ABSTRACT

ROCKHILL, AIMEE PAULINE. Imputation of Missing Observations in Forest Inventories. (Under the direction of Dr. Bronson P. Bullock.)

Imputation techniques are widely used in studies that contain missing data, but they may produce biased parameter estimates and inappropriate variance estimates. If the imputation technique used does not accurately represent the variability in the data, the resulting confidence intervals will be incorrect. This research evaluated numerous imputation methods and tested their ability to predict missing individual total stem height given existing diameter at breast height data. Imputation techniques that were compared include case deletion, mean imputation, Expectation Maximization (EM) algorithm, multiple imputation, and Bayesian multiple imputation. Known tree heights were randomly removed from a loblolly pine inventory dataset to represent missing data for comparing the imputation techniques. Imputations were made on the missing data using each of the aforementioned techniques. Observed and estimated tree heights were compared and prediction intervals for individual heights were computed to ascertain inclusion frequencies. The imputed tree heights were used to calculate volume loss (cu ft.) had the heights been missing due to stand damage.

The multiple imputation and Bayesian multiple imputation techniques maintain the variability in the data without biasing estimates. The Expectation Maximization algorithm produced confidence intervals that most closely simulated the actual dataset, however the technique overestimated the mean and underestimated the standard deviation. The Expectation Maximization algorithm, multiple imputation and Bayesian multiple imputation
slightly overestimated the volume loss by 32, 55, and 6 cubic feet, respectively. The actual total volume loss was 117,378 cubic feet.

The Bayesian proper multiple imputation technique was further applied to data collected after an ice storm in 2002 on the North Carolina State University Hill Demonstration Forest. An estimate of the volume loss from the ice storm using imputed height data was calculated that is unbiased and has a representative variance structure similar to that of the original data. A total volume of 13,537 cubic feet was lost on 502 acres of the Hill Demonstration Forest due to the ice storm.

Use of these imputation techniques give land managers another tool to aid in making management decisions when data are incomplete. The findings suggest that Bayesian multiple imputation can efficiently estimate missing height and, by extension, volume loss through the use of the R program package, however, the technique may be time consuming. Given limited time and resources, the Expectation Maximization algorithm in SAS provides a quick and efficient means of estimating unknown tree height given diameter measurements.
Imputation of Missing Observations in Forest Inventories

by

Aimeé Pauline Rockhill

A thesis submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the degree of Master of Science in Forestry

Raleigh, North Carolina

2008

APPROVED BY:

______________________________  ______________________________
Dr. Steven E. McKeand                            Dr. George R. Hess
______________________________  ______________________________
Dr. Edward L. Boone                             Dr. Bronson P. Bullock
                                  External Member                          Chair of Advisory Committee
BIOGRAPHY

Aimeé Rockhill (formerly Salstead) was born on May 6, 1976 in Ticonderoga, New York. She graduated from Ticonderoga High School in May 1994. Through persistent persuasion by her parents, she continued her education at Plattsburgh State University of New York, Plattsburgh, New York. In January 1998, she was awarded the opportunity to study abroad for a semester at the University of Queensland, Brisbane, Australia. She returned to SUNY Plattsburgh and graduated in December 1998 with a Bachelors of Art in Environmental Science. She worked as a fulltime substitute teacher in the Math and Sciences at Crown Point Central School, Crown Point, New York for 1 year. In 2000, she moved to Raleigh, North Carolina and began work at North Carolina State University in the Genome Research Lab under the direction of Dr. Charles Opperman. After several years of working on the Tobacco Genome Initiative Project, she began to pursue a degree in Forestry. In August 2004, she was officially accepted into the College of Natural Resources, Department of Forestry and Environmental Resources to pursue a Masters of Science in Forestry. Aimee is currently pursuing a Ph.D. in the Department of Forestry and Environmental Resources, Fisheries and Wildlife Program under the direction of Dr. Christopher S. DePerno.
ACKNOWLEDGEMENTS

I have many people to thank for where I am today and where my future takes me. I would like start by extending a sincere thank you to Dr. Charlie Opperman, Dr. Bryon Sosinski, and Dr. Steve Lommel who are the reason I am where I am today. Thank you for giving me my start at North Carolina State and being supportive of my decision to follow my heart and pursue a degree in Forestry. Thank you to my advisor, Bronson Bullock, for his continued support and guidance over the past 2 ½ years. It has been a long journey, but a great one, and I couldn’t have done it without you! My committee, Dr. Edward Boone, Dr. Steve McKeand, and Dr. George Hess; thank you for your guidance and support and for sticking with me through the change of my project. Dr. Lee Allen for the generous use of his dataset and for teaching me my first bird calls on early Raleigh mornings. I would like to say a special thanks to Dr. Bob Abt, Dr. Glenn Catts, and Dr. Barry Goldfarb for their continuous help and guidance on my projects. It is great to be surrounded by faculty that is always willing to take time out to guide and steer students in the proper direction. Thank you to Dr. Chris DePerno for supporting me through the last leg of my thesis and providing me with the opportunity to pursue a Ph.D. in Fisheries and Wildlife. I look forward to having you as my mentor for the next 3-8 years! Thank you to my parents, Debbie and Paul Salstead, who always say that things seem to work out for me but have yet to realize that it’s because of their guidance and support. Thank you to my older sister and her husband, Nicole and Brad Condon, for their constant support and faith in me and for giving me two beautiful nieces, Lauren and Kiersten, that are a constant reminder that we should never forget to have fun in life. Thanks to my younger sister and her husband, Sara and Matt Bowers, for their love and
support and for being great camping partners and getting me out in the woods when I am in desperate need of a get-a-way! Last, but most certainly not least, thank you to my husband Steve who was and is always there for me with a shoulder to lean on, an ear to talk into and a unbiased opinion when I need advice. You are not only my husband, but my best friend and I would not be writing this today without your continued love and encouragement. Thank you for being my Rock!
# TABLE OF CONTENTS

LIST OF TABLES ........................................................................................................................................ vi

LIST OF FIGURES ....................................................................................................................................... vii

1  Background ..................................................................................................................................................1

   1.1  Introduction ..........................................................................................................................................1

   1.2  Objectives .........................................................................................................................................3

   1.3  Literature Review ..............................................................................................................................3

   1.4  Assumptions .....................................................................................................................................10

   1.5  Missing Data Patterns ......................................................................................................................11

   1.6  Missing Data Mechanisms .............................................................................................................13

2  Materials and Methods .............................................................................................................................25

   2.1  Data ..................................................................................................................................................25

   2.2  Methods ..........................................................................................................................................35

3  Results ..................................................................................................................................................43

   3.1  Henderson Study Site ......................................................................................................................43

   3.2  Hill Demonstration Forest Ice Damage Assessment ......................................................................62

4  Discussion and Conclusions .......................................................................................................................65

LITERATURE CITED ....................................................................................................................................68

APPENDIX ..................................................................................................................................................71
# LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>Monotone Missing Data Pattern</td>
<td>12</td>
</tr>
<tr>
<td>Table 2</td>
<td>Arbitr ary Missing Data Pattern</td>
<td>12</td>
</tr>
<tr>
<td>Table 3</td>
<td>Estimate Efficiency for Various Levels of Imputation</td>
<td>19</td>
</tr>
<tr>
<td>Table 4</td>
<td>Combined Plot Data Summary</td>
<td>28</td>
</tr>
<tr>
<td>Table 5</td>
<td>Henderson Data Summary by Age - All Treatments</td>
<td>28</td>
</tr>
<tr>
<td>Table 6</td>
<td>Hill Forest Plot B Non-Missing Data Summary</td>
<td>34</td>
</tr>
<tr>
<td>Table 7</td>
<td>Henderson Dataset Summary</td>
<td>35</td>
</tr>
<tr>
<td>Table 8</td>
<td>Missing Data Summary by Imputation Method</td>
<td>49</td>
</tr>
<tr>
<td>Table 9</td>
<td>Confidence Interval Coverage Frequency by Imputation Method</td>
<td>59</td>
</tr>
<tr>
<td>Table 10</td>
<td>Volume (cubic feet) Summary Chart for Actual Data and Five Imputation Methods</td>
<td>61</td>
</tr>
<tr>
<td>Table 11</td>
<td>Volume Loss Estimates By Technique</td>
<td>62</td>
</tr>
<tr>
<td>Table 12</td>
<td>Hill Forest Imputed Data Statistics</td>
<td>63</td>
</tr>
<tr>
<td>Table 13</td>
<td>Hill Forest Volume Estimates for Block B</td>
<td>64</td>
</tr>
<tr>
<td>FIGURE</td>
<td>TITLE</td>
<td>PAGE</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Figure 1</td>
<td>Multiple Imputation Flow Chart</td>
<td>17</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Independent Sequences of a Markov Chain Simulation before Convergence</td>
<td>20</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Study Site Locations</td>
<td>27</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Diameter Distribution for the Completed Dataset</td>
<td>29</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Height Distribution for the Completed Dataset</td>
<td>30</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Diameter Distribution of Hill Forest Data</td>
<td>33</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Height Distribution of Hill Forest Data</td>
<td>34</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Regression of the Observed Dataset</td>
<td>37</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Time Series Plot using the Worst Linear Function</td>
<td>44</td>
</tr>
<tr>
<td>Figure 10</td>
<td>Autocorrelation Plot using the Worst Linear Function</td>
<td>45</td>
</tr>
<tr>
<td>Figure 11</td>
<td>Time Series Plot using the Mean</td>
<td>46</td>
</tr>
<tr>
<td>Figure 12</td>
<td>Autocorrelation Plot using the Mean</td>
<td>47</td>
</tr>
<tr>
<td>Figure 13</td>
<td>Observed Tree Heights and DBH Measurements for the Data with the Missing Data Deleted for the Case Deletion Technique</td>
<td>51</td>
</tr>
<tr>
<td>Figure 14</td>
<td>Observed Tree Heights and DBH Measurements for the Data with Imputed Mean Heights for the Mean Imputation Technique</td>
<td>52</td>
</tr>
<tr>
<td>Figure 15</td>
<td>Imputed Heights Using the Expectation-Maximization Algorithm Technique</td>
<td>53</td>
</tr>
</tbody>
</table>
Figure 16  Imputed Heights for Imputation 2 of the Bayesian Multiple Imputation Technique ..........................................................54

Figure 17  Imputed Heights for Imputation 5 of the Bayesian Multiple Imputation Technique ......................................................................................55

Figure 18  Imputed Heights for Imputation 3 of the Multiple Imputation Technique ......................................................................................56

Figure 19  Imputed Heights for Imputation 5 of the Multiple Imputation Technique ......................................................................................57
1 Background

1.1 Introduction

Imputation is the process of filling in missing data. Imputation methods have been implemented in many fields, typically in the social sciences, to fill in missing data in surveys. Forest inventory data usually has missing information in the datasets, making them difficult to analyze and compare. Imputation techniques could be used in forest inventory to complete the datasets with missing information or reduce the number of measured trees in the field, saving time and money. This research evaluated five mechanisms for addressing missing data in forest inventories. Imputation techniques can be used to aid in obtaining information from datasets with missing information by filling in the missing data with estimates derived from the observed data. Parameter estimates can be developed for incomplete datasets that reflect the characteristics of complete datasets, saving time and money versus the alternatives of disposing of the missing data or sending field crews out to re-measure the incomplete observations. Imputation techniques may also be used to impute information for datasets where observations were specifically not measured; though precautions would need to be taken to avoid bias due to the missingness of certain observations. This technique allows a reduction of time commitment on measuring sample plots while providing some information for estimating missing values (Robinson and Wykoff 2004).

