Spatial Allocation, Imputation, and Sampling Methods for Timber Product Output Data

John P. Brown

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Forestry

Richard G. Oderwald, Chair
Stephen J. Prisley
Philip J. Radtke
Robert L. "Bob" Smith
Janice K. Wiedenbeck

September 16, 2009
Blacksburg, VA

Keywords: multiple imputation, relative efficiency, spatial allocation, nonlinear repeated measures, timber product output data

Copyright 2009, John P. Brown

Spatial Allocation, Imputation, and Sampling Methods for Timber Product Output Data

John P. Brown

ABSTRACT

Data from the 2001 and 2003 timber product output (TPO) studies for Georgia were explored to determine new methods for handling missing data and finding suitable sampling estimators.

Mean roundwood volume receipts per mill for the year 2003 were calculated using the methods developed by Rubin (1987). Mean receipts per mill ranged from 4.4 to 14.2 million ft^3. The mean value of 9.3 million ft^3 did not statistically differ from the NONMISS, SINGLE1, and SINGLE2 references means (p=.68, .75, and .76 respectively).

Fourteen estimators were investigated to investigate sampling approaches, with estimators being of several means types (simple random sample, ratio, stratified sample, and combined ratio) as well as employing two methods for stratification (Dalenius-Hodges (DH) square root of the Frequency method and a cluster analysis method. Relative efficiency (RE) improved when the number of groups increased and when employing a ratio estimator, particularly a combined ratio. Neither the DH method nor the cluster analysis method performed better than the other.

Six bound sizes (1, 5, 10, 15, 20, and 25 percent) were considered for deriving samples sizes for the total volume of roundwood. The minimum achievable bound size was found to be 10 percent of the total receipts volume for the DH-method using a two group stratification. This was true for both the stratified and combined ratio estimators. In addition, for the stratified and combined ratio estimators, only the DH method stratifications were able to reach a 10 percent bound on the total (6 of the 12 stratified estimators). The remaining six stratified estimators were able to achieve a 20 percent bound of the total.

Finally, nonlinear repeated measures models were developed to spatially allocate mill receipts to surrounding counties in the event of obtaining only a mill's total receipt volume. A Gompertz model with a power spatial covariance was found to be the best performing when using road distances from the mills to either county center type (geographic or forest mass). These models utilized the cumulative frequency of mill receipts as the response variable, with cumulative frequencies based on distance from the mill to the county.

DEDICATION

To my wife and confidant Anna for her loyal support and for my children.

ACKNOWLEDGEMENTS

I would like to first wholeheartedly thank all of my project leaders in the US Forest Service that enabled me to begin and complete my studies, particularly Drs. Bruce Hansen and John Baumgras who both encouraged me to apply to Virginia Tech, as well as Dr. Jan Wiedenbeck who served on my committee and Dr. Beth Adams who was enduring of the additional time I needed to get this project finished.

Many thanks go to my advisor Dr. Rich Oderwald for being my mentor and patiently helping me along the way. I appreciate his guidance in moving my dissertation drafts forward and helping to transform them into a much improved final product.

My thanks as well go to my other committee members. Dr. Steve Prisley provided thoughtful comments throughout, and I additionally benefitted from his excellent GIS instruction. Dr. Phil Radtke provided unthought-of lines of inquiry and I particularly enjoyed his portion of the Q&A during my defense, even though it had to be delivered over the phone. Dr. Bob Smith's comments as well helped to properly shape my dissertation.

Last, I would like to thank two members of the Southern Research Station FIA unit for getting me involved in TPO work and for providing me with the data used in this work. They are Tony Johnson and Carolynn Steppleton. Tony was involved in early discussions on how to improve TPO studies and suggested the data used herein. Carolynn was extremely helpful, gracious, and patient in helping transfer and explain the data to me. Of all the people involved in completing this work, she had the least vested interest and I very much appreciated her help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION AND OVERVIEW OF RESEARCH

1.1 JUSTIFICATION AND OBJECTIVES

Periodic assessments of timber removals are an important component of the United States Department of Agriculture (USDA) Forest Service's national forest inventory and monitoring program. The Organic Administrative Act of 1897 (30 Stat. 11, 34), which is the enabling legislation for the USDA Forest Service, included provisions for the inventory and monitoring of the Forest Reserves. Later, the McSweeney-McNary Forest Research Act of 1928 (P.L. 70-466) directed the Secretary of Agriculture to: "…make and keep current a comprehensive inventory and analysis of the present and prospective conditions of and requirements for the renewable resources of the forest and rangelands of the United States…". These two pieces of legislation formed the basis for the nation's Forest Survey program, which later evolved into the USDA Forest Service's forest inventory and analysis (FIA) program. Periodic changes and additions to public law have continued the mandate for a national forest inventory. Most notable are the Forest and Rangeland Renewable Resources Research Act of 1978 (P.L. 95-307), which replaced the McSweeney-McNary Act, and the Agriculture Research, Extension, and Education Reform Act of 1998 (16 USC 1642(e)), which outlines the current annual 20 percent measurement plan of FIA.

The timber product output (TPO) survey is one method utilized by the FIA program to assess timber removals. A complete canvass of all primary wood-using mills is conducted on a state by state basis based upon each mill within the state. The primary measure of interest is roundwood receipts, which is the cubic foot volume of roundwood harvested in state plus roundwood

imported from other states. The information collected provides important information on the amount, county of harvest, and species composition of roundwood harvested within each state.

These TPO studies are utilized by many clients, including private industry; city, county, state, and federal governments; trade organizations; and environmental groups. Accurate and timely information concerning roundwood removals is important to these organizations.

TPO research within the USDA Forest Service's Northern Research Station (NRS) is currently facing several challenges in regard to complete TPO censuses. Flatter budgets, fewer TPO personnel across both federal and state governments, inconsistent past mill response, and varied survey dates have all made complete censuses problematic. Workable estimation procedures are needed to find solutions in order to provide a sufficiently accurate assessment of industrial roundwood consumption.

This research has several components that explore methodology for addressing incomplete censuses in TPO studies, with "incomplete" defined as a census having missing data. The objectives of this research are fourfold. First, multiple imputation will be utilized to handle missing cases. Second, methods are developed to stratify mill populations to reduce sampling variability in regards to state means and totals. Third, the stratification techniques are then followed by sample size calculations to test the feasibility of the stratification techniques. Finally, spatial models are developed that provide modeling techniques to tie mill receipts back to the land base at the county level. Ultimately, it is hoped that these new methodologies can move TPO studies to a sounder statistical basis and away from problematic incomplete censuses.

1.2 PROPOSED METHODS

There are two fundamental issues arising after general consideration has been given to this data. The first issue is how to handle the missing data, whether it is missing partially or completely. The second issue is whether methods can be developed to move away from a canvass and towards sampling.

After thoughtful consideration of various advanced methods for handling missing data, the methods developed by Rubin (1987) were chosen for further exploration using this data. The values of interest are means and totals, which are very straightforward basic statistics, and therefore should not prove intractable to application of the methods. This analysis will be detailed in Chapter Three.

Another facet of this missing data is that TPO studies are undertaken with the intent to determine counties of origin for roundwood removals. Some information may be more readily obtainable than other types. Specifically, researchers may be able to get the totals receipts for a mill but not the species groups or counties of origin. This situation generates the question of how receipts might be spatially allocated back to counties through a modeling approach. This question is addressed in Chapter Six.

Canvasses can fail to be complete because the resources needed to do so become prohibitive, i.e. cost, time, etc. Consideration of how it can be done with less then arise and sampling is a routine response. Sampling can additionally help to alleviate the missing data problem, as there can be

fewer missing responses and thus potentially sufficient resources to obtain this smaller number of missing observations.

When moving to sampling, care must be given to which techniques might produce estimates that have reasonable bounds on the uncertainty of those estimates. This data exhibited high variability in relation to the mean, suggestion that approaches other than a simple mean that help to reduce variance might be needed to get better bounds on estimates. In Chapter Four, two methods are explored: stratified means and ratio estimators.

It was quickly determined that the class sizes of the readily available classification by mill type were too small to adequately sample from the finite mill population. Cluster analysis and the Dahlenius-Hodges classification method were used to develop an alternative classification variable. This analysis is also included in Chapter Four.

Reducing the variance of the estimators (means and totals) may not create a sampling scheme that will provide acceptable bounds to the estimators. It was known that classes of the original mill types were small (e.g. composite panel, plywood and veneer mills), so perhaps any new classes might also be small. Therefore, examination of sample size allocation strategies is needed to determine whether any desired bounds are achievable. This is covered in Chapter Five.

1.3 DATA PREPARATION

1.3.1 Canvass Data

TPO data was collected for the years 2001 and 2003 from 100 percent canvasses of Georgia primary wood using plants. Data was collected in 2002 for the 2001 assessment and the in 2004 for the 2003 assessment by the Southern Research Station's (SRS) FIA unit. Questionnaires were mailed with additional information provided by telephone or through personal contact as needed for completion (Johnson and Wells 2001, Johnson and Wells 2003). A blank 2003 questionnaire is included in the Appendix.

The goal of the TPO studies performed by USFS is to produce a report detailing the production figures for each state. This requires that the import data for a particular state survey be returned as production to all other states during their TPO surveys. These returned imports are then tallied as exports and the total production can then be reported for that particular state.

Data was entered and checked by the SRS's FIA unit. This information was then provided as two Microsoft Access® databases. One database was strictly for pulpmills and the other was for the remaining mill types: sawmills, veneer, composite panel, plywood, post, pole and others. There were four basic types of tables within each database. A single table included mill location, contact information, number of employees and previous TPO survey status (yes, no). A second table gave receipts totals for each mill by product classes (sawlogs, veneer, OSB, poles, posts, other). The third table provided county locations of roundwood removals. The fourth type of table provided the percentage of a species group removed from each county for each mill. This

fourth type had three tables per year due to multiple species groups and the need for many columns in the table layout.

