AN ABSTRACT OF THE THESIS OF

Donald Anthony Gagliasso for the degree of Master of Science in Sustainable Forest Management presented on October 1, 2012.
Title: Evaluating the Accuracy of Imputed Forest Biomass Estimates at the Project Level.

Abstract approved:

_____

Temesgen Hailemariam

Various methods have been used to estimate the amount of above ground forest biomass across landscapes and to create biomass maps for specific stands or pixels across ownership or project areas. Without an accurate estimation method, land managers might end up with incorrect biomass estimate maps, which could lead them to make poorer decisions in their future management plans.

Previous research has shown that nearest-neighbor imputation methods can accurately estimate forest volume across a landscape by relating variables of interest to ground data, satellite imagery, and light detection and ranging (LiDAR) data. Alternatively, parametric models, such as linear and non-linear regression and geographic weighted regression (GWR), have been used to estimate net primary production and tree diameter.

The goal of this study was to compare various imputation methods to predict forest biomass, at a project planning scale (<20,000 acres) on the Malheur National Forest, located in eastern Oregon, USA. In this study I compared the predictive performance of, 1) linear regression, GWR, gradient nearest neighbor (GNN), most similar neighbor (MSN), random forest imputation, and k-nearest neighbor (k-nn) to estimate biomass (tons/acre) and basal area (sq. feet per acre) across 19,000 acres on the Malheur National Forest and 2) MSN and k-nn when imputing forest biomass at spatial scales ranging from 5,000 to 50,000 acres.

To test the imputation methods a combination of ground inventory plots, LiDAR data, satellite imagery, and climate data were analyzed, and their root mean square error (RMSE) and bias were calculated. Results indicate that for biomass prediction, the k-nn (k=5) had the lowest RMSE and least amount of bias. The second most accurate method consisted of the k-nn (k=3), followed by the GWR model, and the random forest imputation. The GNN method was the least accurate. For basal area prediction, the GWR model had the lowest RMSE and least amount of bias. The second most accurate method was k-nn (k=5), followed by k-nn (k=3), and the random forest method. The GNN method, again, was the least accurate.

The accuracy of MSN, the current imputation method used by the Malheur Nation Forest, and k-nn (k=5), the most accurate imputation method from the second chapter, were then compared over 6 spatial scales: 5,000, 10,000, 20,000, 30,000, 40,000, and 50,000 acres. The root mean square difference (RMSD) and bias were

calculated for each of the spatial scale samples to determine which was more accurate. MSN was found to be more accurate at the 5,000, 10,000, 20,000, 30,000, and 40,000 acre scales. K-nn (k=5) was determined to be more accurate at the 50,000 acre scale.

Evaluating the Accuracy of Imputed Forest Biomass Estimates at the Project Level

by
Donald Anthony Gagliasso

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented October 1, 2012
Commencement June 2013

Master of Science thesis of Donald Anthony Gagliasso presented on October 1, 2012.

APPROVED:

_____

Major Professor, representing Sustainable Forest Management

_____

Head of the Department of Forest Engineering, Resources and Management

_____

Dean of the Graduate School

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

_____

Donald Anthony Gagliasso, Author

ACKNOWLEDGEMENTS

I am especially grateful for the opportunity to work with Temesgen Hailemariam, my major professor, and for his inspiration, patience, and support over the course of my work in the College of Forestry. I am very thankful for the opportunity to work with my committee members and for their input in completing my thesis, Susan Hummel, Professor Glen Murphy and Professor David Hibbs. Also, thank you to the USDA Forest Service, Pacific Northwest Research Station, for making this study possible by providing its funding to the College of Forestry. Thank you to my family, especially my mom and dad, for their moral support over the course of my education. In addition, I would like to thank the research assistants and students of the Forest Biometrics Lab in the College of Forestry at Oregon State University, especially Bianca Eskelson, for their constant support and feedback.

TABLE OF CONTENTS

Page

Chapter 1 – General Introduction 1

Chapter 2 – A comparison of the thematic accuracy of parametric and non-parametric methods 4

Introduction 4

Non-Parametric versus Parametric Models 5

Aerial LiDAR 8

Materials and Methods 12

Project Site 12

Aerial LiDAR Data 13

Ground Data 14

Data Compilation 19

Statistical Analysis 20

Results and Discussion 21

Conclusion 31

Chapter 3 – A comparison of the spatial accuracy of selected imputation methods 32

Introduction 32

Materials and Methods 34

TABLE OF CONTENTS (Continued)

Page

Project Site 34

Aerial LiDAR Data 34

Ground Data 34

Data Compilation 35

Creation of Scale Samples 35

Statistical Analysis 36

Results and Discussion 38

Conclusion 44

Chapter 4 – General Conclusion 45

Literature Cited 48

LIST OF TABLES

Page

Table 1: Number of Plots in Damon Site 14

Table 2: Coefficients and standard errors for linear regression model for ln(biomass) in tons per acre 22

Table 3: Coefficient and standard errors for linear regression model for basal area (ft^2 per acre)..... 23

Table 4: Basic statistics of explanatory and response variables 24

Table 5: Correlation coefficients of biomass vs. selected predictor variables..... 26

Table 6: Correlation Coefficients of basal area vs. selected predictor variables 27

Table 7: RMSE and bias for estimating biomass (tons/acre) by selected method..... 28

Table 8: RMSE and bias for estimating basal area (ft^2/acre) by selected method..... 28

Table 9: Number of Plots in Camp Creek Site..... 35

Table 10: Summary Statistics of Explanatory and Response Variables 39

Table 11: RMSD and bias for estimating biomass (tons/acre) by selected scale size . 41

CHAPTER 1 – GENERAL INTRODUCTION

Forest managers need accurate forest inventory data to develop a forest management plan that will allow them to prepare for future forest activities. Often times these data must cover large areas of land, up to thousands of acres. However, finding the balance of the amount of data to cover these thousands of acres and the cost to collect them can be very difficult. In recent years, the need for cost-effective, accurate forest inventory data has led to new ways of estimating and imputing plot data collected by the United States Department of Agriculture (USDA) Forest Inventory and Analysis Program (FIA). This process has resulted in regional maps of forest cover and vegetation types created from the Gradient Nearest Neighbor (GNN) method (Ohmann and Gregory, 2002), which has been recently used by the Oregon Department of Forestry (ODF), United States Department of Interior (USDI) Bureau of Land Management (BLM), and the USDA Forest Service (USFS). The GNN imputation method has been used in the past few years for analysis and planning efforts across the Pacific Northwest (PNW), and is being used to estimate many aspects of a regular forest inventory, including woody biomass. Woody biomass is becoming a desired forest product due to proposed energy facilities that use it as a renewable resource and an alternative to coal. Other imputation methods, such as the Most Similar Neighbor (MSN) (Moeur and Stage, 1995) and Random Forest (RF) (Crookston and Finley, 2008), were also developed by the USFS and are used in the PNW and throughout the Rocky Mountain region (Hudak et. al., 2008). Despite the

availability of these potential cost-effective imputation methods, they have generally all been used to create vegetation maps at a region scale (>100,000 acres). However, forest managers write forest plans at project level scales (<50,000 acres) and the accuracy of these imputation methods have not been tested at these scales due mostly to a lack of independent data. This can make it very difficult for forest managers to know which imputation methods should be used and when or how to report their accuracy when creating forest plans at the project level.

Growing public concern over the condition of our federal forests has brought proposals for forest silvicultural treatments to thin our forests in areas where insect and disease outbreaks have occurred. Land managers, decision makers, and scientists have asked about cost-effective ways to predict the amount of biomass across our federal forests to increase the amount of small trees that can be removed from the forests through thinning prescriptions and are looking to determine the local prediction accuracy (<20,000 acres) of these imputation methods that have generated maps at regional scales. Determining the spatial accuracy of these imputation methods at the project level will add confidence to the results of their investments in using these methods to impute forest vegetation maps.

This study will assess the predictive accuracy of imputed forest vegetation maps at spatial scales that are suitable for writing forest plans at the project level. It

seeks to quantify the accuracy of selected imputation methods at varying geographic scales.

CHAPTER 2 – A COMPARISON OF THE THEMATIC ACCURACY OF PARAMETRIC AND NON-PARAMETRIC METHODS

Introduction

A map can have many different uses in forestry. Uses of forest maps include a: harvest map with estimated timber volume and ownership boundaries, stand map for inventory data collection, road map for travel across an ownership, and hydrology map for various stream runoff and landslide issues. No matter the purpose of a specific map the question should be asked, how accurate is this map? Whether the map displays property lines, harvest boundaries, forest stand locations, or timber types, different aspects should be addressed to ensure that the map is actually representing what is truly on the ground.