Basic measurements typically needed for total stem volume or weight estimation are diameter at breast height (D) and total stem height (H). Measurements of D are typically quick and easy to obtain, while measurements of H are more time consuming.
Historically, missing heights for individual stems in datasets are generated using published models that exist for a number of species over a large geographic range and are often useful when dealing with even-age single species plantations (Pienaar et al. 1987). When not working with a single age class as a dependent variable in a model, the model is limited to diameter classes as the only dependent variable for determining height. Pienaar et al. (1987) found that one-inch diameter classes of loblolly pine (Pinus taeda L.) ranging from 2 to 14 in. in the Lower Coastal Plain of NC had total tree heights ranging from 15 to 85 ft. Typically, smaller (2-3 in.) and larger (13-14 in.) D trees have smaller height ranges. For example, Pienaar et al. (1987) found that a one-inch diameter class of 13 in. contained heights ranging from 65 to 85 ft. while a one-inch diameter class of 7 in. contained heights that ranged from 25 to 65 ft. When utilizing imputation techniques to account for missing data in forestry applications, it is important to obtain coverage of the distributional range of H in relation to D to maintain the variability present between the two.

Imputation techniques can also aid in estimating volume loss from hurricane, fire, or ice storm damage. Most landowners, both private and public, do not perform annual inventories on their timber. When an ice storm or hurricane strikes, the residual stand of trees may have numerous broken tops. Estimating volume loss on damaged stands can be a challenge when no recent stand data exists to produce a model that reflects the local conditions. Robinson and Wykoff (2004) state that given the previous data, imputed heights are a compromise between missing variation in global models and the limited power of local models. The following research demonstrates that, given an appropriate imputation technique, it is possible to estimate missing observations with an appropriate variance
structure that maintains variability and allows for accurate volume loss estimates.

1.2 Objectives

The objectives of this research are to:

1) Compare various imputation mechanisms by analyzing a simulated missing dataset to determine which imputation technique most accurately predicts total height of trees from observed diameter measurements while maintaining appropriate variability in the dataset

2) Introduce Bayesian Multiple Imputation as an imputation mechanism which will maintain variability in the dataset and produce unbiased volume estimates where data are missing

3) Use the best available imputation technique to analyze a dataset collected after an ice storm and estimate the volume loss on the stand

1.3 Literature Review

General Imputation

Rubin (1976) analyzed a number of datasets containing missing data to determine the weakest simple conditions that needed to be met for the data to be considered Missing At Random or for it to be appropriate to use case deletion. Rubin (1976) states that it is appropriate to ignore the process that causes missing data if the missing data are Missing At Random and the observed data are observed at random. The resulting inferences are
typically conditional on the missing data pattern of the observed data (Rubin 1976). Rubin (1976) concluded that it might be simpler to make proper Bayesian and likelihood inference in cases of missing data instead of assuming proper sampling distribution inferences; however, the same restrictions exist for Bayesian and likelihood inferences that exist for sampling distribution inferences. Rubin (1976) also concluded that statisticians rarely consider the process that causes missing data in a dataset and models for this process do not receive enough attention in the statistical literature.

Rubin and Schenker (1986) compared a number of Multiple Imputation techniques for simple random samples with ignorable nonresponse. Looking at discrete data, missing values drawn from observed values, simple random imputation, Bayesian bootstrap imputation and approximate Bayesian bootstrap imputation techniques were analyzed. These were compared with a number of imputation methods using continuous data: missing values drawn from observed and non-observed values, fully normal imputation, and imputation adjusted for uncertainty in the mean and variance. The methods were compared in terms of coverages (frequentist coverage probabilities) of the resulting interval estimates. Primary findings for large samples with non-normal data or all sample sizes of normal data were that more than one imputation was superior to single imputation and imputation works best when adjusted for uncertainty. In conclusion, Rubin and Schenker (1986) recommend the best imputation technique is one which adjusts for uncertainty in the estimation of the mean and the variance and is computed individually multiple times with the results combined.

Ake and Carpenter (2002) modeled survival data of patients in the Veterans Administration who had contracted HIV. The dataset consisted of four covariates with
missing values. The goal was to present suggestions for optimal use of SAS Version 8 imputation procedures; Proc MI, Proc MIAnalyze and Proc Phreg (SAS Institute Inc. 2006). Proc is short for procedure and is a command in SAS that invokes all of the forms of statistical analysis (SAS Institute Inc. 2006). For example, Proc MI calls upon the multiple imputation statistical analysis procedure described later in the methods section. Ake and Carpenter (2002) developed a loop that indexed the model version number with a call to Proc MIAnalyze in SAS (SAS Institute Inc. 2006) using Proc Phreg in SAS. Proc Phreg was used for individual analysis of the five complete datasets created in Proc MI before they used Proc MIAnalyze. Ake and Carpenter (2002) concluded that the MI and MIAnalyze procedures are powerful tools for imputing and estimating missing values in data, however, a level of complexity is added due to the analysis of multiple datasets and repeated analysis.

Paul et al. (2003) performed a case study in which they analyzed eight of the most commonly used techniques for imputation: casewise deletion, weighted casewise deletion, mean imputation, mean imputation with a dummy variable, conditional mean imputation, hotdeck imputation, approximate Bayesian bootstrap, and full Bayesian imputation. The research was an extension of that performed by Fox et al. (1998) and the same data were used for the analysis. Case deletion and weighted case deletion were found to be the poorer imputation techniques of the eight techniques assessed. Bayesian Multiple Imputation was not found to be superior to the other techniques. When analyzing a real dataset, both Bayesian techniques produced the best based on standard error, coefficient bias, and variance inflation. However, when simulated datasets were used and the predicted data were compared to the observed data, Paul et al. (2003) found the Bayesian techniques performed
well and did not exceed the performance of conditional mean imputation and hotdecking with respect to coefficient bias. An error was found and was due to a problem known as semi-complete separability, which is a result of a faulty imputation model, which was “over-taxing” the data. Semi-complete separability occurs when $Y \times Z$ combinations for data $X$ are the same and each has one missing respondent and one non-missing respondent. This causes the algorithm to choose a nonzero value for the estimated probability of an observed cell proportion for a possible value of a missing element when it should equal zero (Paul et al. 2003). Had the data not been simulated, semi-complete separability would not have been recognized and superiority of the Bayesian multiple imputation technique would have been assumed. In certain cases, it is possible to reduce the model to avoid semi-complete separability; however, it is important to use the proper model (Paul et al. 2003).

Nielsen (2003) looked at proper and improper multiple imputation techniques and argues that Bayesian multiple imputation is inefficient when used outside of the survey world. Nielsen (2003) declared that imputations drawn from a Bayesian predictive distribution are proper when the models used for imputation and analysis are compatible. Nielsen (2003) presents examples of situations in which Bayesian multiple imputation did not provide accurate results. In one example, a gamma prior distribution was used with a complete data model instead of the Maximum Likelihood Estimate, which led the complete data model to produce inflated confidence intervals due to large variances. In another example, Nielsen (2003) simulated a dataset with 20% missing data that were Missing At Random. An unrestricted variance matrix made Maximum Likelihood Estimate difficult which lead to an inefficient complete data estimator. It was found that the variance estimator
may be biased upward, leading to inefficient but correct inference, or downwards, leading to incorrect inference (Nielsen 2003).

Forestry Imputation

Hassani et al. (2004) compared tabular and Most Similar Neighbor imputation techniques to determine which predicted regeneration in cedar-hemlock stands in British Columbia. Imputation techniques were used instead of regression analysis because imputation allows estimates to be imputed based on several missing variables as opposed to regression, which needs one dependent variable. In the tabular imputation technique, regeneration tables were derived from data on 80% of the plots and used to impute regeneration estimates on the remaining 20% of the data. Hassani et al. (2004) then compared observed verses estimated regeneration values to determine the accuracy of the techniques. The Most Similar Neighbor approach utilized a set of 10 indicator variables to represent slope and site preparation and then used a similarity measure to determine the most similar neighbor. The two techniques were compared by calculating bias, Mean Absolute Deviation and root mean-squared error (RMSE). Hassani et al. (2004) found that Most Similar Neighbor performed better than tabular imputation; both methods produced reasonably accurate volume and regeneration estimates. While both methods produced similar RMSE values, the Most Similar Neighbor technique produced lower bias and Mean Absolute Deviation values. The overall results showed that although tabular imputation has the benefit of imputing with multivariate data in one step, Most Similar Neighbor imputation consistently produced better results.
Robinson and Wykoff (2004) compared six imputation strategies for applying height-diameter regression models to height imputation algorithms. They determined which model worked best by comparing the RMSE of the residuals from the imputation techniques and the RMSE of the prediction errors from a 2000-fold cross-validation. The data consisted of three Forest Vegetation Simulators using a published model, an unbiased log scale model, and an unbiased natural scale model. Three mixed effects models were introduced: fixed-effects only, stand-level best linear unbiased predictions, and point-level best linear unbiased predictions. Imputation strategies were compared and applied to the Forest Vegetation Simulator data. Robinson and Wykoff (2004) found that the Forest Vegetation Simulator published model over-predicted heights with a positive bias of one meter. The Forest Vegetation Simulator log scale and natural scale both produced less variable results that were similar. When analyzing the mixed-effects models, the fixed-effects only model resulted in a higher RMSE than the published model when estimating volume. Robinson and Wykoff (2004) also concluded that the most preferable model for both height and volume was a log-scale correction model and that intercept and mixed-effects slope corrections within stand and species are preferred to intercept only correction.

Van Deusen (1997) obtained annual measurement of forest plot data from the USDA Forest Service Inventory and Analysis Units with the objective of producing updated Forest Service Inventory and Analysis Units datasets. A number of data analysis techniques were analyzed including standard systematic sampling, time trend incorporation, single imputation, double sampling for regression and multiple imputation. Van Deusen (1997) determined that multiple imputation was the best technique to use in exchange for measuring
100% of the measurement plots as long as an equal proportion of the plots were sampled each year. Thus allowing companies an opportunity to reduce monitoring efforts and have little loss in precision.

Gartner and Reams (2001) compared a number of imputation techniques for estimating the volume of live trees per acre in Georgia to attempt to update inventory data that is typically taken on a 5-10 year cycle. Forest Service Inventory and Analysis Units data were taken 1988 and 1996 and stand volume data were deleted from 40 percent of the 1996 dataset containing unit, county plot, forest type, and stand origin data from. Several techniques were compared to determine the best estimator of volume including double sampling ratio estimator, single imputation group means, and multiple imputation using growth model predictions. Comparing standard errors and mean square error of the various techniques, Gartner and Reams (2001) determined that the matching techniques provided better estimates than the modeling techniques. Single imputation tended to underestimate the standard error and the double sampling ratio estimator produced standard errors close to the multiple imputation techniques. However, multiple imputation tended to outperform the rest of the techniques due to the higher standard errors that are more representative of the original data and the flexibility that allows for the use of all tables (i.e. diameter distribution tables) simultaneously.

