/ FS PNW Research at http://www.fs.fed.us/pnw/fia/; last accessed Sep.10, 2008.

Van Deusen, P.C. 1996. Incorporating predictions into an annual forest inventory. Canadian Journal of Forest Research 26(9):1709-1713.

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

Van Deusen, P.C. 1999. Modeling trends with annual survey data. Canadian Journal of Forest Research 29:1824-1828.

Van Deusen, P.C. 2000. Pros and cons of the interpenetrating panel design. P. 14-19 in *Proceedings of the First Annual Forest Inventory and Analysis Symposium*, McRoberts, R.E., G.A. Reams, and P.C. Van Deusen (eds.). USDA Gen. Tech. Rep. NC-213. St. Paul, MN.

Van Deusen, P.C. 2002a. Issues related to panel creep. P. 31-35 in *Proceedings of the Third Annual Forest Inventory and Analysis Symposium*, R.E. McRoberts, G.A. Reams, P.C. Van Deusen, and J.W. Moser (eds.). USDA For. Serv. Gen. Tech. Rep. NC-230, Traverse City, MI.

Van Deusen, P.C. 2002b. Comparison of some annual forest inventory estimators. Canadian Journal of Forest Research 32:1992-1995.

Van Deusen P.C., S.P. Prisley, and A.A. Lucier. 1999. Adopting an annual inventory system. User perspectives. Journal of Forestry 97(12): 11-14.

Weiskittel, A.R., S.M. Garber, G.P. Johnson, D.A. Maguire, and R.A. Monserud. 2007. Annualized diameter and height growth equations for Pacific Northwest plantation-grown Douglas-fir, western hemlock, and red alder. Forest Ecology and Management 250:266-278.

Welsh, A.H., R.B. Cunningham, C.F. Donnelly, and D.B. Lindenmayer. 1996. Modelling the abundance of rare species: statistical models for counts with extra zeros. Ecological Modelling 88:297-308.

Wisdom, M.J. and L.J. Bate. 2008. Snag density varies with intensity of timber harvest and human access. Forest Ecology and Management 255:2085-2093.

Zeileis, A., C. Kleiber, and S. Jackman. 2008. Regression models for count data in R. Journal of Statistical Software, 27(8):1-25.

APPENDIX A

LIST OF SYMBOLS USED IN THE TEXT

Symbol	Definition	Units
%	percent	-
AFIS	Annual Forest Inventory System	-
AIC	Akaike's information criterion	-
ANNPRE	Annual precipitation	ln cm (scaled * 100)
ANNTMP	Mean annual temperature	°C (scaled * 100)
BA	Basal area	m^2/ha
BAL	Basal area in larger trees	m^2
BALocc1	Basal area in larger trees measured at occasion 1	m^2
BAocc1	BA measured at occasion 1	m^2/ha
BIC	Schwarz's information criterion	-
BIOT	Total gross oven dry weight biomass	Tons/ha
BIOTocc1	BIOT measured at occasion 1	Tons/ha
CANOPY	Tree canopy cover	%
CART	Classification and regression tree	-
CCA	Canonical correspondence analysis	-
cm	centimeter	-
CVS	Current Vegetation Survey	-
DBH	Diameter at breast height	cm
DBHocc1	Diameter at breast height measured at occasion 1	cm

LIST OF SYMBOLS USED IN THE TEXT (Continued)

D_{ij}^2	Distance metric for nearest neighbor imputation	-
d_k	Difference between predicted probabilities and observed frequencies	-
EL	Elevation	m
est_i	Refers to the estimated or imputed value for unit *i*	-
FHM	Forest Health Monitoring	-
FIA	Forest Inventory and Analysis	-
GNN	Gradient nearest neighbor	-
ha	hectare	-
HT	Total tree height	m
HTocc1	Total tree height measured at occasion 1	m
IMP	Refers to nearest neighbor imputation method	-
k	Count class	-
L()	Log-likelihood of model	-
LiDAR	Light Detection and Ranging	-
ln(EL)	Natural logarithm of EL	-
m	meter	-
m	Number of count classes (Chapter 5)	-
m	Number of iterations of randomly splitting data (Chapters 2-4)	-

LIST OF SYMBOLS USED IN THE TEXT (Continued)