Stehman and Czaplewski (1998) describe a fundamental structure for assessing the accuracy of thematic maps. Their structure has three basic steps in the process to determine the accuracy of a map: a response design, a sample design, and an estimation and analysis protocol. A response design is defined as the process in determining the reference classification for each sampling unit, generally a pixel or a polygon. The reference classification is defined as the "true" classification of that sampling unit, and can be determined by some combination of aerial imagery or visiting the sampling unit on the ground. The sampling design is the process by which the reference samples are selected for analyses. The sampling design consists of

defining a sampling frame, which includes a list or map of the entire target population, and defining the sampling units, which includes a list of individual points or areas from the sampling frame to be analyzed for an accuracy assessment. In order to determine the overall accuracy of the final map, an estimation or analysis protocol should be implemented by creating an error matrix to compare the actual and estimated values of each sampling unit or pixel.

Two main types of map error can occur, attribute error and locational error (Stehman and Czaplewski, 1998). Attribute error occurs when a thematic attribute, such as timber type, is inaccurate. Locational error occurs when a boundary is inaccurate. Locational error can be assessed by using a line intersect sample design to estimate the length of the boundary. Attribute error can be assessed by selecting a random sample of points to determine if the specified attribute was mapped correctly or incorrectly (Skidmore and Turner, 1992). Both of these types are critical in creating an accurate map; however, this project will only look at the attribute error of a map.

Non-Parametric versus Parametric Models

To derive forest cover types for a thematic map one can combine satellite imagery with data from field plots and impute a raster dataset showing a continuous map of the different cover types across the landscape. Previous studies have used both

non-parametric and parametric methods to predict forest attributes, including: Gradient Nearest Neighbor (GNN), Most Similar Neighbor (MSN), k-MSN, and the random forest nearest neighbor methods, and linear regression and geographic weighted regression.

Gradient nearest neighbor maps are created using a multivariate model that integrates field plot data with ancillary data, such as satellite imagery, and environmental data. This method uses the nearest neighbor, or shortest distance in feature space, from a point to the nearest plot to generate volume and basal area estimates that are then related to a specific timber type. The distance is measured by creating a weight matrix derived by canonical correspondence analysis. Most similar neighbor maps are created using a model that also integrates field plot data with satellite imagery, as well as topographic features such as slope and aspect. This method uses a canonical correlation analysis to derive a similarity function, with user specified relationships, to impute data where there are no ground plots. The k-MSN method uses the same methods as MSN, but takes an average of the k nearest plots. The random forest imputation method creates a classification matrix and regression tree in order to find similarities between the explanatory and response variables.

Nearest neighbor imputations have been used to perform multivariate analyses of forested landscapes by associating variables of interest to ground data, satellite

imagery, and light detection and ranging (LiDAR) data. Hudak et al. (2008) found that the random forest nearest neighbor method performed best at predicting various plot level estimates such as basal area and tree density in north-central Idaho. In Finland, Maltamo et al. (2006) compared k-MSN imputations for plot and stand level volume estimates and found that using aerial-laser scanner data resulted in better estimates than using aerial imagery estimates and when used together the resulting root mean square error improved again. Eskelson et al. (2009a) used nearest neighbor models to impute plot-level forest attributes, such as basal area, stems per hectare, volume and total gross oven dry weight biomass, and found that the random forest method performed best when compared to MSN and GNN imputation methods.

An alternative to the nearest neighbor imputation methods to estimate selected variables of interest is the use of parametric models. Linear and non-linear regression models have been used for this purpose in previous studies (Wang et al. 2005, Salas et al. 2010, Crow and Schlaegel 1988). Another option is the use of a geographic weighted regression (GWR) model. Fotheringham et al. (2002) developed the use a GWR model, which takes a global regression model and localizes it to a specific area and allows for relationships between the explanatory and response variables to account for spatial variations, by including a weighting function in the regression model.

Wang et al. (2005) developed an ordinary least squares (OLS) model, a spatial lag model and a GWR model to analyze the amount of net primary production (NPP) in forest ecosystems across China using predictor variables that included forest stand locations, forest inventory data, and remote sensing data. The authors found that the GWR model was superior to both the OLS model and the spatial lag model (SLM) in predicting the NPP, measured by the Akaike Information Criteria (AIC) and r-squared (R^2). The GWR model had an AIC of 4891 and a R^2 of 0.66. The OLS model had an AIC of 5036 and a R^2 of 0.58. Lastly, SLM returned an AIC of 5001 and a R^2 of 0.60.

Salas et al. (2010) modeled tree diameter using forest inventory and ancillary data. The models that the authors compared were OLS, generalized least squares (GLS), GWR, and linear mixed effects (LME). The authors used aerial LiDAR data and forest inventory plots to estimate diameter at breast height on individual trees in Norway. They found that the best performing model was LME; however, the GWR model also performed better than both the OLS and GLS model.

Aerial LiDAR

When current field inventory data are insufficient to assess if maps are accurate on a local-scale, a common practice is to revisit the forest and measure additional ground. This can be both costly and time consuming. A newer practice, becoming more available to forest land managers, is to use LiDAR data to acquire

detailed data over a larger landscape. LiDAR is a tool that forestry researchers and professionals are increasingly using to improve estimates of forest inventory attributes across larger landscapes, at a comparable cost to a traditional ground inventory data collection for some attributes (Hummel et al 2011).

LiDAR data have become a useful tool in obtaining large amounts of forest inventory data due to its precision and relative ease of ground truthing. Ground truthing the LiDAR data consists of installing plots randomly throughout the landscape, measuring trees on the plot, and georeferencing the trees so that one can locate specific trees in the LiDAR data set (Wulder et. al. 2008). LiDAR datasets can also be used to assess much larger areas of forested landscape at one time, rather than installing thousands of field plots.

Nelson et al. (2004) used LiDAR to estimate the amount of biomass and carbon in the state of Delaware. They used parallel flight lines 4 kilometers apart to measure the merchantable forest volume, biomass and above ground carbon. Using four explicitly linear models the authors predicted merchantable forest volume and above ground biomass across the state. The authors found that merchantable volume estimates were within 22% of USFS estimates county wide and 15% statewide. Additionally, the authors found that their biomass estimates were within 22% of USFS estimates county wide and 20% statewide. The USFS estimates were based on FIA

volume and biomass estimates at the county and state level. They concluded that forest volume and biomass can be estimated using a laser based transect sampling method.

Naesset (2004) reported on the first Nordic stand-based forest inventory using LiDAR. The author predicted six stand variables from LiDAR data: mean tree height, dominant height, mean diameter, basal area, stem volume and stem number. Plot and tree level data were collected, including tree diameter at breast height (dbh), tree height, and the spatial location of the tree. With the plot data the author calculated: mean height, dominant height, mean diameter by basal area, plot basal area, number of trees per hectare, and total plot volume. From the LiDAR data, a digital elevation model and canopy height model was determined. The author found that 85-95% of the variability was explained by the regression models for mean height and dominant height. Additionally, 72-85% of the variability was explained by the regression models for basal area and stand volume and 49-63% of the variability was explained by the regression models for mean diameter and stem number. Validation of the models revealed the mean differences between the ground truth data and the predicted values were statistically significant in 5 of 24 cases and no bias was detected.

Using LiDAR derived metrics and other remote sensing data as predictor variables, the current study assesses the accuracy of parametric and non-parametric methods for estimating the amount of standing tree biomass across the Malheur

National Forest, in Eastern Oregon, USA. The models were assessed for their accuracy by comparing measured ground plot values to the model estimates.

Materials and Methods

Project Site

The project site consists of two non-adjacent blocks of land on the Malheur National Forest, located in the Blue Mountains of eastern Oregon (Figure 1). The northern site contains 106,600 acres of the Camp Creek LiDAR data set. The southern site consists of 112,240 acres, consisting of the Damon and a portion of the LLP LiDAR data sets.