Ek et al. (1997) developed and tested imputation models to estimate forest stand characteristics to predict regeneration conditions post harvest in Minnesota. Using data acquired from the Forest Service Inventory and Analysis Units database for Minnesota, tables were created that contained forest plot, stand origin, cover type, site index, stand age,
tree species and D. To test the models, Ek et al. (1997) measured variability using standard deviation statistics, graphed diameter and species distributions, and plotted trees per hectare and basal area between the models and over time. Ek et al. (1997) found that by imputing individual plot values they were able to produce more realistic estimates of the variation of tree species diversity than by imputing plot averages.

1.4 Assumptions

Missing At Random

The ability to use missing data mechanisms depends heavily on the reason for missingness (Rubin 1976). As explained by Schafer (1997), given an observed variable Y as \( Y_{\text{obs}} \) and a missing variable Y as \( Y_{\text{mis}} \), it can be said that \( Y = (Y_{\text{obs}}, Y_{\text{mis}}) \). Data are said to be Missing At Random if the probability that an observation is missing is not dependent on \( Y_{\text{mis}} \), although it may be dependent on \( Y_{\text{obs}} \) (Rubin 1976). That is, the missing data are assumed a random sample of the sample of values within a subclass of the observed data \( Y_{\text{obs}} \) (Rubin 1976). For example, consider a dataset, \( Y_1 = \) diameter at breast height (D) and \( Y_2 = \) total height (H) in which diameters are fully observed and H are missing for some cases. Missing At Random assumes that the probability a total height measurement is missing for an individual is not dependent on the H of the tree, however it can be dependent on the D of the tree (Little and Rubin 1987).

Missing Completely at Random

Little and Rubin (1987) defined data that are Missing Completely at Random as data
that are Missing At Random and observed at random or data in which the probability of response is independent of \( Y_{\text{obs}} \) and \( Y_{\text{mis}} \) (Schafer 1997). In other words, \( Y_{\text{mis}} \) is said to be a simple random subsample of \( Y \). In this case, given the previously mentioned forestry example, the probability that \( H \) is missing cannot be dependent on \( H \) or \( D \) (Schafer 1997). To test for Missing Completely At Random, the distribution of the complete data case is compared to the distribution of the incomplete data case (Schafer 1997). Missing Completely At Random is more restrictive than Missing At Random and generally harder to prove because the reason for missing observations is usually not known (Schafer 1997).

Non-Ignorable

In both cases, Missing At Random and Missing Completely At Random, the missing-data mechanism is said to be ignorable because it is possible to adjust for the missingness (Little and Rubin 1987). In some cases, even if it is possible to control for the remaining variables in the analysis, the probability of an observation in \( Y_{\text{mis}} \) being missing is still dependent on \( Y_{\text{mis}} \) (Schafer 1997). In this case, the missing-data mechanism is said to be non-ignorable and the imputation techniques needed require complicated analytical methods (Schafer 1997).

1.5 Missing Data Patterns

Monotone

Given variables \( Y_1 \), \( Y_2 \), and \( Y_3 \), a dataset is said to have a monotone missing pattern if the variables are ordered in some way. For example, Table 1 represents variables \( Y_1 \), \( Y_2 \),
and $Y_3$ with observations 1, 2, and 3. An X denotes missing data and a . denotes available data. When element $Y_3$ is missing, so are elements $Y_2$ and $Y_1$, representing a monotone missing data pattern (Rubin 1974).

Table 1: Monotone Missing Data Pattern

<table>
<thead>
<tr>
<th></th>
<th>$Y_1$</th>
<th>$Y_2$</th>
<th>$Y_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>X</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

Arbitrary

An arbitrary missing data pattern is one in which no missing data pattern is apparent. For example, in Table 2, given variables $Y_1$, $Y_2$, and $Y_3$, where an X denotes missing data and an . denotes available data, no pattern of missingness can be determined such that if $Y_2$ is missing, so is $Y_1$.

Table 2: Arbitrary Missing Data Pattern

<table>
<thead>
<tr>
<th></th>
<th>$Y_1$</th>
<th>$Y_2$</th>
<th>$Y_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.</td>
<td>X</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>.</td>
<td>X</td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>
1.6 Missing Data Mechanisms

Case Deletion

Case Deletion is the process of deleting observations with missing data and analyzing the remaining data as a complete dataset. In doing this, there is an assumption that the incomplete dataset is in-fact complete (Little and Rubin 1987). Little and Rubin (1987) refer to case deletion as complete case analysis or available case analysis, where the data being analyzed are the data available where none of the observations have missing data. This is the easiest imputation method because there are no extra steps to be taken aside from discarding the observations that contain missing data. It is also the default treatment in most statistical packages and therefore used more frequently than other missingness techniques (Paul et al. 2003). Proc Reg in SAS (SAS Institute Inc. 2006) follows this methodology; observations with missing data are automatically discarded and the remaining data are analyzed as a complete dataset. An obvious disadvantage to this approach is a reduction in sample size, i.e. a loss of data. It is possible to use this approach if the missing data being analyzed are Missing At Random and are a random subsample of the observed data (Little and Rubin 1987). Analysis is subject to bias if the missing-data mechanism is non-ignorable, which occurs when the probability of $Y_i$ being observed is in any way dependent on $Y_i$ (Little and Rubin 1987). According to Chantala and Suchindron (2003), if a random subsample is drawn from an original sample that is Missing Completely At Random, then Case Deletion will produce unbiased estimates. This is believed to be the case if the proportion of missing samples is small. Case deletion, in most instances, produces standard errors in a subsample that are larger than would be seen in a non-reduced sample (Chantala and Suchindron 2003).
If the sample data are not Missing Completely At Random, then this does not hold. When the sample data are Missing At Random, Chantala and Suchindron (2003) found that estimates would be biased. Another concern with case deletion in multivariate problems is the loss of power due to a large portion of data being discarded (Schafer 1999).

Mean Imputation

Mean imputation is a form of single imputation in which missing values are replaced with the estimated mean of the dataset. Mean imputation requires the imputer to calculate a mean and impute it into the missing values. Most of the data management packages available today are capable of performing this analysis procedure quickly (i.e. SAS, R, S+). Little and Rubin (1987) do not recommend this method of analysis because the resulting estimates are unreliable and often require ad hoc adjustments. Little and Rubin (1987) also state that mean imputation is not practical when studying the shape of the distribution of Y because imputing a mean often distorts the empirical distribution; because of this, histogram and plot shapes will be difficult to analyze correctly. Schafer (1997) stated that while mean imputation preserves the observed sample mean, it often biases the estimated variances and covariances toward zero. Chantala and Suchindron (2003) also found that standard deviations and standard errors are underestimated when using mean imputation.

Expectation-Maximization Algorithm

Single imputation using the Expectation-Maximization algorithm is the process of calculating and imputing a value for each missing variable based on best prediction models
(Dempster et al. 1977; Little and Rubin 1987). The Expectation-Maximization approach is used often in survey non-response (Rubin 1987). One advantage of the Expectation-Maximization algorithm is that once the imputation is performed, the dataset may be analyzed using standard complete-data methods (Rubin 1987). Another advantage is that the ease of the process allows the imputation to be performed only one time and, in most cases, by the data collector who is likely to have more knowledge of the basis for non-response (Dempster et al. 1977). Because of this, the data collector is likely to receive less biased results when imputing values than an external analyst using a public-use database (Rubin 1987).

Expectation-Maximization monopolizes on the fact that information is contained within missing data \((Y_{\text{mis}})\) that is relevant to estimating a parameter \((\theta)\) that is used to find possible values of \((Y_{\text{mis}})\) (Schafer 1997). The maximum likelihood estimates of the missing data are computed based on the observed data by determining a likelihood function that has been averaged over a predictive distribution for the missing values (SAS Institute Inc. 2006). Schafer (1997) explains the Expectation-Maximization computation as follows: obtain an initial parameter estimate \(\theta\) and fill in the missing data \((Y_{\text{mis}})\), re-estimate the parameter \(\theta\) based on the observed data \((Y_{\text{obs}})\) and the estimated missing data \((Y_{\text{mis}})\), repeat this process until the estimates converge. The calculation can be broken down into two main steps (Schafer 1997), the Expectation step and the Maximization step. In the Expectation step, \((Y_{\text{mis}})\) is unknown and the function of a parameter \((\theta)\) given \(t\), where \(t\) is the sequence of iterates 1,2,3…and represents a preliminary estimate of the unknown parameter, is calculated by taking an average \((Q)\) of the estimates (written \(Q(\theta|\theta^{(t)})\)) (Schafer 1997). By averaging
the complete-data log likelihood \( l(\theta|Y) \) over the prediction interval of missing data (\( Y_{\text{mis}} \)) given observed data (\( Y_{\text{obs}} \)) of parameter \( t \) (written \( P(Y_{\text{mis}}|Y_{\text{obs}},\theta^{(t)}) \)), parameter estimates of (\( Y_{\text{mis}} \)) can be obtained (Schafer 1997). Schafer (1997) calls \( P(Y_{\text{mis}}|Y_{\text{obs}},\theta^{(t)}) \) the predictive distribution of the missing data given \( \theta \), and notes that it plays a vital role in the EM algorithm because it accentuates the interdependence between \( Y_{\text{mis}} \) and \( \theta \). In the Maximization step, maximizing \( Q(\theta|\theta^{(t)}) \) allows you to find \( 1 + \) the maximum number of iterates of a parameter (\( \theta^{(t+1)} \)). Because single imputation imputes just one value for each missing variable it tends to underestimate the variance of estimates (Little and Rubin 1987). Single imputation may be adequate if the proportion of missing samples is small, say less than 5% (Schafer 1997). In addition, Schafer (1997) states that inferences from Expectation-Maximization are very similar to inferences from multiple imputation given a large number of complete datasets.

Multiple Imputation

Multiple imputation is similar to single imputation in that it imputes a set of likely values from a distribution for each missing variable (Rubin 1987). Multiple imputation, however, imputes several values, \( m \geq 2 \), for each missing datum (Rubin 1987). Each individual \( m \) is analyzed using standard complete-data procedures (Rubin 1987). Pooling \( m \) complete datasets and calculating the between imputation variances accounts for the uncertainty of the parameters about the proper value to impute (Rubin 1987). Pooling data from a number of imputations allows Multiple Imputation to produce more accurate
estimates of the missing data (Rubin 1987). The three stages of multiple imputation are summarized by Rubin (1987) as: 1) Imputation - create \( m \) datasets from \( m \) imputations of missing data drawn from a different distribution for each missing variable, 2) Analysis - analyze the \( m \) imputed datasets as complete datasets using standard statistical analysis, and 3) Data Pooling - combine \( m \) datasets into one dataset and analyze the results as one complete dataset. Figure 1 shows an example of multiple imputation where \( m = 3 \).