MA	Moving Average	-
MAC	Mean annual change	-
MSN	Most similar neighbor	-
MSPE	Mean square prediction error	-
n	Number of observations	-
NB	Negative binomial	-
NCRS	North Central Research Station	-
NDVI	Normalized difference vegetation index	-
NN	Nearest neighbor	-
obs_i	Refers to the observed value for unit i	-
p	Number of explanatory variables	-
P1	Panel 1	-
P2	Panel 2	-
P3	Panel 3	-
P4	Panel 4	-
PE	Prediction error	-
PNW	Pacific Northwest	-
RF	randomForest	-
RMSE	Root mean square error in percent of mean	%
SAFIS	Southern Annual Forest Inventory System	-
SAMPLE25	Estimator using data from Panel 4 only	-

LIST OF SYMBOLS USED IN THE TEXT (Continued)

SPH	Stems per hectare	-
SPHocc1	SPH measured at occasion 1	-
sqrt()	Square root	-
SRS	Southern Research Station	-
TC	Tasseled Cap	-
TM	Thematic Mapper	-
US	Unites States	-
USDA	United States Department of Agriculture	-
VOL	Gross cubic-meter volume	m^3/ha
VOLocc1	VOL measured at occasion 1	m^3/ha
w	Sum of absolute differences d_k	-
w_{t-3}, w_{t-2}, w_{t-1},, w_t,	Panel weights of *P4*, *P3*, *P2*, and *P1*	-
W	Weight matrix in distance metric D_{ij}^2	-
WMA	Weighted Moving Average	-
x	Vector of explanatory variables	-
X	Ancillary variable	-
Y	Variable of interest	-
z	Vector of explanatory variables	-
ZANB	Zero-altered negative binomial	-
ZAP	Zero-altered Poisson	-
ZINB	Zero-inflated negative binomial	-

LIST OF SYMBOLS USED IN THE TEXT (Continued)

ZIP	Zero-inflated Poisson	-
ZTNB	Zero-truncated negative binomial	-
ZTP	Zero-truncated Poisson	-
α	Degree of overdispersion in NB regression model	-
β	Vector of regression coefficients	-
γ	Vector of regression coefficients	-
μ	Mean for Poisson and NB regression models	-
π	Probability of belonging to point mass component in ZINB regression model	-
χ^2	Chi-square	-
Γ	Matrix of standardized canonical coefficients for the ancillary variables	-
Λ^2	Diagonal matrix of squared canonical correlations between ancillary and ground variables	-

APPENDIX B

PARAMETER ESTIMATES AND AIC AND BIC VALUES FOR THE CAVITY TREE AND SNAG MODELS

Appendix B.1: Summary of fitted negative binomial (NB), zero-inflated NB (ZINB), zero-altered NB (ZANB) regression models for cavity tree abundance: coefficient estimates from count and zero models with standard errors in parentheses.

	NB	**ZINB**	**ZANB**
Count Model			
(Intercept)	-1.0860	-0.1403	-0.7352
	(0.1571)	(0.2070)	(0.3955)
Average stand age	0.0011		
	(0.0003)		
% conifer	-0.7259	-0.5429	-0.4959
	(0.1105)	(0.1603)	(0.2171)
Height class midpoints (m)	0.0351	0.0179	0.0129
	(0.0030)	(0.0031)	(0.0041)
Elevation (m)	-0.0002	-0.0003	-0.0003
	(0.0001)	(0.0001)	(0.0001)
Slope (%/100)	-0.5644	-0.8624	
	(0.3378)	(0.4026)	
Slope*cosine(aspect)	0.1524		
	(0.0820)		
Slope*ln(elevation)	0.1593	0.1755	0.0870
	(0.1593)	(0.0568)	(0.0267)
Forest type:	0.3965	0.2990	0.2239
fir/spruce/mountain hemlock	(0.0930)	(0.0931)	(0.1507)
Forest type: other conifers	-0.0384	-0.0758	-0.0116
	(0.0711)	(0.0709)	(0.1189)
Forest type: hardwoods	0.3002	0.2377	0.3546
	(0.0947)	(0.1068)	(0.1760)
Owner: other federal	-0.2791	-0.2711	-0.2623
	(0.0895)	(0.0886)	(0.1525)
Owner: state and	-0.1642	-0.2334	-0.3886
local governments	(0.1060)	(0.1058)	(0.1709)
Owner: private	-0.2160	-0.2453	-0.3032
	(0.0636)	(0.0657)	(0.1067)
Site class midpoints	0.0171		
(m^3/ha/year)	(0.0068)		
Zero Model			
(Intercept)		0.3525	-1.4610
		(0.2825)	(0.1387)