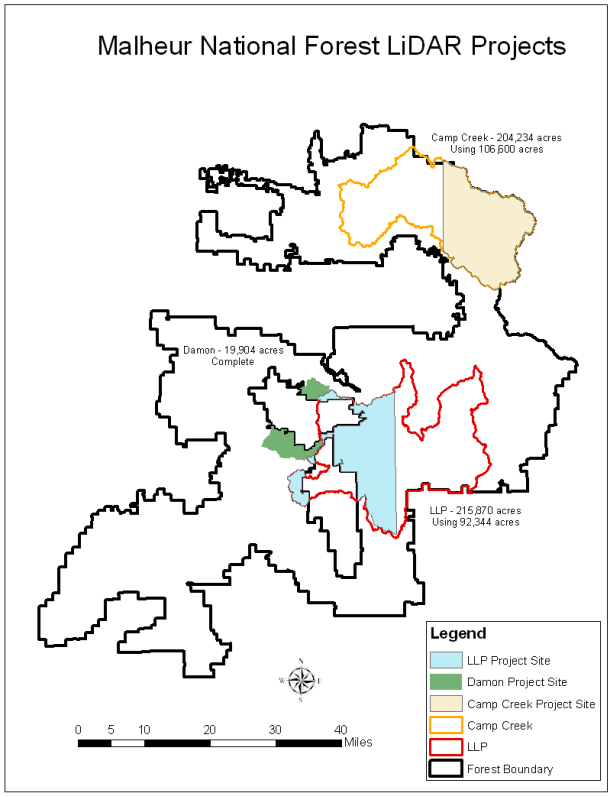

Figure 1: LiDAR datasets on the Malheur National Forest

Aerial LiDAR Data

The LiDAR data was collected from late 2007 through late 2008 by Watershed Sciences, Inc. Each of the three separate acquisition areas were obtained separately, during "leaf-off" conditions: the Damon site was collected on September 15 and 16, 2007, the Camp Creek site was collected from August 19th through August 27, 2008, and the LLP site was collected from November 19th through December 11, 2008.

The LiDAR acquisition used a Leica ALS50 Phase II laser mounted on a Cessna Caravan 208B. The scan angle was ±14° from nadir with a pulse rate designed to obtain an average number of pulses emitted by the laser of ≥4 points per square meter. The Leica ALS50 Phase II laser system is designed for up to four range measurements per pulse, and all laser returns were processed for the dataset. The Damon dataset had an average pulse density of 6 points per square meter, the Camp Creek dataset had an average pulse density of 8 points per square meter, and the LLP dataset had an average pulse density of 8 points per square meter.

Aircraft position was recorded by an onboard differential GPS unit, which measured the x, y, and z location of the aircraft twice per second (2 Hz). Aircraft altitude was measured 200 times per second (200 Hz) as pitch, roll and heading from an onboard inertial measurement unit.

Multiple GPS units were used for the ground real-time kinematic portion of the data collection process. The GPS base stations were set up on the monuments in order to broadcast a kinematic correction to a roving GPS unit. This allowed the ground surveyors to collect precise location measurements ($\sigma \leq 1.5$ cm). A total of 1,007 real-time kinematic ground points were recorded throughout the Damon site and were then compared to the LiDAR data for accuracy assessment.

Ground Data

Previously collected ground data consists of USFS Stand Exams from 2008 and current vegetation survey (CVS) plots measured between 1998 and 2007. The previously collected stand exams and CVS plots were grown forward to 2009 in the Forest Vegetation Software (FVS), Blue Mountain variant (Keyser and Dixon, 2008). Ten additional plots were measured during the summer of 2009 (Table 1).

Table 1: Number of Plots in Damon Site

Source	Number of Plots
USFS Current Vegetation System	10
USFS Stand Exams	98
Summer 2009	8

Recent research has shown that stratifying the landscape using LiDAR data is an efficient and effective way to group the landscape into similar forest type and structure for further analysis (Sullivan 2008, Koch et al. 2009, Leppanen et al. 2008).

Forested stands were delineated using differences in height and canopy closure characteristics. Percent canopy closure, 25th and 75th height percentiles were used for this process. Following the process outlined by Sullivan (2008), stand delineations were created using two software packages, FUSION (McGaughey 2009) and Spring (Câmara et al. 1996). Spring is a user-based classification software package; for this study, the stand density index (SDI) of USFS stand exam plots measured in 2006 was used for the training data of the user-based classification process.

Comparing different inventory designs is an important part of laying out a plan to collect inventory data. Various sample designs can be used and all different types should be considered and evaluated for a specific project. Stehamn (2009) discusses, in detail, the necessary pieces to have a proper sample design for assessing map accuracy. Although primarily describing a sample design for determining the accuracy of land cover classification, the theory behind what a statistically sound sample design should consist of remains the same for any type of map accuracy assessment. Additionally, the author weighs the advantages and disadvantages of different types of sample designs. Based on how a specific project is to be completed, the author lists three questions to answer in order to assist in determining what kind of sample design is ideal for that specific project: (1) Are pixels individual sample units or are they grouped in clusters and the clusters the sample units? (2) Are the sample units stratified? (3) Is the process to select a sample unit a simple random design or a systematic design? Of the ten sample designs identified, and seven design criteria, a

stratified random sample design was rated as having the most strengths and least weaknesses and cluster sample designs having the least strengths. However, the one area where cluster sample designs ranked higher than all other designs was in the cost effective criteria. This suggests that if a project is restricted by time and/or budget, a cluster sample design may be the best sample design to implement due to being the most cost effective, while still being a statistically sound design.

The 10 plots measured during the summer of 2009 followed the cluster sample design. Each linear cluster (CLUS) of plots consisted of four rectangular fixed radius subplots. Moisen et al. (1992) showed that linear clusters of plots was a cost efficient way of distributing forest inventory plots for assessing map accuracy, while accounting for spatial autocorrelation. The advantage of using a CLUS design is less cost in traveling to each plot, while the disadvantage for CLUS is that there is more potential for spatial autocorrelation. The main reason for using this plot design consisted of the limited amount of time to collect field plots. Additionally, the primary goal of collecting additional ground data was to assess the accuracy of the LiDAR dataset. By using a CLUS design, it was possible to sample more ground area in a limited amount of time, while not sacrificing the amount of plot estimates due to the availability of previously collected inventory data.

The linear clusters consisted of four 1/10-acre rectangular fixed area plots. In order to assure a random sample a grid of 1/10-acre plots was placed over the project

area and a random location was selected based on the plot allocation information previously computed. The other three plots were located by obtaining a random azimuth in one of the four cardinal directions, from the first plot center, and installing the three additional plots in a linear fashion.

In a plot each tree that was greater than or equal to 4.5 feet tall was measured for diameter-at-breast height (DBH), species, and crown dominance (dominant, co-dominant, intermediate, or over-topped). The first, third and fifth tree per species per plot were measured for height, crown diameter, and crown ratio. Additionally, two to three of the tallest trees per plot were geo-referenced for LiDAR ground truthing purposes. Crown diameter was measured by taking a random azimuth and measuring the diameter of the crown at that azimuth, then taking the diameter of the crown perpendicular to the first measurement and averaging the two. Dead trees and snags, greater than five inches DBH, were measured for DBH and height. All trees with broken tops were measured for height.

These ground data were collected on a TDS Ranger handheld computer, with the USFS Stand Exam software and the output was analyzed using the SAS Software (SAS Institute Inc., v9.2). Missing heights were estimated by re-fitting the FVS height-diameter equations for the Blue Mountains. The USFS published coefficients were used as a starting point to determine height-diameter equations specific to the

project site. The localized height-diameter equations were found using the PROC NLIN function in SAS Software (Appendix A).

The USFS Stand Exam plots consist of 98 plots that were measured in the summer of 2008. Stand Exam plots are a nested plot design that consists of a variable radius plot for large trees and fixed radius plots for small trees and seedlings. Ninety-eight stands were measured with this process in the Damon project site, then 1 plot in each stand was chosen at random and a professional forester from the USFS re-measured the plot so that a $1/10^{th}$ acre fixed plot was used for the large trees, instead of the previously measured variable radius plot design. These data were analyzed internally by the Forest Service within their plot compiler.

CVS plot data were also supplied by the USFS. The CVS data system is a database of permanent forest inventory plots in Region 6 (Pacific Northwest) of the USFS. Each plot is re-measured once every ten years. Within this study site, CVS plots are on a 1.7 mile systematic grid. The plots consist of a 2.47- acre circular plot with 5 sub-plots. Each sub-plot is a set of 3 plots: (1) 1/5.3- acre plot, (2) 1/24- acre plot, and (3) 1/100- acre plot. Each plot has set criteria for which data should be collected and recorded, including live and dead tree measurements, down woody debris, shrub and understory components, and general geographical and slope position information of the plot (US Forest Service, 2001).