1.3.2 Canvass Data Preparation

Tables were imported to SAS as database tables for the first two types of tables--mill descriptive data and total receipts. However, the other two table types--county of origin and percentage by species group--did not have primary keys and were first exported to Microsoft Excel® for preparation. Each mill had one or more rows associated with it, depending on whether the mill drew roundwood from more than twelve counties. This is an artifact of the questionnaire (see Appendix), as only twelve columns were provided to label the counties. A thirteenth (or twenty-fifth, etc.) county created a new line in the data table. This required very careful checking of the alignment of the rows within a table, after which a primary key was then added and data exported to SAS.

Data resolution was at the species group level. For each species, the percentage of that species from each county is recorded for each mill. Every species had twelve columns of data with values ranging from 0 to 100, e.g. white pine had columns WPINE1-WPINE12. There was at least one row per mill and sometimes more than one row. Macros were written in SAS to rearrange the data into a mill, county, species, and percentage columnar format. This was facilitated by having created a primary key above.

Once the data was loaded into a SAS library, it was necessary to convert all totals to a standard unit of measure, in this case cubic feet Receipts totals, which were given in a variety of units,

were converted into cubic feet totals using conversion factors from Johnson (2001), (Table 1.1) . Conversion factors not provided in that report were obtained from C. Steppleton in the SRS's FIA unit (personal communication April 17, 2006). These conversion factors were then utilized to calculate the total cubic feet removed of each species at the mill and county level.

Two summations were then performed to create receipts totals needed for later analysis. The first summation was at the mill level, with all species receipts totaled for each mill. The second summation was for each county's contribution to each mill, where all species receipts were totaled for each county.

Table 1.1: Conversion factors for TPO data (Johnson 2001, C. Steppleton, personal communication, April 17, 2006)

Class	Type	From	To	Conversion Factor
Saw Logs	Hardwood	Board Feet (International 1/4-inch rule)	Cubic Foot	0.16807
	Softwood	Board Feet (International 1/4-inch rule)	Cubic Foot	0.18349
Veneer Logs	Hardwood	Board Feet (International 1/4-inch rule)	Cubic Foot	0.16260
	Softwood	Board Feet (International 1/4-inch rule)	Cubic Foot	0.17094
Pulpwood	Hardwood	Cords	Cubic Foot	75.0
	Softwood	Cords	Cubic Foot	72.6
Posts	Hardwood	Pieces	Cubic Foot	0.636
	Softwood	Pieces	Cubic Foot	0.635
Poles	All	Pieces	Cubic Foot	1.714

1.3.3 Geographic Data

Mill geographic locations in the state of Georgia were collected with GPS units and provided by the Southern Station FIA unit. Three mills were missing coordinates and these were georeferenced with checks made to Google® map satellite images. There were ten states for which TPO mill receipts were recorded: Alabama, Florida, Georgia, Kentucky, North Carolina, Ohio, South Carolina, Tennessee, Virginia and West Virginia (Figure 1.1). Previously calculated mill receipts were added as a table to ARCMap® and then joined to the mill layer.

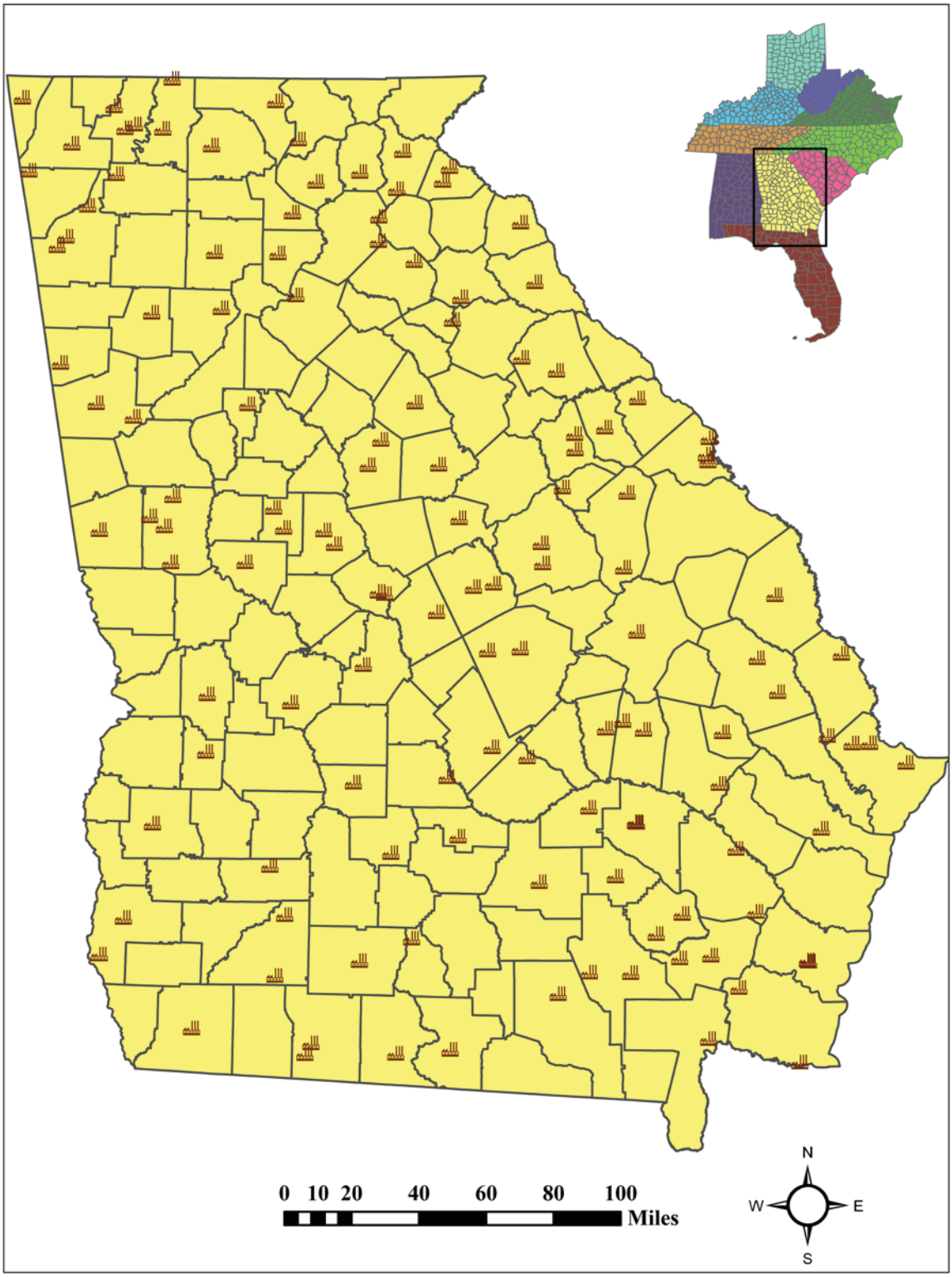

Figure 1.1: Spatial distribution of Georgia primary mills. Inset shows ten state region contributing industrial roundwood. Map created by author.

County polygons and roads layers were obtained from the U. S. Census Bureau's TIGER database. County shapefiles were merged into one ten state county layer. Roads were also merged into a single layer as well.

Forest cover data was derived from the U.S. Forest Service's Enhanced FIA Program. Forested plots from the inventories prior or up to 2001 were first selected from the FIA database. A plot is considered forested by FIA if it has ten percent or greater stocking. The latitude and longitude of each plot along with its representative area was retained in a table after querying the database. The table was imported to ARCMap, with all observations added as points to each of the counties in the ten state study area using the Add XY menu command in ARCMap. These plot locations are subject to privacy law and plot location accuracy is affected by FIA's "fuzz and swap" rules. Plot locations are "fuzzed" by altering the coordinates by up to one mile of distance. Up to 20 percent of private plots have coordinates "swapped" within a county, which masks ownership but still maintains county level data accuracy (USDA 2004).

Given that all mill data is from Georgia, a customized azimuthal equidistant projection was developed centered in Georgia. The mean center tool in ARCMap was utilized to find this center and the coordinates were used as the central meridian and latitude of origin in the customized projection.

1.4 CANVASS DATA DESCRIPTION

There were a total of 170 mills canvassed for the 2001 survey. Of these 170 mills, 85 were non-respondents for a response rate of only 50.0 percent. In 2003, 187 mills were canvassed. Six mills were non-respondents giving a response rate of 96.8 percent. The distribution of non-respondent mills by mill type is shown in Figure 1.2. For the year 2001, the 50.0 percent missing rate is roughly distributed across all mill types (Figure 1.2A). Veneer and Other mills are missing more than 50 percent whereas pulp mills were all accounted for. In 2003, non-respondent mills do not predominate for any class (Figure 1.2B) and no class is missing more than two mills. (Hereafter, non-respondent mill receipts will be referred to interchangeably as duplicates, as previous survey receipts are substituted for the missing mill receipts.)

The distribution of mill receipts by volume class for both 2001 and 2003 is illustrated in Figure 2.3. Years are divided into two categories: Response=Yes and Response=No. The receipts distribution for all mills is left skewed in both years with a high count of mills having receipts of five million cubic feet or less. The non-respondents (which are duplicated from the year 1999 data) show a slightly different distribution of receipts values for the year 2001 (Figure 1.3A). Over 50 mills in the zero to one million cubic feet category are duplicates from previous inventories, while the remaining duplicate mills appear to follow a similar distribution as the actual mills. Non-respondent mills for 2003 are single occurrences scattered throughout the distribution of the actual mills (Figure 1.3B).