![Image of Multiple Imputation Flow Chart]

**Figure 1: Multiple Imputation Flow Chart**

The results are combined using the following calculations and annotation used by Rubin (1987) and Schafer (1997). Given \( \overline{C} = \) the multiple-imputation point estimate, \( C = \) the complete data and \( \hat{C} = \) the complete data point estimate using the \( t^{th} \) set of imputed data,
For $t = 1, 2, \ldots, m$, we can determine the multiple-imputation point estimate ($\bar{C}$) for $C$ by averaging the complete data point estimates with,

$$
\bar{C} = \frac{1}{m} \sum_{t=1}^{m} \hat{C}^{(t)}
$$

(1)

Given $U$ = the variance estimate associated with $\hat{C}$, let $\hat{C}^{(t)} = \hat{C}(Y_{obs}, Y_{mis}^{(t)})$ and $U^{(t)} = U(Y_{obs}, Y_{mis}^{(t)})$. Given $m$ imputations, we are able to calculate $m$ versions of $\hat{C}$ and $U$. The within-imputation variance is calculated as,

$$
\bar{U} = \frac{1}{m} \sum_{t=1}^{m} U^{(t)}
$$

(1)

The between-imputation variance is calculated as,

$$
B = \frac{1}{m-1} \sum_{t=1}^{m} (\hat{C}^{(t)} - \bar{C})^2
$$

(2)

The total variance estimate associated with $\bar{C}$ is

$$
T = \bar{U} + (1 + \frac{1}{m})B
$$

(3)

Relative efficiency is estimated in units of variance and is approximately a function of missing information (Rubin 1987, p. 114). Knowing the rate of missing information helps determine the number of imputations needed to reach a desired efficiency of estimates. As the rate of missing information increases, the number of imputations needed to maintain efficiency must increase as well (Table 3).

From a non-Bayesian perspective, the efficiency of an estimate is shown by Rubin (1987) with
where $\gamma$ is the rate of missing information and $m$ is the number of imputations (Table 3).

Table 3: Relative Efficiency for Various Levels of Imputation

<table>
<thead>
<tr>
<th>$m$</th>
<th>rate of missing information ($\gamma$)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20%</td>
<td>30%</td>
<td>40%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.938</td>
<td>0.909</td>
<td>0.882</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.952</td>
<td>0.930</td>
<td>0.909</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.962</td>
<td>0.943</td>
<td>0.926</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.968</td>
<td>0.952</td>
<td>0.938</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.972</td>
<td>0.959</td>
<td>0.946</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.976</td>
<td>0.964</td>
<td>0.952</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.978</td>
<td>0.968</td>
<td>0.957</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.980</td>
<td>0.971</td>
<td>0.962</td>
<td></td>
</tr>
</tbody>
</table>

A number of multiple imputation mechanisms have been developed. This study utilizes the Markov chain Monte Carlo (MCMC) method because the missing data pattern is arbitrary. This method assumes data are from a multivariate normal distribution and can only be used for analysis when data are Missing At Random or Missing Completely At Random (Rubin 1987). According to Gelman et al. (1995), a Markov chain Monte Carlo approach gives reliable results in the simplest manner. Monte Carlo simulation is distinguished from other simulation methods by being nondeterministic, typically by using random numbers (Gelman et al. 1995). A Markov chain is described as a random walk in the space $\theta$ that creates a joint posterior distribution or target distribution $p(\theta | y)$ by converging to a stationary distribution (Gilks et al. 1996, Berg 2004, Robert and Casella 2004). The Markov
chain is created by selecting a series of random variables in which the distribution of each element is dependent on the value of the previous variable (Schafer 2003).

Figure 2 illustrates five independent sequences of a Markov chain with each black dot representing the starting point of a Markov chain. The space within and surrounding the points represent all possible values of \( \theta \). The Markov chain begins at the center of each dot and continues to make iterations throughout the space of \( \theta \) forming a common stationary distribution \( p(\theta|y) \). The iteration is performed until the five Markov chains converge. Convergence occurs when the change in parameter estimates between iterations is less than \( p \), where \( p = \) a number between 0 and 1 (SAS Institute Inc. 2006). See Gelman et al. (1995) for further details.

![Figure 2: Independent Sequences of a Markov Chain Simulation before Convergence](image)

* Figure adapted from Gelman et al. (1995)

The Markov chain Monte Carlo method may be viewed as a Bayesian analysis technique because it pulls from the posterior distribution (Gelman et al. 1995). A number of frequentists, however, feel that Multiple Imputation performed in this manner is not strictly
Bayesian because the prior distribution used is a non-informative prior (Gelman et al. 1995). A prior probability is a description of what is known about a variable in the absence of some evidence (Schafer 2003). An non-informative prior articulates vague information about a variable and typically yields results similar to conventional statistical analysis (Schafer 2003). They tend to have minimal effect on the analysis (Schafer 2003). The Jeffreys prior (the default in SAS) is a non-informative prior distribution that gives no prior information about the mean and covariance estimates and therefore is not considered Bayesian by all (SAS Institute Inc. 2006). The posterior distributions of a Jeffreys prior are

$$\Sigma^{(t+1)} \mid Y \sim W^{-1} (n - 1, (n - 1)S)$$

(5)

where $\Sigma$ is a given covariance matrix, $t$ is the iteration, $Y$ is the data matrix, $W^{-1}$ is an inverted Wishart distribution, $n$ is the degrees of freedom, and $(n - 1)S$ is the corrected sums of squares and cross products matrix

$$\mu^{(t+1)} \mid (\Sigma^{(t+1)}, Y) \sim N(\frac{1}{n} \Sigma^{(t+1)} \bar{Y})$$

(6)

where $\mu$ is the posterior population mean vector and $\bar{Y}$ is the observed mean (SAS Institute Inc. 2006).

If the missing data pattern is monotone, either a regression method that assumes multivariate normality or a propensity score method may be used (Rubin 1987). A regression approach is one where a regression model is fit for each variable with missing values, with the previous variables as covariates (Rubin 1987). Based on the resulting model, a new regression model is then fit and used to impute the missing values for each variable (Rubin 1987). Another approach is known as the propensity score method. As
defined by Rosenbaum and Rubin (1983), a propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates. Applying this technique to multiple imputations, a propensity score is generated for each variable with missing values to indicate the probability of that observation being missing (Rosenbaum and Rubin 1983). The observations are then grouped based on these propensity scores, and an approximate Bayesian bootstrap imputation (Rubin 1987) is applied to each group (Lavori et al. 1995).

Rubin (1987) states that a major advantage to using one of the aforementioned Multiple Imputation techniques over single imputation is that the user can use complete data analysis on \( m \) imputed datasets and combine the results using simple formulas. Using this technique increases the efficiency of estimation because the estimates are randomly drawn in a manner that represents the distribution of the data (Rubin 1987). Rubin (1987) also states that by combining complete data inferences, it is possible to reflect additional variability due to missing values as long as the imputations represent repeated random draws under a model of non-response.

An obvious disadvantage to Multiple Imputation is that more work is required to produce and analyze the imputations. In addition, more space will be needed to store the datasets during and after imputation (Gelman et al. 1995). However, assuming the amount of missing data are relatively low, say less than 30%, \( m \) can be kept modest (between 3 and 5) and the amount of space needed, minimal.
Bayesian Multiple Imputation

According to Neal (1998), Bayesian inference is an approach to statistics in which all forms of uncertainty are expressed in terms of probability. Neal (1998) explains a Bayesian approach to a problem as formulating a model that is adequate to describe the data of interest, formulating a prior distribution over the unknown parameters of the model, applying Bayes's Rule to obtain a posterior distribution for these unknowns that takes into account both the prior and the data, and computing predictive distributions from this posterior distribution. Rubin (1987) applies this theory to Multiple Imputation and lists a number of stipulations Multiple Imputation analysis must meet to be considered Bayesian. Rubin (1987) states that a parametric model from the complete data must be specified, a prior distribution (informative or non-informative) must be applied to unknown model parameters, and m independent draws must be simulated from the conditional distribution of the missing data given the observed data by Bayes’s Theorem. Given stochastic events H (height) and D (diameter), and probability p, Bayes’s theorem is stated as

\[
p(H|D) = \frac{p(D|H)p(H)}{\int (p(D|H)p(H))}
\]

(7)

According to Rubin (1987), Bayes’s theorem is used to relate marginal distributions and conditional distributions. Schafer (1997) defines Bayesianly proper multiple imputations as independent realizations of the posterior predictive distribution of the missing data given a complete-data model and prior defined by

\[
P(Y_{mis}|Y_{obs}) = \int P(Y_{mis}|Y_{obs}, \theta)P(\theta|Y_{obs})d\theta
\]

(8)

The Inverted Wishart distribution is an informative prior often used for Bayesian
calculations and is the matrix variate generalization for the inverted gamma distribution (Gupta and Nagar 2000). An informative prior conveys explicit information about a variable. 

\[ A = X^T X \], the matrix of sums of squares and cross-products, is said to have a Wishart distribution where \( X \) is an \( n \times p \) data matrix whose rows are iid \( N(0, \Lambda) \) (Schafer 1997). The Wishart distribution is written as

\[ A \sim W(n, \Lambda) \]  \hspace{1cm} (9)

Where \( n \) is the degrees of freedom and \( \Lambda \) is the scale.

The inverted-Wishart is \( B = A^{-1} \) or

\[ B \sim W^{-1}(n, \Lambda) \]  \hspace{1cm} (10)

This forms a conjugate prior distribution that is the simplest way to perform Bayesian inference in the complete-data case (Schafer 1997). Rubin (1996) explains that the properness of a multiple imputation procedure depends on the complete-data estimates and variance-covariance matrices. Given population values \((X,Y)\) and the intended sample that are fixed, proper imputation deals with the fixed but missing values of the complete data statistic as if they were estimands (Rubin 1996). The values of the complete-data statistic and the variance-covariance matrix must be approximately unbiased (Rubin 1996). A disadvantage to Bayesian multiple imputation is that operational difficulties seem to arise because of the increased volume of data management and handling due to repeated imputation. Also, additional random noise is added to the data through the numerous imputations (Rubin 1996). Both of these disadvantages are minor compared to the relative increase in efficiency of data estimates. Bayesian proper Multiple Imputation also captures uncertainty due to missing data given parameters of the complete data (Schafer 1997).
2 Materials and Methods

2.1 Data

Data from two sources were used for this study: (1) Henderson Long-term Site Productivity Study (Figure 3), maintained by the North Carolina State University and Virginia Polytechnic Institute and State University Forest Nutrition Cooperative and (2) North Carolina State University’s Hill Demonstration Forest (Figure 3).

Henderson Study Site Data

The Henderson Long-Term Site Productivity Study was established in 1980 in Vance County, near Henderson, NC (North Carolina State University Forest Nutrition Coop 2006). The site is located on Cecil soils in the Piedmont of NC and is approximately 100 km north of Raleigh, NC. Average temperatures near Henderson range from 2.5°C in January to 23°C in July with approximately 143.3 cm of rain falling annually. In 1981, one year loblolly pine seedlings were hand planted to establish a second rotation study on abandoned agricultural fields. Four regeneration treatments were installed and implemented during the first five years after establishment using two levels of vegetation control (1) no further treatment and (2) mechanical and herbicide applications and two levels of site preparation (1) chop and burn techniques and (2) shear/pile and disk technique. Each of these treatments was replicated six times within the site. The objectives of the study were to assess the effects of harvesting, site preparation, and vegetation control. The measurements used for the present study are D and H measurements taken annually during the dormant season for 22 years. 29,386 trees were measured over the 22 year period (Table 4). The mean D of all trees measured was 5.25 in with the smallest diameter equal to 0.10 in and the largest diameter...
equal to 12.40 in (Table 4). The mean H of all trees measured was 36.98 ft and H ranged from 2.60 ft to 79.50 ft (Table 4). At age 5, tree diameter has a range of approximately 2 in and at age 22, tree diameter has a range of approximately 10 in. At age 5, tree height has a range of approximately 16 ft and at age 22, tree height has a range of approximately 52 ft. The imputed data set should be representative of these ranges.