Data Compilation

The total standing tree woody biomass (tons per acre) was estimated for each ground inventory plot. In this study, standing tree woody biomass is defined as the biomass of the bole, bark, and branches of all standing dead and live trees that are greater than or equal to 4.5 feet tall. Volume and biomass estimates were calculated using the USFS Forest Inventory Analysis (FIA) cubic volume, including top and stump, and biomass equations for the Blue Mountains (Appendix B). All results found in this study assume that the USFS FIA equations are correct and that all assumptions of the volume and biomass models will therefore pertain to this research as well.

LiDAR data analysis was performed with FUSION (McGaughey 2009). Using the batch processing tools within FUSION, the raw LiDAR data files were clipped to each individual ground inventory plot and attributes such as a digital elevation model (DEM), height percentiles, and their variances were obtained. Additionally, using the GridMetrics batch processing tool these same estimates were obtained for all other areas within the project level that did not have ground inventory data. The percent cover, percent slope, aspect, and elevation of each plot were found using the LiDAR derived DEM and analyzed with the Spatial Analyst Extension within ArcGIS (ESRI ArcGIS, 2010).

Landsat Thematic Mapper (TM) data was downloaded from the United States Geological Survey Global Visualization (GloVis) website for the entire project data.

All of the seven bands were brought into ArcGIS and the normalized difference vegetation index (ndvi) was determined using bands three and four.

Climate data from the DAYMET website (Thornton 2003) was downloaded for the entire project data. Variables of interest consisted of: average daily maximum temperature, average daily minimum temperature, average temperature, number of growing degree days, number of frost days, and total precipitation. All variables were merged into one large table on a 20x20 meter pixel grid using Hawth's Tools (Beyer 2004) and the SurfaceSpot command line function in ArcGIS. Additionally, each of the ground inventory plots was added as a separate row to the table.

Statistical Analysis

There are multiple methods in determining which explanatory variables should be used in running the nearest neighbor models (Latifi et al. 2010 and Goerndt et al. 2010). For this study, explanatory variables were determined for the nearest neighbor imputations and geographic weighted regression, by implementing a stepwise regression technique, as outlined by Goerndt et al. (2010), using the *regsubsets*() tool within the *leaps*() R-package (R Development Core Team, 2011). This tool returns the best fitting linear model, based on the independent variables that are determined using the Bayesian information criteria (BIC).

Using the eight independent variables found by the best fitting linear model, a geographic weighted regression (GWR) model was fit using the *gwr*() tool within the *spgwr*() R-package. Before a back transformation of the natural log biomass estimate was performed, a bias-correction factor of 0.5 times the mean square error was added to the estimates (Baskerville 1972, Goerndt et al. 2010). Most similar neighbor (MSN), gradient nearest neighbor (GNN), k-nearest neighbor (k-MSN), and random forest (RF) were performed using the *yai*() and *impute*() functions within the *yaImpute* R-package.

The accuracy of each model was assessed using the 116 plots located within the Damon project site and measured by calculating the root mean square error (RMSE) and bias using a leave one out cross-validation, with the following equations:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}}, \quad (1)$$

$$bias = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)}{n}, \quad (2)$$

where Y_i is the observed value, $\hat{Y}_I$ is the imputed estimate, and n is the sample size (number of plots).

Results and Discussion

The best linear model, for estimating biomass (tons per acre) on a plot included the following explanatory variables: the minimum value from the LiDAR height

percentile profile (Min_Elev), 80th percentile value of the height profile from the LiDAR data (P80), the longitudinal location of the plot (UTM_Y), the reflective property value of Landsat TM band 2 (LandsatB2), normalized difference vegetation index (ndvi), 18-year average daily minimum temperature (MinTemp), 18-year average of the number of growing degree days (DegDay), and the 18-year average of the annual precipitation (TotPrecip). The results of this linear model are summarized in Table 2.

Table 2: Coefficients and standard errors for linear regression model for ln(biomass) in tons per acre.

Variable	Coefficient	SE
Intercept	1115	154
LiDAR height percentile profile	-0.6847	0.2353
80th percentile value from the LiDAR height profile	0.0525	0.0165
UTM northing	-0.0003	0.0000
Reflective property of Landsat TM band 2	-0.1705	0.0411
Normalized Difference Vegetation Index	-6.382	1.359
18 year average of the daily minimum temperature	5.052	0.2276
18 year average of the number of growing degree days	0.0329	0.0049
18 year average of the annual precipitation	1.231	0.1741

Basal area per acre was used as a second response variable due to the GNN and random forest methods needing two y-variables to work properly. The best fitting

linear model, for estimating basal area per acre included the following variables: the standard deviation of all LiDAR returns on the plot (StdDev), the 95th percentile value of the height profile from the LiDAR data (P95), and the reflective property value from Landsat TM band 5 (LandsatB5). The results from this linear model are summarized in Table 3.

Table 3: Coefficient and standard errors for linear regression model for basal area (ft^2 per acre)

Variable	Coefficient	SE
Intercept	50.12	22.32
Standard deviation of all LiDAR returns on the plot	-27.79	5.212
95th percentile value from the LiDAR height profile	11.88	1.634
Reflective property of Landsat TM band 5	-0.7082	0.1908

The inventory plots varied in cover type, from non-forest meadows, to highly dense pine forests. Biomass measured on the inventory plots ranged from zero tons per acre to 103.7 tons per acre, with a standard deviation of 15.9 tons per acre. The basal area of the inventory plots ranged from zero square feet per acre to 248.7 square feet per acre, with a standard deviation of 55.6 square feet per acre (Table 4).

Table 4: Basic statistics of explanatory and response variables[1]

Biomass (tons per acre) Explanatory variables

units	Min_Elev meters	P80 meters	UTM_Y	LandsatB2 µm	ndvi	MinTemp celsius	DegDay degree days	TotPrecip cm
Minimum	1.00	0.00	4,882,625.7	23.0	0.2	-4.2	1,895.7	46.2
Maximum	4.42	33.9	4,901,661.6	39.0	0.7	-2.2	2,541.2	64.5
Mean	1.14	14.5	4,890,903.5	27.8	0.4	-2.9	2,298.9	54.0
Median	1.02	14.8	4,888,069.0	27.0	0.4	-2.8	2,312.5	53.9
Standard Deviation	0.38	6.34	6,759.8	3.5	0.1	0.5	168.0	4.2

Basal Area Explanatory Variables

units	Biomass tons per acre	Basal Area square feet per acre	StdDev meters	P95 meters	LandsatB5 µm
Minimum	0.0	0.0	0.0	0.0	47.0
Maximum	103.7	248.7	13.6	42.6	134.0
Mean	8.9	79.3	4.7	18.2	80.0
Median	2.9	77.1	4.4	18.2	75.0
Standard Deviation	15.9	55.6	2.3	7.7	19.8

[1]Min_Elev = Minimum value of the LiDAR percentile height profile. P80 = 80th percentile of the LiDAR height profile. UTM_Y = UTM northing coordinate. LandsatB2 = reflective property of Landsat TM band 2. Ndvi = normalized difference vegetation index. MinTemp = 18 year average of the minimum temperature. DegDay = 18 year average of the number of degree days. TotPrecip = 18 year average of the annual precipitation. StdDev = standard deviation of all LiDAR values on the plot. P95 = 95th percentile of the LiDAR height profile. LandsatB5 = reflective property of Landsat TM band 5.

Nearest neighbor imputations rely on explanatory variables being correlated with the response variables. Thus, the higher the correlation coefficient the better the imputation model should perform. The highest correlation between the predictor variables and biomass per acre comes from the LiDAR derived P80 variable, a correlation coefficient of 0.44 (Table 5).

Table 5: Correlation coefficients of biomass vs. selected predictor variables[2]

	ln_Biomass	ln_BA	Min_Elev	P80	UTM_Y	LandsatB2	ndvi	MinTemp	DegDay
ln_BA	0.4339								
Min_Elev	-0.2310	-0.0870							
P80	0.4368	0.5303	-0.0827						
UTM_Y	-0.3135	-0.1858	0.1243	-0.2442					
LandsatB2	-0.3320	-0.4832	0.1873	-0.5834	0.4484				
ndvi	-0.0516	0.1673	-0.0568	0.3494	-0.0614	-0.5555			
MinTemp	0.0321	-0.0374	0.2089	-0.0473	0.4424	0.2309	-0.2012		
DegDay	-0.1544	-0.0488	0.1485	-0.1336	0.4563	0.1331	-0.1158	0.5835	
TotPrecip	0.1158	-0.0064	-0.1164	0.0742	-0.1848	0.0158	0.1085	-0.4904	-0.9529

[2]Min_Elev = Minimum value of the LiDAR percentile height profile. P80 = 80th percentile of the LiDAR height profile. UTM_Y = UTM northing coordinate. LandsatB2 = reflective property of Landsat TM band 2. Ndvi = normalized difference vegetation index. MinTemp = 18 year average of the minimum temperature. DegDay = 18 year average of the number of degree days. TotPrecip = 18 year average of the annual precipitation.