Appendix B.1 (continued)

Average stand age		-0.0179	0.0013
		(0.0033)	(0.0004)
% conifer		1.3573	-0.7126
		(0.3293)	(0.1202)
Height class midpoints (m)		-0.0513	0.0388
		(0.0141)	(0.0028)
Slope (%/100)		-0.7928	0.3351
		(0.5067)	(0.1117)
Slope*cosine(aspect)		-0.7570	0.1699
		(0.2920)	(0.0902)
Forest type:			0.3530
fir/spruce/mountain hemlock			(0.0958)
Forest type: other conifers			-0.0803
			(0.0734)
Forest type: hardwoods			0.2109
			(0.1032)
Owner: other federal			-0.2569
			(0.0971)
Owner: state and			-0.0508
local governments			(0.1113)
Owner: private			-0.1574
			(0.0638)
# estimated parameters	16	22	28
AIC	14,218	14,117	14,159
BIC	14,329	14,271	14,355

Appendix B.2: Summary of fitted negative binomial (NB), zero-inflated NB (ZINB), zero-altered NB (ZANB) regression models for snag abundance: coefficient estimates from count and zero models with standard errors in parentheses.

	NB	**ZINB**	**ZANB**
Count Model			
(Intercept)	0.9057	1.0580	-0.7352
	(0.1078)	(0.1119)	(0.3955)
Average stand age	-0.0006	-0.0015	
	(0.0003)	(0.0002)	
% conifer	-0.4089	-0.1617 (0.094)	-0.4959
	(0.0786)		(0.2171)
Height class midpoints (m)	0.0290	0.0202	0.0129
	(0.0021)	(0.0022)	(0.0041)
Elevation (m)	0.0002	0.0001	-0.0003
	(0.0000)	(0.0000)	(0.0001)
Slope (%/100)	0.1527		
	(0.0730)		
Slope * ln(elevation)			0.0870
			(0.0267)
Forest type:	0.5235	0.4849	0.2239
fir/spruce/mountain hemlock	(0.0648)	(0.0624)	(0.1507)
Forest type: other conifers	-0.1026	-0.0625	-0.0116
	(0.0493)	(0.0477)	(0.1189)
Forest type: hardwoods	-0.0888	0.0464	0.3546
	(0.0683)	(0.0806)	(0.1760)
Owner: other federal	-0.3603	-0.2676	-0.2623
	(0.0601)	(0.0591)	(0.1525)
Owner: state and	-0.2860	-0.2616	-0.3886
local governments	(0.0750)	(0.0716)	(0.1709)
Owner: private	-0.6635	-0.0658	-0.3032
	(0.0438)	(0.0419)	(0.1067)
Site class midpoints	-0.0182		
(m^3/ha/year)	(0.0049)		
Zero Model			
(Intercept)		-1.5600	-1.4610
		(0.6450)	(0.1387)
Average stand age		-0.0197	0.0013
		(0.0029)	(0.0004)
% conifer		1.952 (0.3855)	-0.7126
			(0.1202)

Appendix B.2 (continued)

Height class midpoints (m)		-0.0778	0.0388
		(0.0147)	(0.0028)
Elevation (m)		-0.0004	
		(0.0001)	
Slope (%/100)		-1.3375	0.3351
		(0.3343)	(0.1117)
Slope * cosine(aspect)			0.1699
			(0.0902)
Forest type:		-0.0750	0.3530
fir/spruce/mountain hemlock		(0.3466)	(0.0958)
Forest type: other conifers		0.2611	-0.0803
		(0.2369)	(0.0734)
Forest type: hardwoods		0.8396	0.2109
		(0.3586)	(0.1032)
Owner: other federal			-0.2569
			(0.0971)
Owner: state and			-0.0508
local governments			(0.1113)
Owner: private			-0.1574
			(0.0638)
Site class midpoints		0.1212	
(m^3/ha/year)		(0.0199)	
# estimated parameters	16	19	24
AIC	14,218	14,114	14,156
BIC	14,329	14,247	14,324