The Henderson Long-Term Site Productivity Study data are presented by age in Table 5. Data from the first four years have been removed for analysis because diameters were not taken at breast height until age 5. Diameter and height distributions for the completed dataset are depicted in Figure 4 and 5, respectively. The empirical data are presented in the histogram for comparative purposes.
Figure 3: Study Site Locations
### Table 4: Combined Plot Data Summary

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diameter (in)</td>
<td>29,386</td>
<td>5.25</td>
<td>2.28</td>
<td>0.10</td>
<td>12.40</td>
</tr>
<tr>
<td>Height (ft)</td>
<td>29,386</td>
<td>36.98</td>
<td>17.01</td>
<td>2.60</td>
<td>79.50</td>
</tr>
</tbody>
</table>

### Table 5: Henderson Data Summary by Age – All Treatments

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>Diameter (in)</th>
<th>Height (ft)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Mean</td>
</tr>
<tr>
<td>5</td>
<td>1718</td>
<td>1.51</td>
</tr>
<tr>
<td>6</td>
<td>1742</td>
<td>2.29</td>
</tr>
<tr>
<td>7</td>
<td>1745</td>
<td>2.92</td>
</tr>
<tr>
<td>8</td>
<td>1742</td>
<td>3.46</td>
</tr>
<tr>
<td>9</td>
<td>1702</td>
<td>4.05</td>
</tr>
<tr>
<td>10</td>
<td>1690</td>
<td>4.54</td>
</tr>
<tr>
<td>11</td>
<td>1676</td>
<td>4.95</td>
</tr>
<tr>
<td>12</td>
<td>1660</td>
<td>5.33</td>
</tr>
<tr>
<td>13</td>
<td>1649</td>
<td>5.66</td>
</tr>
<tr>
<td>14</td>
<td>1636</td>
<td>5.94</td>
</tr>
<tr>
<td>15</td>
<td>1616</td>
<td>6.19</td>
</tr>
<tr>
<td>16</td>
<td>1613</td>
<td>6.42</td>
</tr>
<tr>
<td>17</td>
<td>1545</td>
<td>6.62</td>
</tr>
<tr>
<td>18</td>
<td>1531</td>
<td>6.84</td>
</tr>
<tr>
<td>19</td>
<td>1565</td>
<td>7.07</td>
</tr>
<tr>
<td>20</td>
<td>1548</td>
<td>7.25</td>
</tr>
<tr>
<td>21</td>
<td>1517</td>
<td>7.39</td>
</tr>
<tr>
<td>22</td>
<td>1491</td>
<td>7.57</td>
</tr>
</tbody>
</table>
Figure 4: Diameter Distribution for the Completed Dataset
Figure 5: Height Distribution for the Completed Dataset
Hill Demonstration Forest Data

The North Carolina State University Hill Demonstration Forest is a research and education forest located 21 km north of Durham in Durham County, NC (Figure 3). The average temperature ranges from 2.9°C in January to 25°C in July and the area receives an average of 115.3 cm. of rain annually. In December 2002, an ice storm caused damage to the forested stands on this 992-hectare forest. Many trees suffered broken tops, bent stems, and snapped boles. There was interest in estimating the amount of timber loss due to the ice storm by faculty at North Carolina State University. An inventory of a subset of the forest was performed using 178 BAF 20 point samples. Characteristics that were recorded include species, D, H, merchantable height, and a damage assessment value. The damage condition code ranged from 0-3 where 0 = dead due to ice storm damage, 1 = live with heavy damage, 2 = live with some damage, and 3 = live with no damage. Where broken tops were found, height to the broken top was recorded.

This research used the individual stem measurements from the point samples installed after the ice storm. All observations with missing D were deleted. Only stems with a D greater than or equal to 4.5 inches were retained for estimating the volume loss. This step was taken because trees with D less than 4.5 are considered non-merchantable in standard forestry practices. If the merchantable height (MerHt) is equal to H, then there is a broken top because the merchantable limits were to an approximate 4 in. top diameter outside bark. The total stem height was set to missing when MerHt was equal to H. Condition codes 0 and 1 predominantly indicated broken tops. Trees with condition code 2 could either have broken branches, broken branches and top, or broken top. For this reason, it was assumed
that all code 2 trees had broken tops and should be imputed. When the condition code was
equal to 0, 1, or 2, H was set to missing. The remaining dataset consisted of 951 observed
trees in which 333 of the trees had missing H, approximately 35.02% of the dataset. This
was a higher percentage of missing data than anticipated. Diameter and height distributions
for the Hill Forest dataset are depicted in Figures 6 and 7, respectively. Notice the diameter-
classes begin at 4.5 inches, reflecting the merchantable timber. A high proportion of the
trees fall within the 7.5-13.5 diameter-class. Most of the trees ranged from 52-84 ft tall with
the average tree height being 73 ft (Figure 7). Summary statistics for D and H of the non-
missing data are shown in Table 6.
Figure 6: Diameter Distribution of Hill Forest Data
Figure 7: Height Distribution of Hill Forest Data

Table 6: Hill Forest Plot B Non-missing Data Summary

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diameter (in)</td>
<td>618</td>
<td>12.27</td>
<td>5.22</td>
<td>4.60</td>
<td>31.80</td>
</tr>
<tr>
<td>Height (ft)</td>
<td>618</td>
<td>73.40</td>
<td>15.82</td>
<td>50.16</td>
<td>132.55</td>
</tr>
</tbody>
</table>
2.2 Methods

Henderson Dataset

To test the accuracy of the methods using a forestry dataset, a complete dataset containing diameter and height information was used. Heights were removed randomly for 20% of the data to simulate a dataset with missing data so that the true distribution could be compared with results from various imputation techniques. A random number generator was used to obtain the missing values so that approximately 20% of the data was classified as missing. The dataset with missing values was used to compare the following imputation methods: case deletion, mean imputation, Expectation Maximization algorithm, multiple imputation, and Bayesian multiple imputation. The mean, standard deviation, minimum height, and maximum height of the missing data do not deviate from the original dataset values (Table 7).

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Total Height (ft)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>All Data</td>
<td>29,386</td>
<td>36.98</td>
</tr>
<tr>
<td>20% Missing</td>
<td>5,818</td>
<td>36.97</td>
</tr>
<tr>
<td>80% Observed</td>
<td>23,568</td>
<td>36.96</td>
</tr>
</tbody>
</table>

The model used to relate height as a function of diameter for all imputation techniques was

\[ H = \beta_0 + \beta_1 D + \xi. \]  

(11)
where $\xi$ is an error term, $\beta_0$ and $\beta_1$ are parameters to be estimated. A scatter plot of the original dataset and the fitted regression line where $\hat{\beta}_0 = 1.584$ and $\hat{\beta}_1 = 6.7428$ was computed for comparison of imputation techniques (Figure 8).

Because the thrust of this research was to compare imputation procedures, a simple model was used to characterize the relationship between $H$ and $D$. A combined variable volume equation derived by Burkhart (1977) was used to estimate the total cubic foot volume of each stem. The equation for plantation-grown loblolly pine, outside bark in total cubic-foot volume given by Burkhart (1977) is:

$$\hat{V} = 0.34864 + 0.000232(D^2 H)$$

(12)

This volume model will give a representation based on stem volume that will aid in determining which imputation technique does the best job at representing the missing data.
Figure 8: Regression of the Observed Dataset
Case Deletion

A regression model was run in SAS with the missing data using Equation (12). The Proc Reg procedure in SAS will delete any observations with missing values automatically. The output dataset contains a 95% confidence interval for each observation that was determined by model (12). Each observation in the output dataset was assigned a code of 1 or 0 to determine which observations had parameter estimates that were contained within the 95% confidence interval. If the actual height was contained in the 95% confidence interval, it was assigned a 1 and if the actual height was not contained in the confidence interval, it was assigned a 0. The frequency of coverage was then determined by calculating the percentage of 0’s to 1’s. In the case deletion method, this step is only useful for verifying that the 80% non-missing data estimates are consistent with the remaining imputation techniques.

Mean Imputation

The analysis for mean imputation was carried out in SAS using Proc Reg. A mean estimate of H was calculated from the observed data where H was observed (80% dataset). The 20% of data that was missing H was replaced with the estimated mean. A regression was run using Equation (12) to determine 95% confidence intervals for the mean imputation technique. (See Appendix 1 for the code used in this research.) The frequency of coverage was then calculated using the same method previously described in the case deletion step. Equation (13) was used to estimate the total volume loss of the 20% missing data imputed with a mean variable.
Expectation Maximization algorithm

A single imputation utilizing the Expectation Maximization algorithm was calculated using the Proc MI procedure (SAS Institute Inc. 2006). A Jeffreys prior was used to generate the missing data (see section 1.6). The Expectation Maximization algorithm was used to compute the Maximum Likelihood Estimate for \((\mu, \Sigma)\) the means and covariance matrix, then Markov chain Monte Carlo was used to produce \(m = 1\) imputation (SAS Institute Inc. 2006). An initial 200 iterations were run before taking the first imputation to attain a stationary distribution and then 100 iterations were run between each imputed observation to reduce any dependency on the next imputation of an individual observation. A regression equation was run using Equation (12) to determine the % confidence interval for each observation with a missing data value. From the resulting confidence interval, it was possible to check the coverage frequencies as previously described. Volume estimates were calculated using Equation (13).

Multiple Imputation

The dataset used for the multiple imputation procedure had \(D\) measurements available for 100% of the data and \(H\) measurements available for 80% of the data. Because the missing data pattern is arbitrary, the Markov chain Monte Carlo method of multiple imputation was used. Given approximately 20% missing data, Table 3 finds \(m=5\) to be sufficient for multiple imputation with a greater than 95% estimate efficiency. Proc MI was run in SAS using a Jeffreys prior, also used in the Expectation-Maximization analysis, to generate the missing data. Proc MI initiates the MI procedure, which is similar to the
Expectation-Maximization analysis performed previously except $m=5$ datasets are imputed instead of $m=1$. Initially, 200 burn-in iterations were run before taking the first imputation to attain a stationary distribution and then 100 iterations were run between each imputation to reduce any dependency on the next imputation (SAS Institute Inc. 2006). Convergence was checked in SAS using the time series and autocorrelation function plots. Schafer (1997) describes the Worst Linear Function of parameters as a scalar function of parameters $\mu$ and $\Sigma$ whose function values converge the slowest among parameters in the Markov chain Monte Carlo process. Given parameters $\theta = (\mu, \Sigma)$, a Worst Linear Function of $\theta$ has the highest rate of missing information which is resultant from the posterior mode in the Expectation Maximization algorithm (Schafer 1997). Based on the convergence of the Worst Linear Function, the convergence of the remaining parameters is assumed (Schafer 1997) (see section 3.1).

Once convergence was verified, a regression equation was run using Equation (12) to determine the % confidence interval for each datum. Multiple imputation in SAS is usually complete at this stage and the between, within, and total imputation variance is calculated from the equations presented in section 1.6. These estimates are between and within each of the $m$ imputations. To determine the between and within imputation variance from each datum, an additional step was introduced. To do this, the variance from each datum was pulled out of the confidence intervals provided by SAS using the following equation

$$\hat{\sigma} = \left( \frac{\hat{Y} - ucl}{1.96} \right)$$

where $\hat{\sigma}$ is the variance estimate, $\hat{Y}$ is the estimate, and $ucl$ is the upper confidence limit.
The average complete data point estimate, within variance estimate and total variance
estimate were calculated using Equations (1), (2) and (4), respectively.