The highest coefficient in the basal area prediction methods was the P95 variable, correlation coefficient of 0.69 (Table 6).

Table 6: Correlation Coefficients of basal area vs. selected predictor variables

	Biomass per acre	Basal area per acre	Standard Deviation of LiDAR returns	95th percentile value of LiDAR height profile
Basal area per acre	0.4372			
Standard Deviation of LiDAR returns	0.1691	0.5749		
95th percentile value of LiDAR height profile	0.1883	0.6870	0.9651	
Reflective property of Landsat TM band 5	-0.2282	-0.6225	-0.4757	-0.5477

The RMSE and bias for the nearest neighbor imputations and regression for biomass (tons per acre) and basal area (in square feet per acre) models are reported in Tables 6 and 7, respectively.

Table 7: RMSE and bias for estimating biomass (tons/acre) by selected method

Model	RMSE	Bias
Linear regression	12.7	-2.41
Geographic Weighted Regression	11.6	-0.67
Gradient Nearest Neighbor	16.31	-0.008
Most Similar Neighbor	13.96	-0.08
Random Forest	12.22	-1.87
k-MSN (k=3)	11.53	0.24
k-MSN (k=5)	11.24	-0.004

Table 8: RMSE and bias for estimating basal area (ft^2/acre) by selected method

Model	RMSE	Bias
Linear regression	33.15	0.0029
Geographic Weighted Regression	33.08	0.0082
Gradient Nearest Neighbor	58.65	-4.79
Most Similar Neighbor	50.99	0.13
Random Forest	39.03	2.82
k-MSN (k=3)	39.02	0.67
k-MSN (k=5)	38.62	0.71

For the biomass prediction, the k-MSN, k=5, has the lowest RMSE and least amount of bias. The second most accurate method consisted of the k-MSN, k=3,

followed by the GWR model and the RF imputation. The GNN method has the least amount of accuracy (Table 7). For the basal area prediction, the GWR model has the lowest RMSE and the least amount of bias. The second most accurate method consisted of the k-MSN, k=5, followed by the k-MSN, k=3 and the random forest model. The GNN method, again, has the least amount of accuracy (Table 8).

Possible reasons for GNN performing poorly, compared to the other models, consists of a very small sample size, the entire area of the project site is fairly small compared to previous uses of the GNN method, or the explanatory variables not being highly correlated with the response variables. The GWR method may be performing better than the non-parametric approaches due to only predicting one response variable, biomass. In contrast, the nearest neighbor methods are predicting both biomass and basal area simultaneously. Therefore, GWR may be sufficient for estimating biomass per acre if that is the only variable of interest; while, the nearest neighbor imputations are preferred when multiple response variables of interest are present in the analysis. When predicting a single variable, Eskelson et. al. (2009b) also reported that the parametric method resulted in more accurate estimates than the non-parametric nearest neighbor imputation methods.

The results of this study suggest that the current method being used to implement forest management activities on the Malheur National Forest, MSN, may not be the best method to predict total standing tree biomass. A better nearest neighbor

model may be k-MSN or RF. Whereas, if forest managers are only interested in a single response variable, total standing tree biomass, GWR may be a more suitable model.

Conclusion

If forest managers only need to predict standing tree biomass at a pixel-level, a GWR model may perform better than any of the non-parametric imputation methods (RMSE = 11.6, bias = -0.67) with predictor variables coming from LiDAR, Landsat TM imagery, climate data and ground forest inventory plots. If the desire is to predict more than one variable at a time, biomass and basal area, the k-MSN (k=5) model performed best (RMSE = 11.24, bias = -0.0004) of all nearest neighbor methods tested (GNN, MSN, k-MSN (k=3), and Random Forest). K-MSN and MSN will be further examined for their predictive abilities at varying geographic scales within the project site. Although MSN was not found to be the best performing method, it will be examined because it is the current model used by the USFS for forest planning purposes on the Malheur National Forest.

CHAPTER 3 – A COMPARISON OF THE SPATIAL ACCURACY OF SELECTED IMPUTATION METHODS

Introduction

Foresters are constantly writing forest plans to describe the activities to perform on a specific forest over time. When creating forest plans, foresters generally wish to use data that is specific to the location of their forest. Determining an estimate of volume or biomass can be an expensive and time-intensive task. Estimation methods such as the imputation methods previously discussed in Chapter 1 can assist foresters in determining estimates of forest attributes across their forests.

The accuracy of imputed maps is important to determine the best estimate of woody biomass in the most cost-efficient manner. An inaccurate estimate will result in poor planning and can lead to anything from loss in revenue on timber sales, treating areas that are a lower risk of large insect and disease outbreaks than other areas, and a loss in public trust. An accurate imputed woody biomass estimate at a project-level scale can lead to better forest plans. Forestland managers can be more confident in using these plans and the public can be ensured that forest managers have the proper tools to manage our federal forests. Additionally, accurate biomass estimates can result in a forest vegetation map that forest managers can use to pinpoint smaller areas, compared to large regional maps, for future treatments to reduce the amount of potential forestland that is susceptible to insect and disease outbreaks.

Using LiDAR derived metrics and other remote sensing data as predictor variables, in this study I compare the accuracy of the best performing imputation method from Chapter 2 (k-MSN, k=5) and the imputation method that the Malheur National Forest is currently using (MSN), for estimating the amount of standing tree biomass across a project area on the Malheur National Forest, in Eastern Oregon, USA. Forest managers need to know if these methods generate accurate results of imputing forest biomass at the project level (<50,000 acres), in order to write forest plans for their management areas and even a specific district on a forest. The selected methods were assessed over six different scale samples (5,000, 10,000, 20,000, 30,000, 40,000, and 50,000 acres) to determine if one method performed better than the other based on the size of the sample area.

Materials and Methods

Project Site

The project site for this study consists of the Camp Creek LiDAR project site on the Malheur National Forest, located in the Blue Mountains of eastern Oregon (Figure 1). The Camp Creek project site comprises of 106,600 acres.

Aerial LiDAR Data

The LiDAR data were collected in August 2008 and provided by Watershed Sciences, Inc., during "leaf-off" conditions.

The LiDAR acquisition used a Leica ALS50 Phase II laser mounted on a Cessna Caravan 208B. The scan angle was ±14° from nadir with a pulse rate designed to obtain an average number of pulses emitted by the laser of ≥4 points per square meter. The Leica ALS50 Phase II laser system is designed for up to four range measurements per pulse, and all laser returns were processed for the dataset. The Camp Creek dataset had an average pulse density of 8 points per square meter.

Ground Data

Ground data for this study were collected within the Camp Creek site following the same protocols as those described in Chapter two within the Damon site. Previously measured USFS CVS plots were grown forward in FVS to 2009 and additional CLUS plots were measured during the summer of 2009 (Table 9).

Table 9: Number of Plots in Camp Creek Site

Source	Number of Plots
USFS Current Vegetation System	53
Summer 2009	20

Data Compilation

The total standing tree biomass (tons per acre) was estimated for each ground inventory plot using the same methods as described in chapter two. Volume and biomass estimates were calculated using the USFS Forest Inventory Analysis (FIA) cubic volume, including top and stump, and biomass equations for the Blue Mountains (Appendix B). All results found in this study assume that the USFS FIA equations are correct and all assumptions of the volume and biomass models will therefore pertain to this research as well.

LiDAR data analysis was performed with FUSION (McGaughey 2009), as described in chapter two. Landsat Thematic Mapper (TM) was downloaded from GloVis and climate data was downloaded from the DAYMET website.

Creation of Scale Samples

Within ArcGIS a random sample of 50 pixels was taken six times, once for each of the six scale samples. For each of the six samples, each pixel was buffered according to a predetermined scale size (e.g. a specific pixel that was selected in the 5,000 acre scale sample was buffered in ArcGIS so that it had a 5,000 acre circle

around that specific pixel. Additionally, a 10,000 acre buffer was created for the 10,000 acre samples, 20,000 acre buffers for the 20,000 acre samples, 30,000 acre buffers for the 30,000 acre samples, 40,000 acre buffers for the 40,000 acre samples, and 50,000 acre buffers for the 50,000 acre samples). For each buffered area that fell across the edge of the project area, the buffered area was split along the project area boundary and moved back into the project area directly across from the originally randomly selected pixel. Each pixel within a buffered area was then selected and all the LiDAR metrics, satellite metrics, and climate metrics were exported for each pixel in the buffered area. This created a list of pixels with each of the explanatory variables that were chosen in the models from Chapter one. Each ground plot was then selected within each of the previously formed buffered areas and merged together with the pixels with all the explanatory variables. This resulted in 300 tables, 50 samples of each of the six scale samples, which would serve as the input values for the imputation runs.