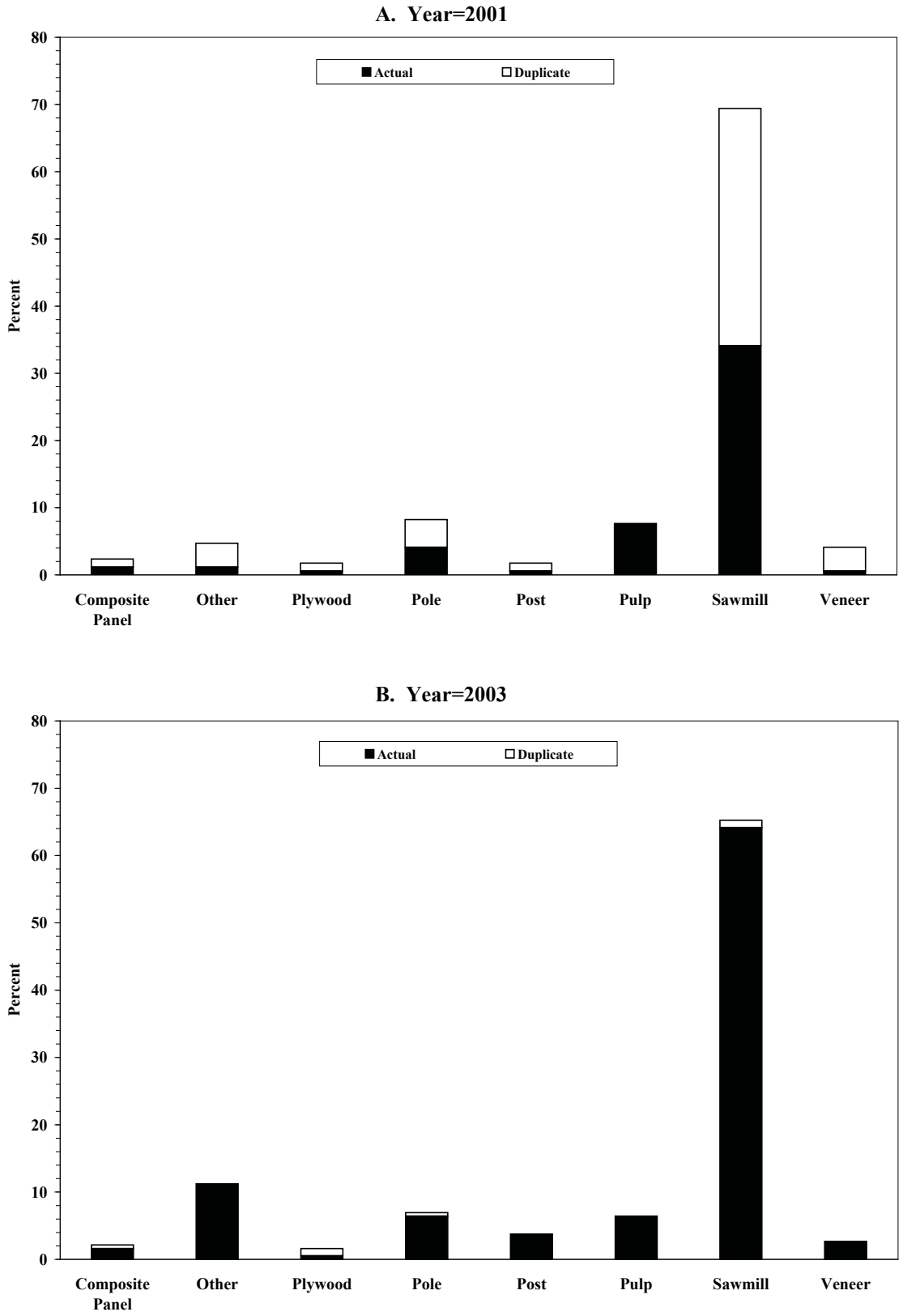

Figure 1.2: Mill types and missing data information for the 2001 and 2003 canvasses.

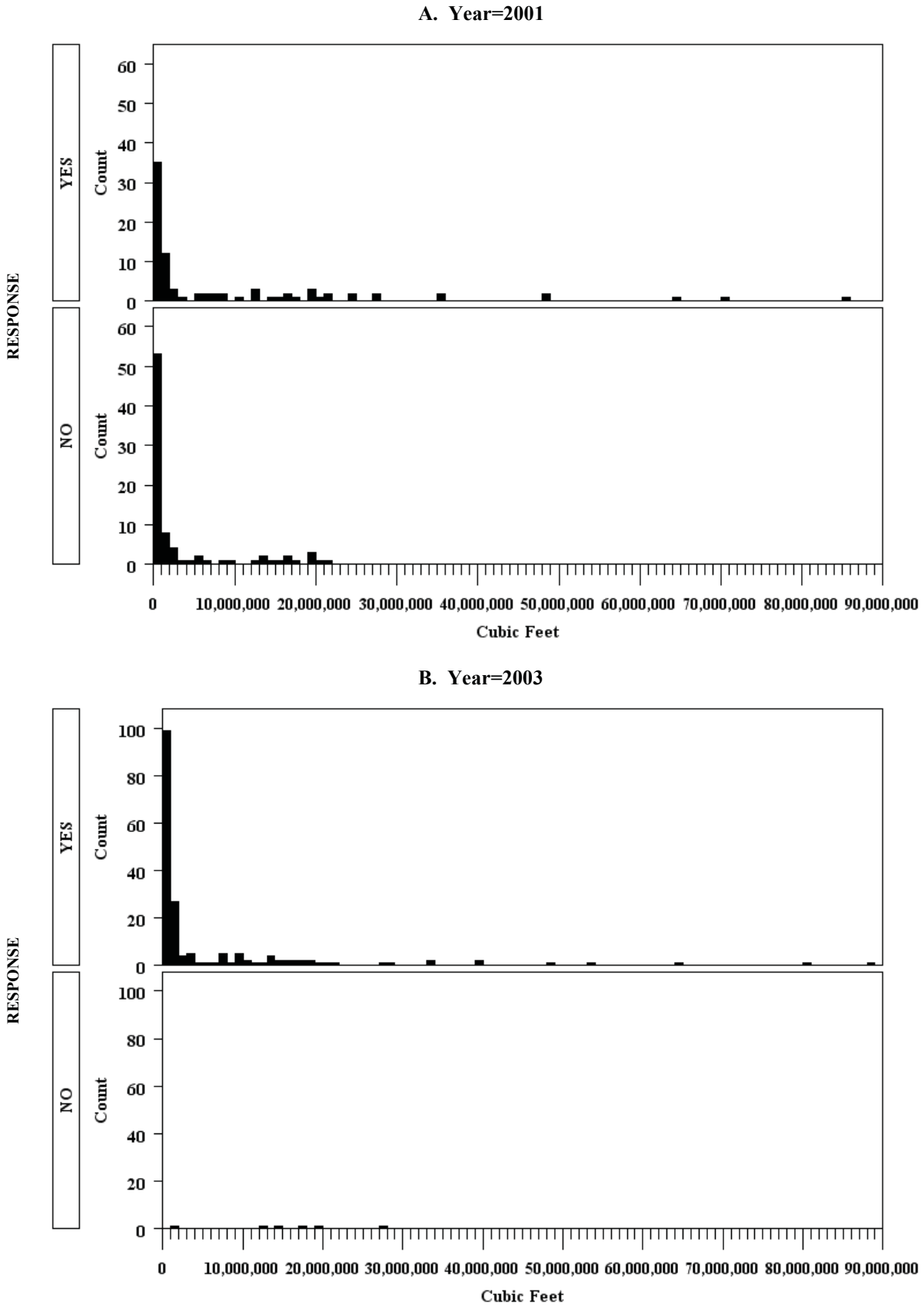

Figure 1.3: Distribution of mill receipts by volume class for the years 2001 and 2003.

Mean receipts per mill for the year 2001 were 9.9 million cubic feet with a standard deviation of 16.5 million cubic feet for the actual mills. With the included duplicate values, the mean per mill drops to 6.7 million with a standard deviation of 12.8 million cubic feet. In 2003, the mean receipts per mill were 5.9 million cubic feet with a standard deviation of 13.1 million cubic feet. This mean changes slightly to 6.2 million cubic feet with a standard deviation of 13.0 million cubic feet when the duplicates are added in.

CHAPTER 2: LITERATURE REVIEW

2.1 INTRODUCTION

While TPO data are available for any a number of states, data from Georgia was selected due to a near complete census in 2003. In Georgia, forest surveys span the years 1934-2003, (Table 2.1). Industrial roundwood production peaked in 1995 at 1.31 billion ft^3, and has declined slightly to 1.15 billion ft^3 in 2003. In TPO terminology production is the roundwood volume harvested and used within state (retained) plus the volume of roundwood exported to other U.S. states. Receipts data, which are separately included in Georgia's TPO reports after 1971, show steady increases up until 1997, reaching a high of 1.36 billion ft^3(Table 2.2). . Then there is a slow decline through 2003, when a value of 1.17 billion ft^3 was recorded Receipts are the retained volume plus the roundwood volume imported from other U.S. states which is consumed by the state's primary mills.

Table 2.1: Historical Georgia roundwood production 1937-2003.

Year of study	*Production (thousand ft^3)*	*Source*
1937	247,090	Spillers 1943
1952	621,888	McCormick and Cruikshank 1954
1961	548,840	Larson and Spada 1963
1971	801,806	Welch and Bellamy 1976
1974	851,669	Welch and Bellamy 1976
1986	1,202,584	Tansey and Steppleton 1991
1989	1,113,594	Tansey and Steppleton 1991
1992	1,229,921	Johnson 1994
1995	1,311,507	Johnson, Jenkins, and Wells 1997
1997	1,280,581	Johnson and Wells 1999
1999	1,244,541	Johnson and Wells 2002
2001	1,122,661	Johnson and Wells 2004
2003	1,152,854	Johnson and Wells 2005

Table 2.2: Georgia roundwood receipts 1971-2003.