The between variance estimate \( B \) was calculated by

\[
B = \hat{\sigma}^2 (\hat{Q} - \bar{Q})^2
\]

where \( \hat{Q} \) is the complete data point estimate for \( Q \) and \( \bar{Q} \) is the multiple imputation point
estimate for \( Q \).

The combined standard error is

\[
\sigma = \sqrt{T}
\]

where \( T \) is the total variance.

The 95% upper confidence limit is calculated by

\[
ucl = \bar{Q} + 1.96\sigma
\]

and the 95% lower confidence limit is calculated by

\[
lcl = \bar{Q} - 1.96\sigma
\]

The percentage of observed heights that fell within the 95% confidence interval was
calculated. The combined volume estimate \( \bar{V} \) is the average of the \( m \) complete-data
estimates:

\[
\bar{V} = \frac{1}{m} \sum_{t=1}^{m} V'\]

where \( V \) is the individual volume estimate and \( t \) is the imputation.

The volume from the 20% missing data was calculated to estimate the total volume loss on
the stand using Equation (13) for the Hill Forest Data.
Bayesian Multiple Imputation

The data set used for Bayesian multiple imputation was the same data set used for multiple imputation where H measurements were available for 80% of the data. SAS is unable to perform the Bayesian multiple imputation analysis using the appropriate prior distribution; the analysis was performed in the R statistical package using the Norm library (R Development Core Team 2006). R-Norm is a package that allows the analysis of multivariate normal datasets with missing data (Schafer 2002). As discussed in section 1.6, the prior distribution used for the Multiple Imputation technique must be an informative prior distribution to be considered Bayesian. The R-Norm package performs similar steps for Multiple Imputation as SAS except it allows for the use of a normal-inverted Wishart distribution. This distribution is discussed in detail in section 1.6. When using a normal-inverted Wishart prior distribution that does not have the parameters of a Jeffreys prior, the Expectation Maximization algorithm step uses the posterior of the distribution curve to obtain the Maximum Likelihood Estimate. The same additional step used to calculate prediction intervals was taken for the Bayesian Multiple Imputation analysis in R that was taken in the Multiple Imputation analysis performed in SAS. Prediction intervals were calculated by using the variance from the confidence intervals calculated in R by using Equation (14) (Appendix 1). From the new prediction intervals, it was possible to calculate the percentage of observed heights that fell within the calculated predicted intervals. Volume estimates were also derived from each imputed $m$ with Equation (19) and combined for an overall volume estimate on the missing data.
3 Results

3.1 Henderson Study Site

Convergence Verification

Figure 9 is a time series plot that is a scatter plot of parameter estimates against the iteration number that examines the convergence behavior of the estimation algorithm (SAS Institute Inc. 2006). The importance of convergence is discussed in section 1.6. It is apparent that convergence occurs using the time series plot Worst Linear Function for the Henderson dataset. Figure 10 is an autocorrelation plot of the Worst Linear Function that examines relationships between successive parameter estimates, \( \xi \) (SAS Institute Inc. 2006). In SAS (SAS Institute Inc. 2006), this relationship is determined for a stationary series, \( \xi_i \), \( i \geq 1 \) where \( k \) is the autocorrelation function at lag, by the formula

\[
p_k = \frac{\text{Cov}(\xi_i, \xi_i + k)}{\text{Var}(\xi_i)}
\]

(19)

And the sample \( k^{th} \) order autocorrelation is computed as

\[
r_k = \frac{\sum_{i=1}^{n-k} (\xi_i - \overline{\xi})(\xi_{i+k} - \overline{\xi})}{\sum_{i=1}^{n} (\xi_i - \overline{\xi})^2}
\]

(20)

The dashed lines in Figure 10 show approximate 95% confidence intervals for autocorrelation of the Henderson dataset. It is apparent that no strong positive or negative autocorrelation exists when using the Worst Linear Function. Mean convergence was also checked with a time series plot and autocorrelation function plot and is seen in Figures 11 and 12, respectively.
Figure 9: Time Series Plot using the Worst Linear Function
Figure 10: Autocorrelation Plot using the Worst Linear Function
Figure 11: Time Series Plot using the Mean
Figure 12: Autocorrelation Plot using the Mean
Initial data analysis allowed for a comparison of the estimated mean, standard deviation (SD), and minimum and maximum imputed estimates of the missing data for all techniques (Table 8). Case deletion deletes the 20% missing data and it is assumed that the remaining 80% of the data are the complete dataset. Performing the analysis in this way caused the loss of almost 6,000 observations. The loss of information prevented the comparison of the case deletion procedure with the remaining procedures in the initial analysis. Mean imputation had estimates that are centered around the mean and allowed for the sample mean to be preserved. There is a bias that strongly underestimates the standard deviation as 0 and causes bias in estimated variances and covariances. The Expectation Maximization algorithm overestimated the mean and underestimated the standard deviation compared to the actual data. The range of estimates in the Expectation Maximization algorithm is the largest of all imputation methods, estimating heights 18 ft lower than the lowest observed height and 15 ft higher than the highest observed height in the actual data. Multiple imputation and Bayesian multiple imputation did not differ by a large amount and had estimates that were very similar to the actual data.
Table 8: Missing Data Summary by Imputation Method

<table>
<thead>
<tr>
<th></th>
<th>Total Height (ft)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>Actual Data</td>
<td>36.97</td>
<td>17.05</td>
</tr>
<tr>
<td>Case Deletion</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Mean Imputation</td>
<td>36.99</td>
<td>0</td>
</tr>
<tr>
<td>Expectation Maximization algorithm</td>
<td>37.06</td>
<td>16.96</td>
</tr>
<tr>
<td>Multiple Imputation</td>
<td>36.90</td>
<td>15.38</td>
</tr>
<tr>
<td>Bayesian Multiple Imputation</td>
<td>36.87</td>
<td>15.38</td>
</tr>
</tbody>
</table>

Figures 13 – 19 show the observed data and imputed data as well as the upper and lower bound confidence intervals for case deletion, mean imputation, Expectation Maximization algorithm, multiple imputation (imputations # 3 and #5), and Bayesian multiple imputation (imputations #2 and #5), respectively. Recall, Figure 8 shows the original dataset and can be reviewed for comparison purposes with the following figures.

Figure 13 has nearly 6,000 observations removed and because the data are missing at random the altered data should appear similar to the original data and the shape of the plotted regression should not change greatly. Figure 14 shows the observed and imputed data to help better visualize the effects of imputing a mean. It can be seen from the figure that estimated variances are biased and the resulting imputed estimates of the mean will be affected.

Figures 15, 16 and 17 graph individual imputations for Expectation Maximization algorithm, Bayesian Multiple Imputation where $m = 2$, and Bayesian Multiple Imputation where $m = 5$, respectively.
respectively. Both techniques appear to have larger prediction intervals than those observed in case deletion, mean imputation, and multiple imputation. It is apparent that imputed data were taken from the posterior of the distribution because more individual imputations (+) are observed outside of the confidence interval limits. Figures 18 and 19 graph individual imputations for Multiple Imputation where $m = 3$ and $m = 5$. It can be seen that Multiple Imputation also has larger prediction intervals than case deletion and mean imputation, but does not seem to pull imputations from the posterior as much as Expectation Maximization algorithm or Bayesian Multiple Imputation. Also, note that the Henderson Site dataset included trees with a D less than 4.5 inches. There were instances where a negative value was imputed causing the Expectation Maximization algorithm, Multiple Imputation, and Bayesian Multiple Imputation techniques to provide results with negative imputations, however it is necessary to include these data for the comparison of techniques. In forestry standard practices trees with D less than 4.5 inches are considered non-merchantable timber and it can be seen that all imputed negative numbers came from the distribution that falls below the 2.5 in D-class. Deleting the negative numbers will affect the distribution of H and not allow us to compare the techniques properly. For this reason, all stems have been left in the data set for the initial analysis on the Henderson Study Site dataset.
Figure 13: Observed Tree Heights and DBH Measurements for the Data with the Missing Data Deleted for the Case Deletion Technique
Figure 14: Observed Tree Heights and DBH Measurements for the Data with Imputed Mean Heights for the Mean Imputation Technique
Figure 15: Imputed Heights using the Expectation Maximization algorithm Technique
Figure 16: Imputed Heights for Imputation 2 of the Bayesian Multiple Imputation Technique
Figure 17: Imputed Heights for Imputation 5 of the Bayesian Multiple Imputation Technique
Figure 18: Imputed Heights for Imputation 3 of the Multiple Imputation Technique
Figure 19: Imputed Heights for Imputation 5 of the Multiple Imputation Technique
The ability of each imputation mechanism to create confidence intervals that will contain the observed height is presented in Table 9. Notice that the actual data produces confidence intervals that contain the observed heights approximately 95% of the time. Case deletion performs well because the dataset is considered Missing Completely At Random. The disadvantage to case deletion is visible when volume loss estimates are derived in the following section. Mean imputation has wider confidence intervals and is, therefore, able to capture a higher frequency of observed data contained within those confidence intervals. Imputed means cause confidence intervals to be inflated and produce inaccurate data summaries. The Expectation Maximization algorithm produces wider confidence intervals and has a slightly higher percentage of observed data that are contained within the confidence intervals produced. Multiple Imputation performs well, creating confidence intervals that contain the observed data at the same frequency as the actual data. Bayesian multiple imputation also creates confidence intervals that contain the imputed missing data at the same frequency of the actual data. Multiple imputation places a slightly lower percentage of imputed heights within the confidence interval and Bayesian multiple imputation places a slightly lower percentage of observed heights within the confidence interval. Comparing the combined 80% observed data and 20% missing data, the Expectation Maximization algorithm produces confidence intervals that most closely replicate the actual data set.
Table 9: Confidence Interval Coverage Frequency by Imputation Method

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100% all data</td>
</tr>
<tr>
<td>Actual Data</td>
<td>94.93</td>
</tr>
<tr>
<td>Case Deletion</td>
<td>94.90</td>
</tr>
<tr>
<td>Mean Imputation</td>
<td>97.96</td>
</tr>
<tr>
<td>Expectation</td>
<td>94.94</td>
</tr>
<tr>
<td>Maximization</td>
<td></td>
</tr>
<tr>
<td>algorithm</td>
<td></td>
</tr>
<tr>
<td>Multiple Imputation</td>
<td>94.92</td>
</tr>
<tr>
<td>Bayesian Multiple</td>
<td>94.90</td>
</tr>
<tr>
<td>Imputation</td>
<td></td>
</tr>
</tbody>
</table>