Statistical Analysis

The nearest neighbor imputation methods were run using the same methods and variables as described in chapter two. The MSN imputation method was used because it is the current imputation used by the Malheur National Forest for imputing stand variables to write forest plans. The k-MSN (k=5) imputation method was used because it was the best performing method found in chapter two.

Each of the 300 scale sample tables were brought into Microsoft Access® and each scale sample was merged into one large table, keeping track of the original randomly selected pixel as the sample number, with one row in the table representing a single imputed pixel.

The accuracy of each method was assessed using the base run as the observed values and the imputed runs as the predicted values. The base run is defined as using all plots to impute all pixels within the project site. The root mean square difference (RMSD) and bias were calculated to compare the accuracy of the two imputation methods at the scale sample. The RMSD and bias were calculated on each of the individual buffered areas and then the average RMSD and bias were determined for each scale size, on each of the two imputation methods. The RMSD (Equation 3) and bias (Equation 4) were calculated using the following:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}}, \quad (3)$$

$$bias = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)}{n}, \quad (4)$$

where Y_i is the base run value, $\hat{Y}_I$ is the imputed estimate from the scale sample, and n is the sample size (number of pixels within the specific scale size).

Results and Discussion

The best linear model from Chapter 2, for estimating biomass (tons per acre) on a plot included the following explanatory variables: the minimum value from the LiDAR height percentile profile (Min_Elev), 80th percentile value of the height profile from the LiDAR data (P80), the longitudinal location of the plot (UTM_Y), the reflective property value of Landsat TM band 2 (LandsatB2), Normalized Difference Vegetation Index (ndvi), 18-year average daily minimum temperature (MinTemp), 18-year average of the number of growing degree days (DegDay), and the 18-year average of the annual precipitation (TotPrecip). However, for this analysis, the Min_Elev and UTM_Y variables were removed due to the range of values not covering a large enough gradient in values to be substantially different from one another.

The inventory plots varied in cover type, from non-forest meadows, to highly dense pine forests and mixed conifer forests. Biomass measured on the inventory plots ranged from zero tons per acre to 150.2 tons per acre, with a standard deviation of 28.1 tons per acre. The basal area of the inventory plots ranged from zero square feet per acre to 295.6 square feet per acre, with a standard deviation of 49.6 square feet per acre (Table 10).

Table 10: Summary Statistics of Explanatory and Response Variables[3]

	LandsatB2 (μm)	ndvi	Min Temp (Celsius)	DegDay	TotPrecip (cm)	P80 (meters)	Tons Biomass per acre	Basal Area square feet per acre
Minimum	20.0	0.2	-5.2	1518.3	52.5	8.2	0.0	0.0
Maximum	34.0	0.6	-1.3	2702.3	98.9	4836.3	150.2	295.6
Mean	25.5	0.4	-2.5	2329.9	65.8	92.4	47.9	104.1
Median	25.0	0.5	-2.4	2387.6	63.3	56.1	44.1	101.9
Standard Deviation	2.9	0.1	0.8	247.7	9.7	414.9	28.1	49.6

[3]LandsatB2 = reflective property of Landsat TM band 2. Ndvi = normalized difference vegetation index. MinTemp = 18 year average of the minimum temperature. DegDay = 18 year average of the number of degree days. TotPrecip = 18 year average of the annual precipitation. P80 = 80th percentile of the LiDAR height profile.

The results for the nearest neighbor imputations for biomass (tons per acre) for each of the six scale samples, while imputing biomass and basal area per acre, are reported in table 11.

Table 11: RMSD and bias for estimating biomass (tons/acre) by selected scale size

Scale Size	MSN RMSD	k-MSN RMSD	MSN Bias	k-MSN Bias
5,000	36.1	36.7	-14.0	-0.5
10,000	36.1	36.9	-8.9	-0.9
20,000	34.3	34.7	-3.3	-3.0
30,000	33.0	33.3	-2.6	-4.2
40,000	32.0	33.0	0.1	-2.3
50,000	40.0	38.7	2.8	3.2

Based on the RMSD, the MSN imputation method resulted in more accurate estimates of tons of biomass per acre than k-MSN (k=5) for all but the 50,000 acre scale sample (smaller RMSD). The smaller scale samples have slightly less bias for k-MSN, with MSN having less bias than k-MSN for the 30,000 and larger scale samples. These results suggest that MSN, the current imputation method being used by the Malheur National Forest, is slightly more accurate than k-MSN for smaller planning level scales (<50,000 acres). The k-MSN imputation method may be more accurate when creating biomass estimates for larger scales (50,000 acres).

A possible reason for k-MSN predicting more accurate results of biomass per acre at the largest scale sample is due to a larger sample size of plots being available for selection. Because the k-MSN method is averaging the 5 most similar plots to a given pixel, we would expect a lower average as we include more plots, which occurs as we increase the scale size.

Although the results suggest that MSN predicts the amount of standing tree biomass per acre slightly more accurately than the k-MSN method at smaller scales, we cannot say for sure that these results are conclusive and actually result in different estimates. This is due to the potential for multiple sources of error within the methods: different plot sizes, different plot shapes, and the error in the regression models used to measure biomass on a specific plot can all lead to various amounts of error in both estimating the amount of biomass on a specific plot and mapping the amount of

biomass across the project site. For example, the USFS CVS plots are circular forest inventory plots; however, the pixels imputed across the project site are square. This can result in a mapping registration error that is difficult to measure.

Conclusion

The results of this study suggest that MSN is a slightly more accurate imputation method than k-MSN (k=5) for smaller scales (<50,000 acres); however, both MSN and k-MSN (k-5) imputation methods result in unbiased estimates and therefore we cannot say conclusively that one method is better than the other. This suggests that the method the Malheur National Forest is currently using is just as accurate as the k-MSN (k=5) method when imputing the amount of biomass per acre across the Camp Creek project site for each tenth-acre pixel across various sampled scale sizes (<50,000 acres).

CHAPTER 4 – GENERAL CONCLUSION

This study suggests that depending on what forest managers want to know about their forest, various imputation methods can be used. If a forest manager would like to know just one piece of information, a non-parametric method, such as GWR, could be used in a cost-effective way to determine the amount of woody biomass over an approximately 20,000 acre area. However, if multiple forest inventory variables are desired, a forest manager may use the k-MSN or MSN imputation method to predict woody biomass and basal area at a project level scale (<50,000 acres), on the Malheur National Forest. These results could be beneficial to the Malheur National Forest in future forest plans due to the recently opened biomass facility that the Malheur Lumber Company opened in December of 2010. Using maps generated from these imputation methods could help to locate areas with high amounts of smaller timber that is more susceptible to insect or disease outbreaks and allow forest managers to complete thinning treatments to increase the health of these forests. This study also suggests that the use of LiDAR data as an explanatory variable in a regression model or nearest neighbor imputation method can increase the accuracy of estimated biomass per acre.

In the second chapter, I saw that k-MSN (k=5) and GWR imputed the most accurate, and unbiased, estimates of woody biomass in a pine dominated landscape that was less than 20,000 acres. This suggests that, for this study, if a forest manager has a relatively small area of land, a cost efficient way to predict multiple forest

inventory variables would be the k-MSN (k=5) imputation method and a cost efficient way to predict just the amount of woody biomass would be the use of GWR. The results for all tested methods resulted in unbiased estimates of woody biomass and basal area and relatively minor differences in the amount of accuracy between the various methods.

In the third chapter we examined a specific example of the best performing imputation method from chapter two and the currently used imputation method to predict multiple forest inventory variables on the Malheur National Forest. The results of this portion of the study suggest that MSN results in slightly more accurate results for smaller scales (<50,000 acres) when imputing biomass per acre and basal area per acre. However, at the 50,000 acre scale the k-MSN imputation method yielded more accurate estimates of biomass per acre. Once again, as in the first part of the analysis, all imputed results were determined to be unbiased and therefore we cannot say for sure that one method will guarantee a more accurate prediction of biomass per acre.