Year of study	Receipts *(thousand ft^3)*	Source
1971	828,853	Welch and Bellamy 1976
1974	829,206	Welch and Bellamy 1976
1989	1,135,335	Tansey and Steppleton 1991
1992	1,282,373	Johnson 1994
1995	1,354,378	Johnson et al. 1997
1997	1,358,020	Johnson and Wells 1999
1999	1,265,888	Johnson and Wells 2002
2001	1,174,655	Johnson and Wells 2004
2003	1,171,419	Johnson and Wells 2005

2.2 TIMBER PRODUCT OUTPUT STUDIES

State-wide forest surveys began soon after the passage of the McSweeney-McNarey Act and continue to this day under the auspices of the FIA program. Currently, the USFS maintains an online database at http://ncrs2.fs.fed.us/4801/fiadb/rpa_tpo/wc_rpa_tpo.ASP which allows access to the 1997, 2002, and 2007 RPA assessments. Source data for the 1997 assessment is listed at http://ncrs2.fs.fed.us/4801/timberproducts/DATASOURCES.HTM, but this list has not been updated for the 2002 or 2007 assessment. Study design and implementation is handled at the USFS station level, with responsibility given to each station's FIA unit.

TPO designs vary both by station and sometimes by state within a station. What is common to all is that a complete census is attempted by canvassing all primary mills in each state (T.G. Johnson, C. E. Keegan, R. J. Piva, E. H. Wharton, personal communications, 11/15/2005). Data are collected into reports and nonrespondents are estimated by unpublished internal methods or they are not reported--suggesting a 100 percent census or that they were ignored. Some examples describing these internal methods follow. For Georgia, "In the event of a nonresponse,

data collected in previous surveys were updated using current data collected for mills of similar size, product type, and location (Johnson and Wells 2005, p i.)." For Arizona, "Published sources and data from various land management agencies were used to make estimates of any nonrespondent firms (Keegan et. al. 2001, p 4)." For Indiana, "IDNR utilization and marketing specialists provided estimates based on prior knowledge and contacts for a few Indiana mills that did not furnish complete data (Blythe, McGuire, and Smith 1987 p 5)." Summary reports of TPO data therefore are the general publication type, with these reports typically lacking any emphasis on statistical inference.

2.3 MULTIPLE IMPUTATION IN THE FORESTRY SETTING

Imputation in statistical terminology refers to the replacement of a missing value with some other value (hopefully a plausible one). If this process is performed just once for each missing value in the dataset, then the process is referred to as *single imputation*. An example of this approach is substituting the mean of the non-missing values. If more than one complete dataset is generated, then the method is called *multiple imputation.* Much of the theory was developed by Rubin (1987). See Scheuren (2005) for an interesting first-hand history of multiple imputation.

While multiple imputation is a prevalent and widely used method within non-forestry fields, it has had limited application within the field of forestry (McRoberts 2001) . Van Deusen (1997) details several missing data methodologies, including imputation, for handling nonoverlapping, annual, systematic samples. However, in that paper no forest survey studies are cited which employed multiple imputation and the focus of the paper was describing the methods rather than actually employing them.

The Southern FIA unit of the USFS explored using multiple imputation (Reams and McCollum 2000) to develop means for the five panels in a cycle of what would later become the Enhanced FIA Program (Bechtold and Patterson 2005). When examined under this framework, each panel essentially is missing eighty percent of the plots. This study used hot deck imputation to generate three separate datasets. The three data sets were then combined using the methods of Rubin (1987). Interesting to note is that no multiple imputed means fell within the confidence limits generated by just the measured plots. It is not clear whether there were true statistical differences however as the methods for parameter hypothesis testing presented by Rubin (1987) were not utilized.

McRoberts (2003) examined several methods for handling missing forest plots, several of which included multiple imputation techniques. This study centered on finding estimates for mean volume per unit area. Five classes of replacing observations (imputation) were utilized: PREVIOUS, STRATUM-U, MODEL+U, IMPUTE and IMPUTE-S. The PREVIOUS model imputed the plot value from the previous inventory. The last four listed employed the multiple imputation methods of Rubin (1987) to calculate the mean, with differing imputation strategies. For the STRATUM-U imputation, a random value u was added to the substratum mean, where u was randomly generated number ($N(0,\sigma)$). The value for σ incorporated variability from both the substratum mean and variability around this mean. MODEL-U followed a similar strategy, instead using a regression model to make a prediction for the mean to which a similar randomly generated u was added. Both IMPUTE AND IMPUTE-S utilized five of the most similar plots to the missing plot and randomly selected one. The Impute imputation was drawn from all private

plots and the IMPUTE-S imputation was drawn from the stratified plots. Beyond those, a mean generated by ignoring the missing plots, termed IGNORE, was included as well. This method increased the expansion factors and differs slightly from simply ignoring the missing plots completely. All methods produced mean estimates that did not statistically differ from the comparison mean.

Outside of forestry research, there are many available studies utilizing multiple imputation. The above references serve to demonstrate how infrequently these methods are used by forestry researchers, given the dearth of available studies. Exposure to methods for handling missing data therefore appear warranted in the field of forestry.

2.4 ROUNDWOOD PROCUREMENT DISTANCE STUDIES

TPO studies are typically performed at the statewide level and record the mills' regional procurement patterns based on county of origin data. Models examining the functional relationship of distance to industrial roundwood removal data are typically restricted to fixed procurement radius studies for receipts from mills. Models of this type have limitations in that the spatial aspect is restricted to a fixed distance and therefore not as flexible as a model employing a varying distance factor.

Wagner , Smalley, and Luppold (2004) studied log markets in the southern tier of New York to determine factors influencing the distribution, consumption, and merchandising of hardwood logs. Part of the study focused on the maximum procurement distance reported by the mills.

These maximum distances ranged from 50 to 250 miles. It was concluded that maximum distance was not correlated with size, although this was not statistically tested.

Alderman and Luppold (2005) sampled thirty logging operations in WV as part of a study of roundwood markets. The average haul distances to market were calculated for three regions of the state, and were broken down by several product classes: sawlogs, peeler logs, OSB, pulpwood, and rustic fencing. Haul distance varied considerably by both product and region, with distances from 20 to 95 miles reported. No statistical testing was undertaken to determine if these averages were significant by product or region.

Schwab, Bull, and Maness (2005) developed a demand equation incorporating distance for a study of Finnish sawmills. There were 105 mills sampled for the study over a six region area of Finland. The study utilized a weighted least squares (WLS) analysis with a logarithmic transformation of the cubic meter roundwood volume delivered as the dependent variable. Distance was demonstrated to have a negative impact on roundwood volume. In addition, mill capacity and sawmill size index were found to have a positive effect on roundwood volume.

Anderson and Germain (2007) studied sawmill wood procurement for the northeastern US. This study included three different procument radii: the average distance encompassing ninety percent of log supply, the farthest distance to stumpage, and the farthest distance to logs. These three measures were rigorously tested under a Multivariate Analysis of Variance (MANOVA) model with follow up Analysis of Variance (ANOVA) and Tukey's HSD multiple comparions performed. The MANOVA had six treatment groups based on two mill types (hardwood,

softwood) and three mill sizes (small, medium, and large). It was found that both mill type and mill size were significant in affecting the three measurement radii within the MANOVA analysis. Further, mill size was significant within the three ANOVA analyses while mill type was significant in only two of the three (ninety percent radius and farthest distance to logs). The ninety percent procurement radius for all mills was estimated at 30-70 miles.

Procurement studies at a multistate scale are few. This suggests a need for distance models capable of incorporating distance as a continuous variable to aid in the development of spatial allocation models.

CHAPTER 3: MULTIPLE IMPUTATION FOR THE MEAN AND TOTAL OF A STATE'S RECEIPTS

3.1 INTRODUCTION

Missing data is a problem that occurs for nearly all researchers at some point in the course of their research studies. It could be a subject who drops out, a lost record, a data entry error, an unanswered question, or any of a multitude of other reasons that observations are not available in a dataset. It is important to give careful consideration to the circumstances surrounding the missing data in order to give the variance credibility.

Missing data is said to fall into three categories: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR) (Rubin 1987). MCAR means that the missing data does not depend on auxiliary variables (Allison 2001). An example of an MCAR situation would be randomly deleting cases from a dataset. Missing at random is a weaker assumption and means that the missing data depends on the auxiliary variables. For instance, corporation size is recorded and small corporations fail to respond to an income question. NMAR means that the missing value has a dependence on some unmeasured missing auxiliary variable and the present auxiliary variables. An example might be that wealthy respondents do not answer questions about income in general but this refusal varies according to whether they shelter their income offshore or not. Further information may be found in Rubin (1987), Allison (2002), and Longford (2005).

Multiple imputation (MI) is one method of handling missing data. This method imputes missing values using other variables in the data set. Multiple data sets are produced, with results combined and parameters of interest estimated. There is great versatility with this method as it accepts inputs from various complete data set analyses, e.g. stratified means, ratio estimators, multiple regression, logistic regression, etc. It is also a good choice for situations where the data is MAR, as opposed to NMAR which generally requires other methods. The objective of this analysis will be to utilize multiple imputation to account for missing mill receipts and generate an estimate of the mean and total receipts for the 2003 year for Georgia.

3.2 MATERIALS AND METHODS

3.2.1 Data

The TPO data is from a census of Georgia primary mills for the years 2001 and 2003 and is detailed in Chapter 2. Specifically the type of mill, the number of employees in 2001, and the total mill receipts for each mill in 2001 and 2003 have been included in the missing data analysis. All continuous variables were log transformed in order to both satisfy multivariate normality requirements and to avoid spurious negative imputations. Three reference means were calculated: NONMISS, SINGLE1, and SINGLE2. NONMISS uses only the non-missing 2003 data, SINGLE1 uses the non-missing data plus the 2001 mill receipts imputed for any missing values (single imputation), and SINGLE2 was obtained from the published total for 2003 (missing values also singly imputed)(Johnson 2004). The SINGLE1 and SINGLE2 means are nearly identical and likely differ due to rounding or minor preparation differences.

3.2.2 Multiple Imputation

There are three phases to using multiple imputation. The first phase is to impute the missing data *m* times to generate *m* data sets with complete data. Second, each of the *m* data sets is analyzed by some standard statistical procedure. Last, the results from the statistical analyses of the *m* data sets (i.e the mean) are combined and inferences are made.