**Volume Estimates**

It is not possible to calculate a volume loss estimate for the 20% missing data in case deletion because data with missing observations were discarded. It is possible to look at the total volume estimate for case deletion and see that deleting the missing data causes us to underestimate the total volume of trees by more than 23,000 cubic feet due to the exclusion of missing data. When comparing volume estimates of the remaining techniques, mean imputation performed the poorest of the four techniques. The estimated volume using mean imputation was underestimated by nearly 5,000 cubic feet (Table 10). The mean was underestimated and the standard deviation was biased downward for mean imputation. The Expectation Maximization algorithm performed well, with volume estimates only off by approximately 32 cubic feet. The mean was slightly overestimated and the standard deviation was overestimated for the Expectation Maximization algorithm technique.
Multiple imputation did not estimate volume as accurately as the Expectation Maximization algorithm or Bayes Multiple Imputation, but still only overestimated the volume by 55 cubic feet. The mean determined by the Multiple Imputation technique was overestimated by 0.011 cubic feet and the standard deviation was overestimated by 0.042 cubic feet. Bayesian multiple imputation performed the best of all five imputation methods. The Bayesian multiple imputation technique maintained the variability in the dataset without biasing estimates. The volume loss estimate using Bayesian Multiple Imputation was overestimated by 2 cubic feet. The mean estimate was overestimated by 0.001 cubic feet and the standard deviation was overestimated by 0.03 cubic feet when using Bayesian Multiple Imputation as an imputation technique. A comparison table of total volume (cubic feet), mean and standard deviation by imputation method is given in Table 10.
<table>
<thead>
<tr>
<th></th>
<th>Actual</th>
<th>Case Deletion</th>
<th>Mean Imputation</th>
<th>Expectation Maximization algorithm</th>
<th>Multiple Imputation</th>
<th>Bayesian Multiple Imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sum</strong></td>
<td>23,143</td>
<td>---</td>
<td>18,312</td>
<td>23,175</td>
<td>23,198</td>
<td>23,145</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>3.978</td>
<td>---</td>
<td>3.148</td>
<td>3.983</td>
<td>3.987</td>
<td>3.978</td>
</tr>
<tr>
<td><strong>St.Dev</strong></td>
<td>3.653</td>
<td>---</td>
<td>2.079</td>
<td>3.729</td>
<td>3.695</td>
<td>3.684</td>
</tr>
<tr>
<td><strong>80%</strong></td>
<td>94,026</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>3.990</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>St.Dev</strong></td>
<td>3.617</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>100%</strong></td>
<td>117,169</td>
<td>94,026</td>
<td>112,339</td>
<td>117,201</td>
<td>117,225</td>
<td>117,355</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>3.987</td>
<td>3.990</td>
<td>3.823</td>
<td>3.988</td>
<td>3.989</td>
<td>3.994</td>
</tr>
</tbody>
</table>

Table 11 reflects the volume loss estimate of the 20% data, the percent of total volume that was estimated to be lost, and the total volume estimate. The percent of total volume lost for the mean imputation technique was underestimated by 3.45 percent not accounting for 4,831 cubic feet in the volume loss estimation. The percent of total volume lost was overestimated by 0.02 and 0.03 percent by the Expectation Maximization algorithm and Multiple Imputation respectively. The Expectation Maximization algorithm outperforms the other imputation techniques in estimating the percent of total volume loss (cubic feet), but it overestimated the total volume and volume loss by 32 cubic feet each. Multiple Imputation and Bayesian Multiple Imputation also had estimates of percent total volume loss similar to the actual data and these techniques performed as well as the Expectation Maximization algorithm. Multiple imputation overestimated the total volume and volume loss...
Bayesian multiple imputation overestimated the volume loss and total volume by only 6 cubic feet. Bayesian Multiple Imputation techniques are applied in the following section to the North Carolina State University Hill Demonstration Forest dataset to determine the volume loss of timber due to an ice storm.

### Table 11: Volume Loss Estimates by Technique

<table>
<thead>
<tr>
<th>Imputation Technique</th>
<th>Volume Loss (cubic feet)</th>
<th>% of Total Vol</th>
<th>Total Volume (cubic feet)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>23,144</td>
<td>19.75%</td>
<td>117,170</td>
</tr>
<tr>
<td>Case Deletion</td>
<td>---</td>
<td>---</td>
<td>94,026</td>
</tr>
<tr>
<td>Mean Imputation</td>
<td>18,313</td>
<td>16.30%</td>
<td>112,339</td>
</tr>
<tr>
<td>Expectation Maximization</td>
<td>23,176</td>
<td>19.77%</td>
<td>117,202</td>
</tr>
<tr>
<td>Multiple Imputation</td>
<td>23,198</td>
<td>19.79%</td>
<td>117,225</td>
</tr>
<tr>
<td>Bayes Multiple Imputation</td>
<td>23,150</td>
<td>19.72%</td>
<td>117,378</td>
</tr>
</tbody>
</table>

#### 3.2 Hill Demonstration Forest Ice Damage Assessment

Bayesian multiple imputation was used to predict the total volume loss due to an ice storm on Block B of North Carolina State University Hill Demonstration Forest. Bayesian Multiple Imputation was used in this step because the method produced the closest estimated volume loss for the 20% missing data in the simulated procedure for the Henderson data presented earlier. A total of 951 D and 618 H measurements were taken, resulting in a 35% data loss for H. The mean D of the observed data on the block was 12.3 inches and the mean
H of the observed data on the block was 73.4 feet. Imputing the 35% data that was lost resulted in 333 estimates with a mean H of 74.2 feet and a standard deviation of 17.0 feet. The minimum H imputed was 50.5 feet and the maximum H imputed was 178.0 feet (Table 12).

Table 12: Hill Forest Imputed Data Statistics

<table>
<thead>
<tr>
<th></th>
<th>Diameter (in)</th>
<th>Height (ft)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>35% data with missing heights</td>
<td>65% observed data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35% imputed data</td>
</tr>
<tr>
<td>n</td>
<td>333</td>
<td>618</td>
</tr>
<tr>
<td>Mean</td>
<td>12.54</td>
<td>12.27</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>5.60</td>
<td>5.22</td>
</tr>
<tr>
<td>Min</td>
<td>4.7</td>
<td>4.60</td>
</tr>
<tr>
<td>Max</td>
<td>46.8</td>
<td>31.80</td>
</tr>
</tbody>
</table>

As discussed in section 2.2, volume estimates were calculated using the Bayesian Multiple Imputation technique and Formula (19). The estimated total volume still standing on Block B is 22,382 cubic feet. The individual tree volume has a mean of 36.2 cubic feet and a standard deviation of 42.0 cubic feet. The minimum and maximum individual tree volume estimate of standing timber on Block B is 2.8 cubic feet and 311.3 cubic feet, respectively. The estimated volume loss from Block B on the North Carolina State University Hill Demonstration Forest is 13,537.0 cubic feet. The mean individual tree volume loss on the block is 40.7 cubic feet and the standard deviation is 71.2 cubic feet. A
minimum of 2.9 cubic feet and a maximum of 904.7 cubic feet of individual tree volume loss was estimated. Table 13 represents the sample size, mean individual tree volume, standard deviation, minimum estimate, maximum estimate, and the sum of all the data to get a volume estimate for the available and missing data on the Hill Demonstration Forest.

<table>
<thead>
<tr>
<th></th>
<th>Volume (cubic feet)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>35% missing data</td>
</tr>
<tr>
<td>N</td>
<td>333</td>
</tr>
<tr>
<td>Mean</td>
<td>40.65</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>71.17</td>
</tr>
<tr>
<td>Min</td>
<td>2.93</td>
</tr>
<tr>
<td>Max</td>
<td>904.73</td>
</tr>
<tr>
<td>Sum</td>
<td>13,536.99</td>
</tr>
</tbody>
</table>
4 Discussion and Conclusions

Imputations of missing data should condition on observed variables to avoid distorting data. Multiple imputation outperforms case deletion, mean imputation, and Expectation Maximization algorithm when comparing confidence interval construction and coverage. Findings were similar to those of Rubin and Schenker (1986) in that the best imputation techniques are those that adjust for uncertainty and are computed a number of times with results combined. Bayesian multiple imputation outperformed the case deletion and mean imputation techniques. Bayesian Multiple Imputation performed comparatively to the Expectation Maximization algorithm and multiple imputation. Paul et al. (2003) performed a similar study where simulated datasets were used and the predicted data and observed data were compared. Paul et al. (2003) found that Bayesian techniques did not outperform conditional mean imputation, contradicting the findings of this study. Case deletion performed well in calculating confidence intervals, however, is not feasible for predicting volume estimates on broken stems because the missing data are discarded. Case deletion is sufficient for constructing confidence intervals and calculating mean estimates when data are Missing Completely At Random but resulted in a large amount of data loss. Mean imputation overestimated the standard deviation and produced biased results creating wider confidence intervals. Mean imputation ignored posterior data estimates and therefore underestimated volume loss estimates more than the Expectation Maximization algorithm, Multiple Imputation or Bayesian Multiple Imputation techniques. The Expectation Maximization algorithm overestimated the mean and underestimated the standard deviation. The Expectation Maximization algorithm produced the widest range of estimates and
captured the highest percentage of observed heights in the estimated confidence intervals but was not representative of the actual data because of the wide range of estimates. The minimum and maximum height imputed were 14 and 15 feet larger than the minimum and maximum height in the actual data, respectively. Volume estimates produced by Expectation Maximization algorithm, although slightly underestimated, were closer than the Multiple Imputation volume estimates. Multiple imputation performed well when deriving prediction intervals and mean estimates, producing estimates similar to the actual data. Volume loss estimates using Multiple Imputation were overestimated slightly more than the Expectation Maximization algorithm volume loss estimates. Bayesian multiple imputation developed confidence intervals and mean estimates that were similar to the original data. Volume loss was slightly overestimated, however, closer to the actual volume loss than the other methods had predicted. Although some of the differences were slight, Bayesian multiple imputation outperformed the rest of the imputation techniques.

In conclusion, the Expectation Maximization algorithm, Multiple Imputation, and Bayesian Multiple Imputation all performed well. The Expectation Maximization algorithm is the most time efficient technique to implement in SAS because little code writing is necessary to perform the analysis. Multiple Imputation can also be performed easily in SAS, but may require a higher understanding of code writing to perform the analysis properly. Bayesian Multiple Imputation requires using the program R. Without a previous understanding of the R program, learning and writing R code for use of the Bayesian Multiple Imputation technique will likely be timely and use up personnel resources. For more robust data estimates, Bayesian Multiple Imputation is recommended given that time
and resources are available to analyze the data. The findings suggest that given limited time
and resources, Expectation Maximization algorithm performed in SAS is a quick and
efficient means of estimating unknown tree height given diameter measurements. The results
are consistent with the findings of Van Deusen (1997) who concluded that Multiple
Imputation worked best in forestry practices and allowed for reduced monitoring effort with
little loss in precision. It is possible to obtain accurate volume estimates with diameter
measurements and 20% of the dataset missing height. This methodology will provide
landowners with an accurate means of estimating timber loss due to damage from ice storm
or hurricane. The Expectation Maximization algorithm, multiple imputation, and Bayesian
multiple imputation may also allow fewer measurements of height in the field. A subsample
of height may be taken in the field, cutting down on cost and time, and the non-observed
heights may be imputed later.
LITERATURE CITED


APPENDIX
SAS Code

Data Creation

data test;
set henderson;
  rannum = ranuni(9172005);
run;
proc sort data = henderson;
  by rannum;
run;
data missing;
set henderson;
  tht=height;
  if rannum <= 0.2 then tht = .;
run;