This suggests that, although the results in the second chapter within mostly pine dominated stands were unbiased and accurate; a blanket imputation method may not be suitable for areas of various forest types. The second chapter and third chapter results contradict each other somewhat. In chapter two I saw that k-MSN (k=5) had predicted biomass per acre more accurately than MSN. However, in chapter three I saw that for the smaller scales, MSN resulted in slightly more accurate results than k-

MSN. The forest type in the Damon site, used in chapter two, is dominated by pine stands, whereas the forest type in the Camp Creek site, used in chapter three, ranges from pine stands to mixed conifer, true fir stands.

Future studies could examine how imputation methods perform in different forest types. Whether one imputation method results in more accurate results in a mixed conifer, pine, or Westside Douglas-fir forest type could help land managers determine the best imputation method for their specific forest. This could also help to determine if large scale, regional analyses using a single imputation method are the best alternative for land managers. Additional studies could also inspect how the sample size of reference plots affects the results of selected imputation methods. I would predict that a larger sample size of reference plots could result in more accurate estimates of biomass per acre. However, as the number of ground plots measured increases, so does cost. Determining the most cost efficient number of ground plots to serve as reference plots in these imputation methods could help to reduce the amount of cost a land manager needs to spend while utilizing LiDAR data. Other future studies could examine how selected imputation methods perform based on imputing pixels versus imputing stand level forest variables. These future studies could assist land managers of the Malheur National Forest in determining areas of high amounts of small woody biomass that could be more susceptible to various insect and disease outbreaks and help to provide a regular supply of biomass to the Malheur Logging Company biomass facility in John Day, Oregon.

LITERATURE CITED

Baskerville, G.L. 1972. Use of logarithmic regression in the estimation of plant biomass. Canadian Journal of Forestry. 2: 49-53.

Beyer, H. 2004. Hawth's analysis tools for ArcGIS. Retrieved from: http://www.spatialecology.com/.

Câmara, G., Souza, R., Freitas, U., and Garrido, J. 1996. SPRING: Integrating Remote Sensing and GIS by Object-Oriented Data Modeling. Computers and Graphics. 20(3): 395-403.

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

Crow, T.R. and Schlaegel, B.E. 1988. A Guide to Using Regression Equations for Estimating Tree Biomass. Northern Journal of Applied Forestry. 5(1): 15-22.

Eskelson, B.N.I., Temesgen, H., and Barrett, T.M. 2009a. Estimating current forest attributes from paneled inventory data using plot-level imputation: A study from the Pacific Northwest. Forest Science. 5(1): 64-71.

Eskelson, B.N.I., Temesgen, H., and Barrett, T.M. 2009b. Estimating cavity tree and snag abundance using negative binomial regression models and nearest neighbor imputation methods. Canadian Journal of Forest Research. 39: 1749-1765.

ESRI ArcGIS, 2010. Version 10. http://www.esri.com.

Fotheringham, A.S., Brunsdon, C., and Charlton, M. 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley, Chichester, Hoboken, NJ. 273 pgs.

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

Hudak, A.T., Crookston, N.L., Evans, J.S., Hall, D.E. and Falkowski, M.J. 2008. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. Remote Sensing of Environment. 112(5): 2232-2245. Corrigendum: Remote Sensing of Environment. 2009. 113(1): 289-290.

Hummel, S., Hudak, A.T., Uebler, E.H., Falkowski, M.J., and Megown, K.A. 2011. A comparison of accuracy and cost of LiDAR versus stand exam data for landscape management on the Malheur National Forest. Journal of Forestry. 109 (5): 267-273.

Keyser, C.E. and Dixon, G.E. comps. 2008 (revised February 3, 2010). Blue Mountains (BM) Variant Overview – Forest Vegetation Simulator. Internal Rep. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Forest Management Service Center. 46p.

Koch, B., Straub, C., Dees, M., Wang, Y. and Weinacker, H. 2009. Airborne laser data for stand delineation and information extraction. International Journal of Remote Sensing. 30(4): 935-963.

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

Leppanen, V.J., Tokola, T., Maltamo, M. Pusa, T. and Mustonen, J. 2008. Automatic delineation of forest stands from lidar data. GEOBIA, 2008 – Pixels, Objects, Intelligence: GEOgraphic Object Based Image Analysis for the 21st Century. Proceedings. University of Calgary, Alberta, Canada. Pgs 271-277.

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

McGaughey, R.J. 2009. FUSION/LDV: Software for LIDAR Data Analysis and Visualization, Version 2.9. USFS. Retrieved from http://www.fs.fed.us/eng/rsac/fusion/.

Moeur, M. and Stage, A.R. 1995. Most similar neighbor: An improved sampling inference procedure for natural resource planning. Forest Science. 41(2): 337-359.

Moisen, G. G., Edwards, Jr., T.C., and Cutler, D.R. 1992. Spatial sampling to assess classification accuracy of remotely sensed data, In *Environmental Information Management and Analysis: Ecosystem to Global Scales*. (J. Brunt, S. S. Stafford, and W. K. Michener, Eds.), Taylor and Francis, Philadelphia, PA. Pages 161-178.

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

Næsset, E. 2004. Accuracy of forest inventory using airborne laser scanning: evaluating the first Nordic full-scale operation project. Scandinavian Journal of Forest Research. 19(6): 554-557.

Nelson, R., Short, A., and Valenti, M. 2004. Measuring biomass and carbon in Delaware using an airborne profiling LiDAR. Scandinavian Journal of Forest Research. 19(6): 500-511.

R Development Core Team. 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Salas, C., Ene, L., Gregoire, T.G., Næsset, E., and Gobakken, T. 2010. Modelling tree diameter from airborne laser scanning derived variables: A comparison of spatial statistical models. Remote Sensing of Environment. 114: 1277-1285.

SAS Institute Inc. Output for this paper was generated using SAS software, Version 9.2 of the SAS System for Windows. Copyright © 2011 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

Skidmore, A.K. and Turner, B.J. 1992. Map accuracy assessment using line intersect sampling. Photogrammetric Engineering & Remote Sensing. 58(10): 1453-1457.

Stehman, S. V. and Czaplewski, R. L. 1998. Design and analysis for thematic map accuracy assessment: fundamental principles. Remote Sensing of Environment. 64: 33-344.

Stehman, S. V. 2009. Sampling designs for accuracy assessment of land cover. International Journal of Remote Sensing. 30(20): 5243-5272.

Sullivan, A. 2008. LIDAR Based Delineation in Forest Stands. Master's Thesis. College of Forest Resources, University of Washington. 90 pgs.

Thornton, P.E. 2003. DAYMET climatological summaries for average air temperature and total precipitation (18-year mean for 1980-1997). Retrieved from http://www.daymet.org. University of Montana, Numerical Terradynamic Simulation Group, Missoula, MT, USA.

US Forest Service. 2001. Region 6 Inventory & Monitoring System: Field Procedures for the Current Vegetation Survey. Natural Resource Inventory, Pacific Northwest Region. USDA Forest Service. Portland, OR, USA. Version 2.04. 25 pgs.

Wang, Q., Ni, J., and Tenhunen, J. 2005. Application of a geographically-weighted regression analysis to estimate net primary production of Chinese forest ecosystems. Global Ecology and Biogeography. 14: 379-393.

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

Appendix A

Predicted height of a tree, in feet (dbh = diameter at breast height in inches):

Grand fir

$$4.5 + (804.96279413 * e^{(-6.0757637617 * dbh^{-0.31595287})})$$

True fir

$$4.5 + (135. -7442011 * e^{(-5.1356015043 * dbh^{-0.641053503})})$$

Western juniper

$$4.5 + (1818.1733 * e^{(-6.8482 * dbh^{-0.2535})})$$

Western larch

$$4.5 + (495.64722211 * e^{(-4.859041682 * dbh^{-0.365412454})})$$

Lodgepole pine

$$4.5 + (629.29624111 * e^{(-5.1668423772 * dbh^{-0.31589086})})$$

Engelmann Spruce

$$4.5 + (379.34205546 * e^{(-5.23595711 * dbh^{-0.397073086})})$$

Ponderosa pine

$$4.5 + (4196.2925115 * e^{(-7.521440746 * dbh^{-0.209421035})})$$

Douglas-fir

$$4.5 + (694.18218218647 * e^{(-5.5423617437 * dbh^{-0.307231871})})$$

Appendix B

FIA Volume and Biomass Equations – updated January 13, 2010

Volume calculated is the cubic foot volume, including the top and stump (CVTS). DBH is measured in centimeters and HT is measured as total height of the tree in meters. CVTSL is the log transformed estimate of the cubic foot volume, including top and stump.