For stage one, the regression method for monotone data was utilized to impute values (Rubin 1987, p. 166-167). In this method, parameters from a posterior predictive distribution are simulated based on the non-missing data. This is accomplished through a Bayesian procedure called data augmentation and a noninformative prior is used (see Allison, 2000, p32-36 for details). These new parameters are used to estimate the missing values, along with a simulated standard deviate. This process is repeated iteratively to allow the imputation to stabilize, called the burn in, and then an imputed data set is output. Initial exploratory runs indicated a fast computation time, so the number of imputations was set at one hundred. Therefore, this data has 100 data sets where 94 observations are the same across data sets and the six missing values have been imputed in each data set.

At the second stage, the *m* complete data sets (here m=100) are analyzed for the parameter(s) of interest, *Q*. For each estimate $\hat{Q}_i$, a variance estimate, $\hat{U}_i$ is needed as well. For this analysis, the sample mean and its standard error are the desired Q and U. Again, for this data, there are 100 means and 100 standard errors, one of each from each of the 100 data sets.

At the third stage, these estimates are combined as follows:

Estimate for Q

$$\overline{Q} = \frac{1}{m} \cdot \sum_{i=1}^{m} \hat{Q}_i \qquad (1)$$

where

m=number of imputations

i=1…m

$\hat{Q}_i$=point estimate for the ith imputation (here the sample mean $\overline{Y}$).

Within-imputation variance

$$\overline{U} = \frac{1}{m} \cdot \sum_{i=1}^{m} \hat{U}_i \qquad (2)$$

where

m=number of imputations.

i=1…m.

$\hat{U}_i$=variance estimate for the ith imputation (here Var($\overline{Y}$))

Between-imputation Variance

$$B=\frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{Q}_i-\overline{Q}\right)^2 \qquad (3)$$

where

m=number of imputations

i=1…m

$\hat{Q}_i$=point estimate for the ith imputation (here the sample mean $\overline{Y}$).

$\overline{Q}$=mean of the $\hat{Q}_i$

Total Variance and Standard Error of Q

$$T=\overline{U}+\left(1+m^{-1}\right)B$$

where

m=number of imputations

$\overline{U}$=within-imputation variance

B=between-imputation variance

Finally, inferences are made by noting that

$$\frac{\left(Q-\overline{Q}\right)}{\sqrt{T}}\sim t_{\upsilon} \qquad (4)$$

where

$\overline{Q}$=estimate for Q

T=total variance

Q=hypothesized value.

υ=adjusted degrees of freedom=$\left(\upsilon_m^{-1} + \upsilon_{obs}^{-1}\right)^{-1}$

and

$$\upsilon_{obs}=\left[\frac{\left(1-\left(1+m^{-1}\right)\right)\upsilon_0\left(\upsilon_0+1\right)}{\upsilon_0+3}\right] \qquad (5)$$

$$\upsilon_m = (m-1)\left[1+\frac{\overline{U}}{\left(1+\frac{1}{m}\right)B}\right]^2 \qquad (6)$$

where

m=number of imputations

$\overline{U}$=within-imputation variance

B=between-imputation variance

There are two other statistics that are sometimes reported in multiple imputation studies. The first is the relative increase in variance due to nonresponse, *r*.

$$r = \frac{\left(1+\frac{1}{m}\right)B}{\overline{U}} \qquad (7)$$

where

m=number of imputations

$\overline{U}$=within-imputation variance

B=between-imputation variance

The second is the fraction of missing information about Q, $\hat{\lambda}$.

$$\hat{\lambda} = \left(\frac{r+2}{r+1}\right)\left(v_m + 3\right)^{-1} \qquad (8)$$

where

r= the relative increase in variance due to nonresponse.

v_m=degrees of freedom from above.

3.3 RESULTS

The within imputation variance and the between imputation variance were of similar magnitude (Table 3.2). The receipts mean was 7.31 million cubic feet per mill, with a standard error of 3.45 million cubic feet per million cubic feet. The relative increase in variance was 0.74, while the fraction of missing information was 0.43

Table 3.2: Mean and variance estimates.

		Variance			
Variable	**Mean**	**Within**	**Between**	**Total**	**se**
Receipts	7.31E6	6.86E12	5.02E12	1.19E13	3.45E6

The mean of the receipts obtained from the multiple imputation method as compared to each of the three reference means--NONMISS, SINGLE1, AND SINGLE2--did not differ statistically, p=0.68, 0.76, and 0.75 respectively (Table 3.3). Since totals are calculated directly from the means, this also signifies that the total from the multiple imputation method, 1.37 billion cubic feet, did not differ from the three references total as well (NONMISS=1.10 billion cubic feet, SINGLE1=1.16 billion cubic feet, SINGLE2=1.17 billion cubic feet). The estimate for the mean from the multiple imputation was estimated at from 4.4 to 14.2 million cubic feet (95% confidence interval).

Table 3.3: Tests for comparison of means.

	Mean	**t**	**p**	**df**
NONMISS	5.90E06	0.41	0.68	89.1
SINGLE1	6.20E06	0.32	0.75	89.1
SINGLE2	6.26E06	0.30	0.76	89.1

Both the mean and the median receipts derived from the 100 imputations were calculated for each mill (Table 3.4). Half of the mills have means that are roughly within two standards errors of the duplicate values, while three or four have medians close to the duplicate values.

3.4 DISCUSSION

There were three comparison values of interest: NONMISS, SINGLE1, and SINGLE2. Hypothesis tests indicate that the multiple imputation mean was not statistically different from any of the three means (α=0.05). Given the large standard error of the multiple imputation mean and the relatively few missing observations, this lack of significance is not unexpected. That is, the large standard error produces a wide confidence interval incorporating the comparison values and the six multiply imputed values are unlikely to dramatically affect the point estimate for the mean.

Examination of the six missing mills indicated that five out of the six had previous receipts values in the tens of millions for 2001 year. The last mill had a 2001 receipt total of 1.5 million cubic feet. This leaves open the possibility that mill size was a part of the missing data mechanism, and lends weight to considering the data as MAR and not MCAR. As is often the case, proving that the data is MAR is not generally possible as the data is missing and therefore unavailable for group comparisons.

Several of the mills (107, 172, and 176) had substantially larger computed means as opposed to previous receipts (Table 3.4). These larger mills also had the greatest standard errors for these

Table 3.4: Simple statistics by mills for the one hundred imputations.

Mill	Duplicate Value (ft^3)	Median (ft^3)	Mean (ft^3)	s.e.
3	27,358,359	3,302,199	7,883,252	1,156,785
86	1,494,608	641,481	2,701,589	615,268
107	19,243,627	16,779,245	86,111,145	26,180,735
125	12,562,459	3,369,341	11,815,347	2,630,716
172	14,526,985	14,159,072	41,782,163	8,158,994
176	17,610,912	17,789,935	49,037,553	32,677,409

means. However, the medians for these three mills are very close to the duplicate value. This suggests a few wildly large imputed values which skewed the mean to some degree. On the other hand, some of the means are fairly close (Mills 86 and 125), while the medians are under the duplicate value. This data set had a few auxiliary variables with which to perform the multiple imputation, whereas it is suggested to have available a large pool of auxiliary variables to take advantage of correlations (Rubin 1987).

Several of the mills (107, 172, and 176) had substantially larger computed means as opposed to previous receipts (Table 3.4). These larger mills also had the greatest standard errors for these means. However, the medians for these three mills are very close to the duplicate value. This suggests a few wildly large imputed values which skewed the mean to some degree. On the other hand, some of the means are fairly close (Mills 86 and 125), while the medians are under the duplicate value. This data set had a few auxiliary variables with which to perform the multiple imputation, whereas it is suggested to have available a large pool of auxiliary variables to take advantage of correlations (Rubin 1987).

3.5 CONCLUSION

In this study, multiple imputation was employed to avoid potential biases that may have resulted from imputing previous year's values or that could have resulted from simply ignoring the unit nonresponse. While there were no differences detected between the MI mean and the reference means, the analysis did provide an opportunity to evaluate how difficult it might be to implement the MI method for a set of TPO data. The analysis was performed using SAS, demonstrating that given knowledge of the technique, adopters of the method have readily available commercial applications. Multiple imputation has been sparingly used in forestry, yet does enjoy wider use in other fields. There is a definite opportunity to utilize this method in TPO studies and other resource surveys conducted by the USFS, as nonresponse is a common problem for survey work. Future research could benefit from incorporating MI into the final analysis when surveys contain missing data, particularly if the sampling method incorporates some sort of variance reduction technique, such as stratification or ratio estimation.

CHAPTER 4: ESTIMATORS FOR THE MEAN AND TOTAL OF A STATE'S RECEIPTS

4.1 INTRODUCTION

Complete canvasses can be problematic due to lack of response, limited monetary resources, lack of available personnel, etc. Oftentimes quality information can be obtained more readily and more easily from sampling. TPO studies are currently conducted on a 100 percent canvass basis, yet as has been mentioned previously, actual responses are not 100 percent. Estimation procedures are needed for both the mean and totals for a state's receipts. This analysis will explore several methods for obtaining means and totals with particular emphasis placed upon developing estimators with small variances (hence small bounds). Methods to be considered include: simple random samples, stratification, ratio estimators, and combined (stratified) ratio estimators.

4.2 MATERIALS AND METHODS

4.2.1 Data

Data sources are previously described in Section 2.2.1. This analysis utilizes the receipts and number of employees values from 2001.

4.2.2 Simple Random Sample Estimators

Collecting a simple random sample (SRS) is one of the most basic methods for estimating population parameters such as the mean, variance, and total. This analysis will use SRS estimators as the basis for comparison with other more sophisticated estimators.