Case Deletion

data missing;
set henderson;
  tht=height;
  if rannum <= 0.2 then tht = .;
  V_cd=0.34864+0.00232*((dbh**2)*tht);
run;
title ' ';  
symbol1 c=black;
symbol2 c=red;
symbol3 c=blue;
symbol4 c=blue;
proc reg data=missing;
  model tht=dbh/cli;
  plot tht*dbh / pred nostat overlay symbol = '-';
  label tht = 'Height, ft'
  dbh = 'DBH, in';
  output out=casedel_bounds p=pred ucl=up_cl lcl=lo_cl;
  Title ' '; 
  run;
  quit;
data casecover;
set casedel_bounds;
if height < up_cl and height > lo_cl then cover = 1;
else cover = 0;
run;
proc freq data=casecover;
  tables cover;
  title 'Case Deletion Frequency';
  run;
data casdelcover80;
set casecover;
    if rannum <= 0.2 then cover = .;
    if cover = . then delete;
run;
proc freq data=casdelcover80;
tables cover;
    title '80% Case Deletion Cover Frequency';
run;
data casdelcover20;
set casecover;
    if rannum > 0.2 then cover = .;
    if cover = . then delete;
run;
proc freq data=casdelcover20;
tables cover;
    title '20% Case Deletion Cover Frequency';
run;
proc means n nmiss min max mean std sum data=missing;
    var V_cd;
    title '100% CD Volume';
run;
proc means n nmiss min max mean std sum data=casdelcover20;
    VAR V_cd;
    output out = CDV20;
    TITLE '20% CD Volume';
run;
proc means n nmiss min max mean std sum data=casdelcover80;
    VAR V_cd;
    output out = CDV80;
    TITLE '80% CD Volume';
run;

Mean Imputation

data nonmiss;
set missing;
    if tht = . then tht = 36.98775;
    V_mean_imp=0.34864+0.00232*((dbh**2)*tht);
run;
title ' ';  
    symbol1 c=black;    symbol2 c=red;    symbol3 c=blue;    symbol4 c=blue;    proc reg data=nonmiss;
        model tht=dbh/cli;
        plot tht*dbh / pred nostat overlay symbol = '.';
        label tht='Height, ft'
dbh='DBH, in';
output out=mean_bounds p=pred ucl=up_cl lcl=lo_cl;
    Title ' ';
    run;
quit;

data meancover;
set mean_bounds;
if height < up_cl and height > lo_cl then cover = 1;
else cover = 0;
run;
proc freq data=meancover;
    tables cover;
    title 'Mean Cover Frequency'
    run;

data meancover80;
set meancover;
    if rannum <= 0.2 then cover = .;
    if cover = . then delete;
run;
proc freq data=meancover80;
    tables cover;
    title '80% Mean Cover Frequency'
    run;

data meancover20;
set meancover;
    if rannum > 0.2 then cover = .;
    if cover = . then delete;
run;
proc freq data=meancover20;
    tables cover;
    title '20% Mean Cover Frequency'
    run;
proc means n nmiss min max mean std sum data=meancover;
    var V_mean_imp;
    Title '100% Mean Volume'
    run;
proc means n nmiss min max mean std sum data = meancover20;
VAR V_mean_imp;
output out = MeanV20;
TITLE '20% Mean Volume'
run;
Expectation Maximization algorithm

data singimp;
set outem;
  V_em=0.34864+0.00232*((dbh**2)*tht);
  run;
  title '';
symbol1 c=black;
symbol2 c=red;
symbol3 c=blue;
symbol4 c=blue;
proc reg data=singimp outest=outsi covout noprint;
  model tht=dbh/cli;
  plot tht*dbh / pred nostat overlay symbol = '-';
  label tht='Height, ft' dbh='DBH, in';
  output out=si_bounds p=pred ucl=up_cl lcl=lo_cl;
  Title '';
  run;
quit;
data sicover;
set si_bounds;
if height < up_cl and height > lo_cl then cover = 1;
else cover = 0;
run;
proc freq data=sicover;
  tables cover;
  title 'Expectation Maximization algorithm Frequency';
  run;
data sicover80;
set sicover;
  if rannum <= 0.2 then cover = .;
  if cover = . then delete;
  run;
proc freq data=sicover80;
  tables cover;
  title '80% Expectation Maximization algorithm Frequency';
  run;
data sicover20;
set sicover;
  if rannum > 0.2 then cover = .;
  if cover = . then delete;
  run;
proc freq data=sicover20;
  tables cover;
  title '20% Expectation Maximization algorithm Frequency';
  run;
proc means n nmiss min max mean std sum data=sicover;
  var V_em;
  Title '100% EM Volume';
  run;
proc means n nmiss min max mean std sum data = sicover20;
   VAR V_em;
   output out = EMV20;
   TITLE '20% EM Volume';
run;
proc means n nmiss min max mean std sum data = sicover80;
   VAR V_em;
   output out = EMV80;
   TITLE '80% EM Volume';
run;
title ',';
proc plot data=sicover20;
   plot tht*v_em;
run;

Multiple Imputation

data mimp;
   set outmi;
   V_mi=0.34864+0.00232*((dbh**2)*tht);
run;
title ',';
symbol1 c=black;
symbol2 c=red;
symbol3 c=blue;
symbol4 c=blue;
proc reg data=mimp;
   model tht=dbh;
   plot tht*dbh / pred nostat overlay symbol = '.';
   label tht='Height, ft'
   dbh='DBH, in';
   output out=MI_bounds p=pred ucl=up_cl lcl=lo_cl;
Title ',';
by _imputation_;
run;
quit;
DATA mi_reg_var;
   SET mi_bounds;
   imp_pred_var = ((pred-up_cl)/1.96)**2;
RUN;

data out1;
   set outmi;
   if _imputation_ = 1 then output out1;
run;
data out1 out2 out3 out4 out5;
   set mi_reg_var;
   keep _imputation_ plot row tree age pred imp_pred_var misflag height v_mi;
   if _imputation_ = 1 then output out1;
else if _imputation_ = 2 then output out2;
else if _imputation_ = 3 then output out3;
    else if _imputation_ = 4 then output out4;
    else if _imputation_ = 5 then output out5;
run;
proc sql;
create table combined1 as
    select a._imputation_, a.plot, a.row, a.tree, a.age, a.height, a.misflag,
        a.pred as pred1, b.pred as pred2, c.pred as pred3, d.pred as pred4, e.pred as pred5,
        a.imp_pred_var as var1, b.imp_pred_var as var2, c.imp_pred_var as var3,
        d.imp_pred_var as var4, e.imp_pred_var as var5,
        a.V_mi as vol1, b.V_mi as vol2, c.V_mi as vol3, d.V_mi as vol4, e.V_mi as vol5
    from out1 as a, out2 as b, out3 as c, out4 as d, out5 as e
    where a.plot = b.plot = c.plot = d.plot = e.plot
    and a.row = b.row = c.row = d.row = e.row
    and a.tree = b.tree = c.tree = d.tree = e.tree
    and a.age = b.age = c.age = d.age = e.age;
quit;
data combined2;
set combined1;
    vol_mi = (vol1 + vol2 + vol3 + vol4 + vol5)/5;
    pred = (pred1 + pred2 + pred3 + pred4 + pred5)/5;
    var_wi = (var1 + var2 + var3 + var4 + var5)/5;
    var_bn = std( of pred1-pred5)**2;
    var_combined = var_wi + (1 + 1/5)*var_bn;
    stderr_combined = sqrt(var_combined);
    ucl = pred + 1.96*stderr_combined;
    lcl = pred - 1.96*stderr_combined;
run;
data MIcover20;
set combined2;
    if height < ucl and height > lcl then cover = 1;
    else cover = 0;
    where misflag = 1;
run;
proc freq data=micover20;
    tables cover;
    title 'MI 20% Cover Frequency';
run;
data MIcover80;
set combined2;
    if height < ucl and height > lcl then cover = 1;
    else cover = 0;
    where misflag = 0;
run;
proc freq data=micover80;
    tables cover;
    title 'MI 80% Cover Frequency';
run;
data MIcover;
set combined2;
  if height < ucl and height > lcl then cover = 1;
  else cover = 0;
run;
proc freq data=micover;
  tables cover;
  title 'MI Cover Frequency';
run;
proc means n nmiss min max mean std sum data=micover;
  var vol_mi;
  title '100% MI Volume';
run;
proc means n nmiss min max mean std sum data = micover20;
  var vol_mi;
  output out = MIV20;
  TITLE '20% MI Volume';
run;
proc means n nmiss min max mean std sum data = micover80;
  var vol_mi;
  output out = MIV80;
  TITLE '80% MI Volume';
run;
title '';
proc plot data = micover20;
  plot pred*vol_mi;
run;

R Code

Bayesian Multiple Imputation

library(norm)

XYdata <- read.table("D:\Rockhill.csv",header=TRUE,sep=",")
names(XYdata)

#pull the values needed for imputation
X1 <- XYdata$dbh
Y1 <- XYdata$tht
Z1 <- XYdata$height # added for coverage
rannum <- XYdata$rannum
miss1 <- XYdata$misflag
Volume <- 0.34864 + 0.00232*((X1^2)*Z1)
#combine them together
XY <- cbind(X1,Y1)
n1 <- length(X1)
one1 <- rep(1,n1)

#initialize the multiple imputation scheme
s <- prelim.norm(XY)
params <- list(1,rep(0,ncol(XY)),.5*diag(rep(1,ncol(XY))))
rngseed(917253)
start <- ninvwish(s,params)
thetahat <- em.norm(s)
theta <- da.norm(s,thetahat,steps=20,showits=TRUE) #You can put your prior here
getparam.norm(s,theta)

# create a place to store results
predval <- matrix(0,n1,5)
varval <- matrix(0,n1,5)
volbmi <- matrix(0,n1,5)  ### added for volume estimates

# start imputing
for(i in 1:5){
  newXY <- imp.norm(s,theta,XY)
  newXY2 <- data.frame(x = newXY[,1])
  newlm <- lm(newXY[,2]~newXY[,1])
  predlm1 <- predict(newlm, newXY2, interval="prediction")
  fit1 <- predlm1[,1]
  upr1 <- predlm1[,3]
  prederr <- ((upr1 - fit1)/1.96)^2
  predval[,i] <- fit1
  varval[,i] <- prederr
  volavg <- 0.34864 + 0.00232*((X1^2)*predval[,i])  ### added for vol. est.
  volbmi[,i] <- volavg  ### added for volume estimate
}

# create a place to store results
predvalavg <- matrix(0,n1,1)
varvalavg <- matrix(0,n1,1)
bnvar <- matrix(0,n1,1)
wivar <- matrix(0,n1,1)
volbmiavg <- matrix(0,n1,1)  ### added for volume estimates

# combine the imputations together
for(i in 1:n1){
  predvalavg[i] <- mean(predval[,i])  # predicted valued
  bnvar[i] <- var(predval[,i])  # between variation
  wivar[i] <- mean(varval[,i])  # with in variation
  varvalavg[i] <- wivar[i] + (1+1/5)*bnvar[i]
  volbmiavg[i] <- mean(volbmi[,i])  ### mean volume
}

ucl1 <- predvalavg + 1.96*sqrt(varvalavg)
lcl1 <- predvalavg - 1.96*sqrt(varvalavg)

aboveucl <- ifelse(Z1>lcl1,1,0)  # ***** changed to Z1*****
belowucl <- ifelse(Z1<ucl1,1,0)
miss <- ifelse(rannum<=0.2,1,0)
nonmiss <- ifelse(rannum>0.2,1,0)
coverage20 <- abovecl*belowucl*miss
sum(coverage20)/sum(miss)

coverage80 <- abovecl*belowucl*nonmiss
sum(coverage80)/sum(nonmiss)

coverageall <- abovecl*belowucl
sum(coverageall)/n1

vol20 <- volbmiavg[rannum<=0.2,]
vol80 <- XY[rannum>0.2,]
sum(vol20)
mean(vol20)
sd(vol20)
sum(vol80)
mean(vol80)
sd(vol80)
sum(volbmiavg)
mean(volbmiavg)