Biomass calculations estimate the total live tree biomass of the bole, branches and bark. All Bark and Branch biomass equations result in Kilograms; to convert to tons, multiply by 0.0011023.

All true fir species

$$CVTS = 10^{CVTSL}$$

$$CVTSL = -2.502332 + 1.864963 * \log(DBH/2.54) + 1.004903 * \log(HT/0.3048)$$

$$Biomass\ of\ stem = (CVTS * Wood\ Density)/2000$$

$$Wood\ Density = (Specific\ gravity\ of\ tree\ species) * 62.4 lbs/ft^3$$

Subalpine fir

$$Bark\ Biomass = 1.0 + 17.2 * (DBH/100)^2 * (HT)$$

$$Live\ Branch\ Biomass = 3.6 + 44.2 * (DBH/100)^2 * (HT)$$

Grand fir

$$Bark\ Biomass = 0.6 + 16.4 * (DBH/100)^2 * (HT)$$

$$Live\ Branch\ Biomass = 13.0 + 12.4 * (DBH/100)^2 * (HT)$$

Western juniper

$$CVTS = 0.005454154 * \left[0.30708901 + 0.00086157622 * HT/0.3048 - 0.0037255243 * DBH/2.54 * \frac{HT/0.3048}{HT/0.3048-4.5}\right] * \left(\frac{DBH}{2.54}\right)^2 * HT/0.3048 * \left(\frac{HT/0.3048}{HT/0.3048-4.5}\right)^2$$

$$Biomass\ of\ stem = (CVTS * Wood\ Density)/2000$$

$$Wood\ Density = (Specific\ gravity\ of\ tree\ species) * 62.4 lbs/ft^3$$

$$Bark\ Biomass = e^{(-10.175+2.6333*ln(DBH*\pi))}$$

$$Live\ Branch\ Biomass = e^{(-7.2775+2.3337*ln(DBH*\pi))}$$

Western larch

For DBH>2 inches

$$CVTS = 10^{CVTSL}$$

$$CVTSL = -2.624325 + 1.847123 * \log(DBH/2.54) + 1.044007 * \log(HT/0.3048)$$

$$Biomass\ of\ stem = (CVTS * Wood\ Density)/2000$$

$$Wood\ Density = (Specific\ gravity\ of\ tree\ species) * 62.4 lbs/ft^3$$

$$Bark\ Biomass = 2.4 + 15.0 * (DBH/100)^2 * (HT)$$

$$Live\ Branch\ Biomass = 12.6 + 23.5 * (DBH/100)^2 * (HT)$$

Lodgepole pine

$$CVTS = 10^{CVTSL}$$

$$CVTSL = -2.615591 + 1.847504 * \log(DBH/2.54) + 1.085772 * \log(HT/0.3048)$$

$$Biomass\ of\ stem = (CVTS * Wood\ Density)/2000$$

$$Wood\ Density = (Specific\ gravity\ of\ tree\ species) * 62.4 lbs/ft^3$$

$$Bark\ Biomass = 3.2 + 9.1 * (DBH/100)^2 * (HT)$$

$$Live\ Branch\ Biomass = 7.8 + 12.3 * (DBH/100)^2 * (HT)$$

Engelmann spruce

$$CVTS = 10^{CVTSL}$$

$$CVTSL = -2.539944 + 1.841226 * \log(DBH/2.54) + 1.034051 * \log(HT/0.3048)$$

$$Biomass\ of\ stem = (CVTS * Wood\ Density)/2000$$

$$Wood\ Density = (Specific\ gravity\ of\ tree\ species) * 62.4 lbs/ft^3$$

$$Bark\ Biomass = 4.5 + 9.3 * (DBH/100)^2 * (HT)$$

$$Live\ Branch\ Biomass = 16.8 + 14.4 * (DBH/100)^2 * (HT)$$

Ponderosa pine

For DBH≥5 inches

$$CVTS = e^{CVTSL}$$

$$CVTSL = -8.521558 + 1.977243 * \ln(DBH/2.54) - 0.105288 * (\ln(HT/0.3048))^2 + \frac{136.0489}{(\frac{HT}{0.3048})^2} + 1.99546 * \ln(\frac{HT}{0.3048})$$

$$Biomass\ of\ stem = (CVTS * Wood\ Density)/2000$$

$$Wood\ Density = (Specific\ gravity\ of\ tree\ species) * 62.4 lbs/ft^3$$

$$Bark\ Biomass = e^{(-3.6263+1.34077*ln(DBH)+0.8567*ln(HT))}$$

$$Live\ Branch\ Biomass = e^{(-4.1068+1.5177*ln(DBH)+1.0424*ln(HT))}$$

For DBH<5 inches

$$CVTS = TARIF * TERM$$

$$TERM = \left[\left(1.033 * \left(1.0 + 1.382937 * e^{\left(-4.015292*\left(^{DBH/2.54}/_{10}\right)\right)}\right)\right) * (BA + 0.087266) - 0.174533\right]$$

$$TARIF = TARIF_{TMP} * [0.5 * (DBH_{TMP} - DBH/2.54)^2 + (1.0 + 0.063 * (DBH_{TMP} - DBH/2.54)^2)]$$

$$if\ TARIF \leq 0.0, then\ TARIF = 0.01$$

$$TARIF_{TMP} = \frac{(CV4_{TMP}*0.912733)}{(BA_{TMP}-0.087266)},$$

$$if\ TARIF_{TMP} \leq 0.0, then\ TARIF_{TMP} = 0.01$$

$$CV4_{TMP} = CF4_{TMP} * BA_{TMP} * HT/0.3048$$

$$CF4_{TMP} = 0.402060 - 0.899914 * \left({}^{1}/_{DBH_{TMP}}\right)$$

$$if\ CF4_{TMP} < 0.3, then\ CF4_{TMP} = 0.3$$

$$if\ CF4_{TMP} > 0.4, then\ CF4_{TMP} = 0.4$$

$$BA_{TMP} = DBH_{TMP}^2 * 0.005454154$$

$$BA = DBH/2.54^2 * 0.005454154$$

$$DBH_{TMP} = 6.0$$

$$Biomass\ of\ stem = (CVTS * Wood\ Density)/2000$$

$$Wood\ Density = (Specific\ gravity\ of\ tree\ species) * 62.4 lbs/ft^3$$

$$Bark\ Biomass = e^{(-3.6263+1.34077*ln(DBH)+0.8567*ln(HT))}$$

$$Live\ Branch\ Biomass = e^{(-4.1068+1.5177*ln(DBH)+1.0424*ln(HT))}$$

Douglas-fir

$$CVTS = e^{CVTSL}$$

$$CVTSL = -6.110493 + 1.81306 * \ln(DBH/2.54) + 1.083884 * \ln(HT/0.3048)$$

$$Biomass\ of\ stem = (CVTS * Wood\ Density)/2000$$

$$Wood\ Density = (Specific\ gravity\ of\ tree\ species) * 62.4 lbs/ft^3$$

$$Bark\ Biomass = 3.6 + 18.2 * (DBH/100)^2 * (HT)$$

$$Live\ Branch\ Biomass = 12.6 + 23.5 * (DBH/100)^2 * (HT)$$

Western white pine

$$CVTS = 10^{CVTSL}$$

$$CVTSL = -2.615591 + 1.847504 * \log(DBH/2.54) + 1.085772 * \log(HT/0.3048)$$

$$Biomass\ of\ stem = (CVTS * Wood\ Density)/2000$$

$$Wood\ Density = (Specific\ gravity\ of\ tree\ species) * 62.4 lbs/ft^3$$

$$Bark\ Biomass = 1.2 + 11.2 * (DBH/100)^2 * (HT)$$

$$Live\ Branch\ Biomass = 9.5 + 16.8 * (DBH/100)^2 * (HT)$$

Quaking aspen

$$CVTS = 10^{CVTSL}$$

$$CVTSL = -2.672775 + 1.920617 * \log(DBH/2.54) + 1.074024 * \log(HT/0.3048)$$

$$Biomass\ of\ stem = (CVTS * Wood\ Density)/2000$$

$$Wood\ Density = (Specific\ gravity\ of\ tree\ species) * 62.4 lbs/ft^3$$

$$Bark\ Biomass = 1.3 + 27.6 * (DBH/100)^2 * (HT)$$

$$Live\ Branch\ Biomass = 1.7 + 26.2 * (DBH/100)^2 * (HT)$$