Estimate of Population Mean

$$\hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y} \qquad (1)$$

Estimated Variance of $\bar{y}$

$$\hat{V}(\bar{y}) = \frac{s^2}{n} \cdot \frac{N-n}{N} \qquad (2)$$

where

N=total number of mills in the state.

n=number of mills sampled.

i=1 to n.

y_i=receipts from mill i.

s^2=sample variance=$\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$. (3)

95% Confidence Interval for the Mean

$$\bar{y} \pm z_{0.95}\sqrt{\frac{s^2}{n} \cdot \frac{N-n}{N}}\ . \qquad (4)$$

$\bar{y}$, s^2, n, N as above.

Estimate of SRS Total

$$\hat{\tau} = N\bar{y}. \qquad (5)$$

Variance of $\hat{\tau}$

$$\hat{V}(\hat{\tau}) = \hat{V}(N\bar{y}) = N^2\hat{V}(\bar{y}). \qquad (6)$$

where

N = number of sampling units in the population.

$\bar{y}$ = sample mean.

$\hat{V}(\bar{y})$ as above.

4.2.3 Strata Selection by Cluster Analysis

Cluster analysis was selected as one of the methods to generate classes of mills. Originally, it was thought that the type of mill, e.g. sawmill, pulp mill, veneer, etc, might serve as a useful class variable for a stratified mean. However, the number of classes was too many and too few mills would be included in each class to serve as a useful sample frame.

A cluster analysis is a statistical technique that attempts to gather similar objects into a meaningful collection of groups. Data is collected on *m* variables for *n* objects. This analysis utilizes hierarchical clustering, which is one of many clustering techniques. The similarity matrix (or resemblance matrix) uses Euclidean distance. Agglomeration of clusters is based on the Ward's Minimum Variance linkage, with objects starting out singly and eventually being placed into groups based upon minimizing the sum of the squared distances weighted by cluster size (McGarigal, Cushman, and Stafford 2000). Here, two variables are included in the

clustering, total receipts and number of employees. There are a total of 85 mills with data for both variables.

Once the cases have been placed into groups, the number of groups to use needs to be selected, as the clustering produces from one group up to the total number of cases in groups. Strict criteria for determining the number of clusters does not exist. However, simulation studies performed by Milligan and Cooper (1985) and Cooper and Milligan (1988) indicate that the pseudo-F statistic, the pseudo-t^2 statistic, and the cubic clustering criterion (CCC) all perform well in determining the best number of clusters. Cluster sizes can be chosen by creating a scree plot of the CCC versus the number of clusters and examining the figure for peaks (Sarle 1983). McGarigal et al. (2000) suggest a combination view of the CCC, pseudo-F statistic, and the pseudo-t^2statistic where large pseudo-F statistics and increasing pseudo-t^2 statistics from one cluster to the next are examined at local peaks on the CCC scree plot. This combination then suggests a stopping size for the number of clusters.

4.2.4 Strata selection by the Dalenius-Hodges Cumulative Square Root of the Frequency Method

The second method chosen to generate a class variable for stratification was the Dalenius-Hodges (DH) cumulative square root of the frequency method (Dalenius and Hodges 1959). The author was exposed to this method in a Survey Sampling course and felt that it was perhaps an underutilized method that might prove useful.

The DH method is another method used to generate estimators with small variance. The DH method first bins the observations on the response variable and a frequency table is constructed. The interval between bins is chosen by the practitioner much like a standard histogram bin size is chosen. Once the frequencies for the bins are found, the square root of each frequency is taken. Then a cumulative total is taken for the square roots of the frequency. The practitioner will decide how many strata are desired. The cumulative sum is then divided by the number of strata, arriving at a rough category size, c. Stratum boundaries are then found by picking the bins closest in cumulative frequency to $1 \cdot c$, $2 \cdot c$, …, $(L-1) \cdot c$, where L is the number of strata selected. For consistency, L will be set to the number of strata obtained from the clustering analysis.

4.2.5 Stratified Random Sample Estimators

Stratified estimators will be employed using the groupings resulting from both the cluster analysis and DH method.

Estimate of the population mean

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i \ . \qquad (7)$$

Estimated Variance of $\bar{y}_{st}$

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\frac{s_i^2}{n_i} . \qquad (8)$$

where

L=number of strata.

i=1..L

$\bar{y}_i$ =estimated mean for mills from stratum i.

N_i=number of sampling units in stratum i.

N=number of sampling units in the population.

n_i=sample size for stratum i.

$$s_i^2 = \text{sample variance for stratum i} = \frac{\sum_{j=1}^{n_i}(y_j - \bar{y}_i)^2}{n_i - 1} \qquad (9)$$

Estimate of Stratified Sample Total (τ)

$$\hat{\tau} = N\bar{y}_{st} = \sum_{i=1}^{L} N_i \bar{y}_i \qquad (10)$$

Variance of $\hat{\tau}$

$$\hat{V}(\hat{\tau}) = \hat{V}(N\bar{y}_{st}) = N^2 \hat{V}(\bar{y}_{st}) = \sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\frac{s_i^2}{n_i}. \qquad (11)$$

4.2.6 Ratio Estimators

Ratio estimators utilize a well-correlated auxiliary variable as an aid to estimate either means or totals. Additionally, an estimate of the population ratio may be of interest. While potentially biased, this bias is negligible (of order 1/n) and disappears if the regression of y on x (variable of interest, auxiliary variable) is a straight line through the origin (Schaeffer, Menenhall, and Ott 2006).

Estimate of Population Ratio

$$r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \tag{12}$$

$$\hat{V}(r) = \left(\frac{N-n}{N}\right)\left(\frac{1}{\mu_x^2}\right)\frac{s_r^2}{n} \tag{13}$$

where

$$s_r^2 = \frac{\sum_{i=1}^{n}(y_i - rx_i)^2}{n-1} \tag{14}$$

n=number of mills sampled.

N= total number of mills in the state.

y_i=receipts from mill i.

x_i=number of employees at mill i.

Estimate of Population Mean

$$\hat{\mu}_y = r\mu_x . \tag{15}$$

$$\hat{V}(\hat{\mu}_y) = \hat{V}(r\mu_x) = \mu_x^2 \hat{V}(r). \tag{16}$$

μ_x=employee population mean .

Estimate of Population total

$$\hat{\tau}_y = r\hat{\tau}_x . \tag{17}$$

$$\hat{V}(\hat{\tau}_y) = \hat{V}(r\hat{\tau}_x) = \hat{\tau}_x^2 \hat{V}(r). \tag{18}$$

τ_x=employee population total.

4.2.7 Stratified Ratio Estimators (Combined)

There are two varieties of stratified ratio estimators, separate ratio estimators and combined ratio estimators. Separate ratio estimators estimate r (sample ratio) within the strata and these strata

estimates for r are then combined with weighted strata means and totals to arrive at population estimates for the mean or total. The combined ratio estimator estimates r from the stratified means of both the variable of interest and its auxillary variable. The combined estimate for r is then employed with the weighted stratum means or totals to arrive at the desired population estimates.

For small sample sizes within strata, it is recommended that the combined ratio estimator be used (Schaeffer et al. 2006). For larger sample sizes, the separate estimator generally provides narrower confidence intervals, especially when strata ratios differ markedly. This analysis will employ combined ratio estimators due to smaller sample sizes.

Combined Estimate for r

$$r_c = \frac{\bar{y}_{st}}{\bar{x}_{st}}. \qquad (19)$$

where

$\bar{x}_{st}$ =stratified mean for mill employees.

$\bar{y}_{st}$ =stratified mean for mill receipts.

Estimate of Population Mean

$$\hat{\mu}_{yC} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \mu_x . \qquad (20)$$

where

μ_x =population mean for mill employees.

$\bar{x}_{st}$ =stratified mean for mill employees.

$\bar{y}_{st}$ =stratified mean for mill receipts.

Estimated Variance of $\hat{\mu}_{yC}$

$$\hat{V}(\hat{\mu}_{yC}) = \sum_{i=1}^{j} \left(\frac{N_i}{N}\right)^2 \left(\frac{N_i - n_i}{N_i}\right) \frac{s_{ri}^2}{n_i} \tag{21}$$

$$s_{ri}^2 = \frac{\sum_{j=1}^{n_i} (y_{ij} - r_c x_{ij})^2}{n_i - 1} \tag{22}$$

where

L=number of strata.

i=1..L

n_i=number of mills sampled from the ith stratum

N_i=number of sampling units in stratum i.

N=number of sampling units in the population.

x_{ij}=number of employees for mill j from stratum i.

y_{ij}=ft^3 total for mill j from stratum i.

r_c=combined estimate for r.

Estimate of Population Total

$$\hat{\tau}_{yC} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \tau_x . \tag{23}$$

where

τ_x = population total for mill employees.

Estimated Variance of $\hat{\tau}_{yC}$

$$\hat{V}(\hat{\tau}_{yC}) = \hat{V}(N\hat{\mu}_{yC}) = N^2 \hat{V}(\hat{\mu}_{yC}) \tag{24}$$

4.3 RESULTS

4.3.1 Cluster Analysis Stratification and Estimators

Values for the CCC indicated a local peak at two clusters with a plateau from four to six clusters (Figure 4.1). The pseudo t^2 and pseudo F statistics suggested two, four, and six groups, with the pseudo t^2 statistic having more definition (higher values) for these three group sizes. Therefore, the data was divided into groups of two, four, and six by liberally considering all criteria.

The actual classification of mills into groups is based solely on the number of employees. While the receipts data is a useful auxiliary variable, the receipts data is unknown prior to sampling (particularly for new mills), whereas the employee data is easily obtainable. This restraint

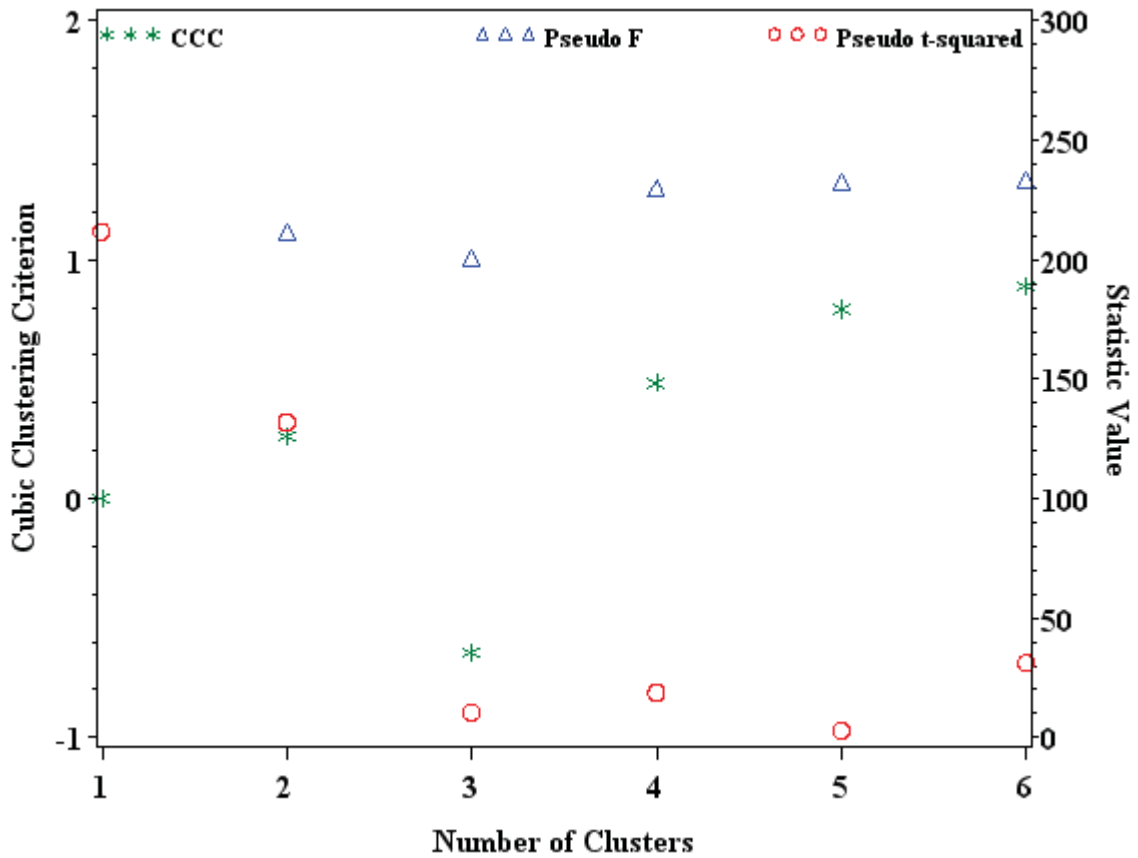

Figure 4.1: CCC, pseudo F and pseudo t^2 values for the cluster analysis.

necessitated examining the cluster solution and developing a stratification which modified the original cluster analysis group characteristics.

The employee ranges obtained from modifying the original cluster solution are presented in Table 4.3. Note that for the six group cluster solution, a sixth group was not obtainable based on the distribution of groups from the initial clustering. There is separation in two dimensions, but not reasonable separation along the employee dimension. Therefore, the modified cluster solution has five groups as the largest groups size examined, and is referred to as the 6/5 group solution.

Table 4.3: Cluster analysis group statistics and solution ranges based on number of employees.

Number of Groups		**Cluster Solution**				**Modified Cluster Solution**	
	Group	**Min**	**Max**	**Mean**	**sd**	**Min**	**Max**
2	1	1	381	73.22	91.48	*1*	*400*
	2	400	907	691.00	159.71	*401*	*907*
4	1	1	66	20.90	17.15	*1*	*40*
	2	37	381	168.15	95.21	*41*	*330*
	3	400	907	657.83	188.50	*331*	*730*
	4	692	800	757.33	57.46	*731*	*907*
6/5	1	1	66	20.90	17.15	*1*	*50*
	2	37	169	113.05	37.45	*51*	*200*
	3	240	381	299.12	47.13	*201*	*360*
	4	400	826	597.50	183.44	*361*	*720*
	5	692	800	757.33	57.46	*720*	*907*
	6	650	907	778.50	181.73	-	-

The sample variances for the various groups formed from clustering into two, four, and six/five groups are presented in Figure 4.2. Several groups had variances smaller than 1.00E12. These are the groups shown as gaps in the figure. The majority of the groups had variances less than that for the SRS variance, with only 5 out of the 23 groups exceeding this value.

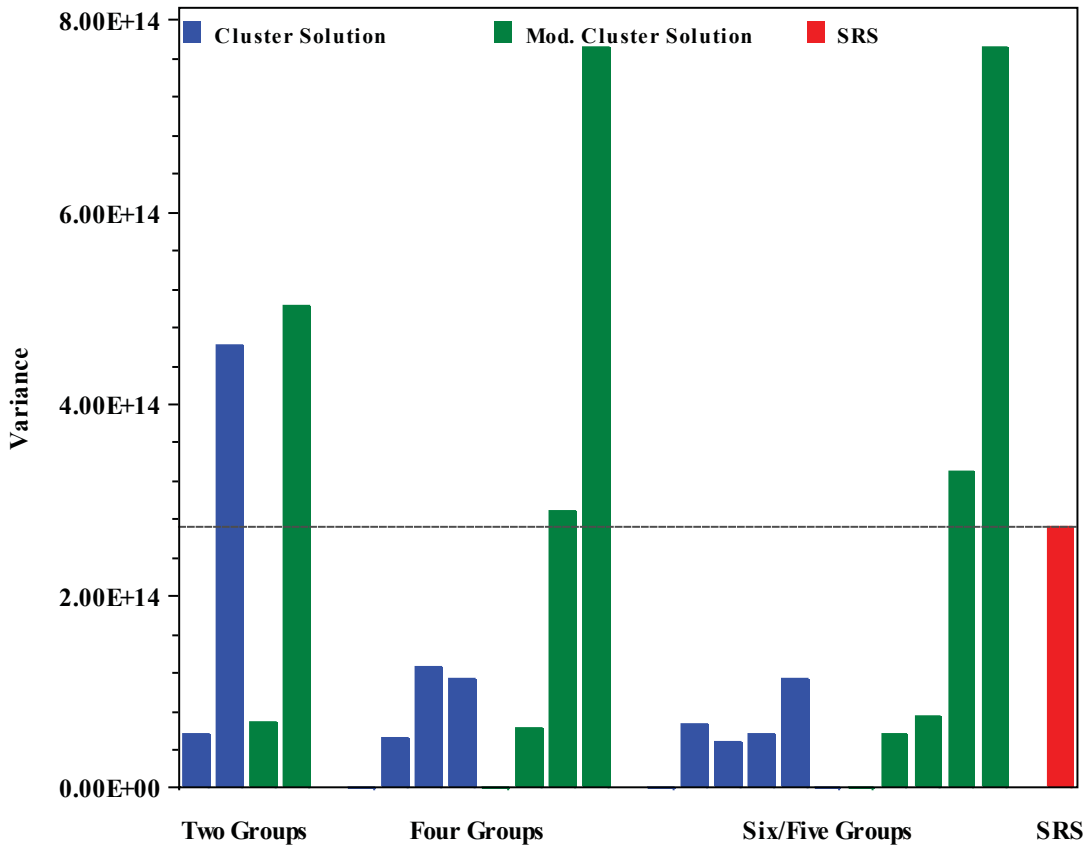

Figure 4.2: Sample variances for the cluster, modified cluster, and SRS groupings.

4.3.2 DH Method Stratification and Estimators

The divisions that led to the groupings by the DH method are presented in Table 4.4. Bin sizes of one million cubic feet were used to develop groupings. Once the cumulative square root of the frequencies (CSQRTF) was calculated, group divisions were made.

Table 4.4: DH method cutoff values for two, four, and six group sizes.

Range (millions ft^3)					Subgroup Totals (mills)		
Lower	**Upper**	**Frequency**	$\sqrt{\textbf{Frequency}}$	**Cumulative** $\sqrt{\textbf{Frequency}}$	**2**	**4**	**6**
0	1	35	5.92	5.92			35
1	2	12	3.46	9.38		47	
2	3	3	1.73	11.11			
3	4	1	1.00	12.11			16
5	6	2	1.41	13.53			
6	7	2	1.41	14.94			
7	8	2	1.41	16.35			
8	9	2	1.41	17.77			
10	11	1	1.00	18.77	60	13	9
12	13	3	1.73	20.50			
14	15	1	1.00	21.50			
15	16	1	1.00	22.50			
16	17	2	1.41	23.92			
17	18	1	1.00	24.92			8
19	20	3	1.73	26.65			
20	21	1	1.00	27.65		12	
21	22	2	1.41	29.06			
24	25	2	1.41	30.48			
27	28	2	1.41	31.89			10
35	36	2	1.41	33.30			
48	49	2	1.41	34.72			
64	65	1	1.00	35.72			
70	71	1	1.00	36.72			
85	86	1	1.00	37.72	25	13	7

As an example, consider the four group solution. The CSQRTF is 37.72, which is divided by 4. This results in an interval of around 9.4. Divisions are then made at 9.38, 18.77, and 27.65 creating four categories of mill sizes. There are 47, 13, 12, and 13 mills in each category respectively. This grouping provides a reasonable number of mills per stratum (and does so for the other two group sizes as well, two and six).

The employee ranges obtained from modifying the original cluster solution are presented in Table 4.5. There were no issues with overlapping groups as this method used one variable to separate the groups.

Table 4.5: DH method group statistics and solution ranges based on number of employees.

		DH Solution				**Modified DH Solution**	
Number of Groups	**Group**	**Min**	**Max**	**Mean**	**sd**	**Min**	**Max**
2	1	1	285	40.30	54.12	*1*	*100*
	2	70	907	374.64	270.58	*101*	*907*
4	1	1	103	21.04	19.59	*1*	*40*
	2	17	285	109.92	78.92	*41*	*150*
	3	70	381	193.33	103.43	*151*	*290*
	4	150	907	542.00	270.94	*291*	*907*
6	1	1	40	13.51	10.95	*1*	*20*
	2	14	285	58.69	64.17	*21*	*90*
	3	37	240	111.78	61.71	*91*	*130*
	4	70	381	167.00	108.66	*131*	*220*
	5	134	826	328.70	231.72	*221*	*530*
	6	400	907	677.58	174.59	*531*	*907*

The sample variances for the two, four, and six group stratifications for both the DH solution and the modified DH solution, along with the SRS variance are presented in Figure 4.3. Again, several of the variances near or below the value 1.0E12 do not appear in the figure. Similar to the cluster solutions, the majority of the variances are below the value for the SRS solution, with only 6 out of 24 groups having variances greater than that of the SRS solution.

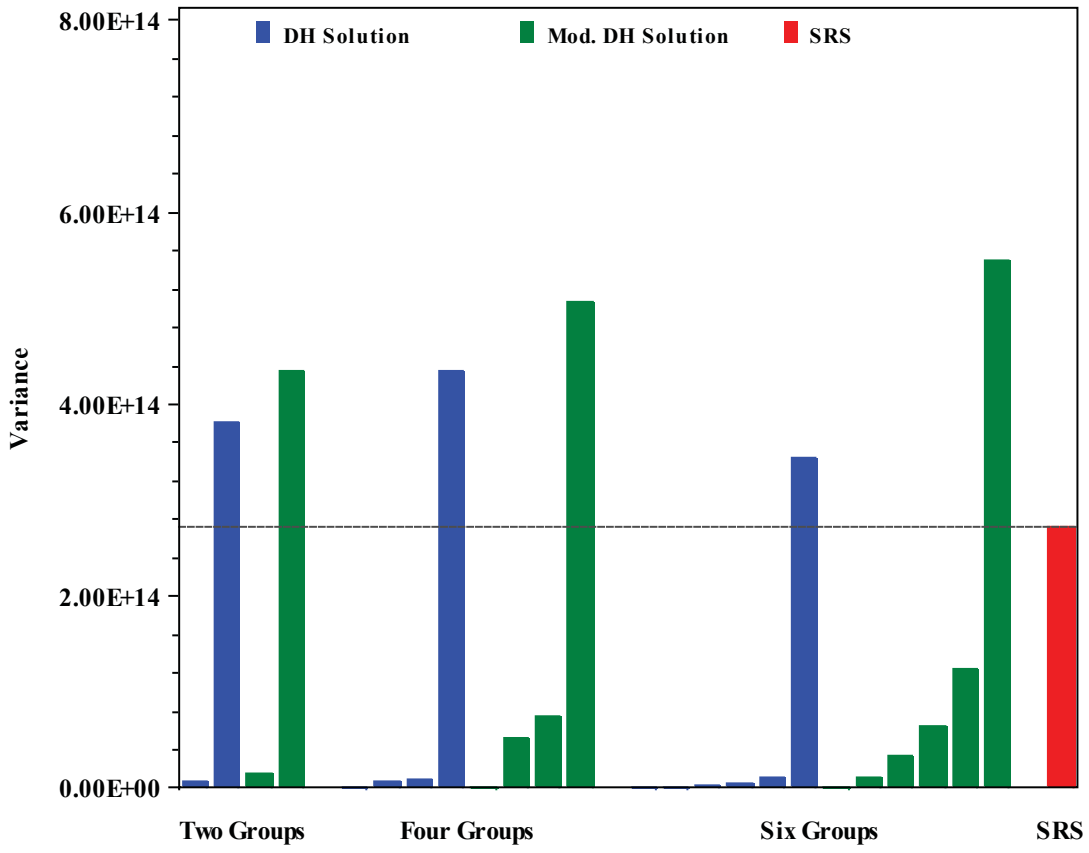

Figure 4.3: Sample variances for the DH, Modified DH, and SRS groupings.

4.3.3 Comparison of the Modified Cluster and DH classifications

The employee ranges obtained for the groupings in both the DH and cluster classifications differed when considering the same number of groups. The two group solution for the cluster analysis was divided nearly in the middle of the employee range, while the DH method had a much lower break at about 100 employees (Figure 4.4). For the larger number of groups, four and 5/6, the cluster solution exhibited more even employee ranges in its group sizes. The DH solution had some narrow range groups and some wider range groups for both its four and six group solutions.

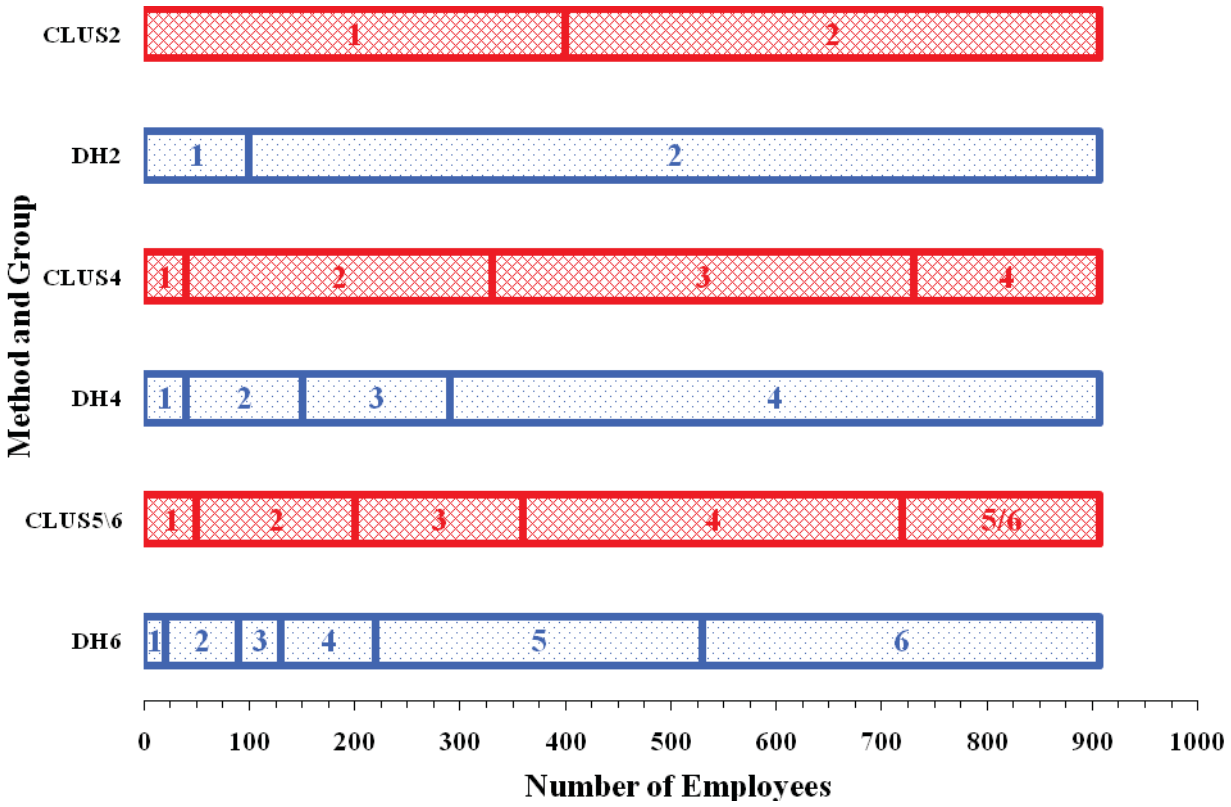

Figure 4.4: Comparison of the modified cluster and modified DH classifications.

The modified cluster solution generally has group variances higher than those of the cluster solution, although it should be noted that the groups do not always comprise the same mills. The groups having the greatest number of employees tended to have the highest variances, and those of the modified cluster solution were larger than those of the cluster solution. The two group solution had the greatest similarity between the cluster solution and the modified cluster solution.

4.3.4 Relative Efficiencies of Estimators

The estimated relative efficiencies of the estimators are included in Table 4.6. Efficiencies greater than one suggest that the estimator in the row position of the table is more efficient that the estimator in the column. However, it should be noted that these are calculated variances and not theoretical variances, and there exists the possibility that any differences are due to random chance.

Many comparisons are presented in Table 4.6. The number of groups is the foremost factor to consider and these comparisons are possible at many levels: within a stratification method and type of means analysis (e.g. DH method, stratified mean analysis), across stratification methods (e.g. cluster stratification vs. DH stratification), and across means methods (cluster stratified mean vs. combined ratio (cluster)).

First, consider the comparisons for group size within a stratification method and type of means analysis. These are found in the 3X3 blocks (blue blocks) along the main diagonal of Table 4.6. For each of the four combinations produced from crossing stratification method and means method (e.g., Combined ratio (cluster)), relative efficiency **improves** for increasing numbers of

groups. That is, six groups are more efficient than four groups which are more efficient than two groups, and these relationships are transitive.

Next, group sizes can be compared within the stratified mean method. When comparing the cluster method to the DH method within stratified means, the relative efficiencies range from 0.8-1.3 (main diagonals, yellow blocks, Table 4.6). A similar range in relative efficiencies, 0.7-1.4, results for the cluster vs. DH comparison within the combined ratio method (main diagonals, red blocks, Table 4.6).

The final group size comparison considered is across means methods. Comparing the cluster analysis method across means methods (main diagonals, green blocks, Table 4.6), RE ranges from 0.4 to 2.4. For the DH method, the range across means methods is 0.4-2.3 (main diagonals, purple blocks, Table 4.6).

The last set of comparisons is each individual estimator vs. the SRS (first column, Table 4.6). All estimators have greater than one relative efficiency vs. the SRS. Additionally, all specific stratification and means method types show increasing efficiency vs. the SRS with increasing group size.

Table 4.6: Estimated relative efficiencies for all estimators.

RE(Row/Cols)	SRS	Ratio	Clus. Stratification 2	Clus. Stratification 4	Clus. Stratification 6	DH Stratification 2	DH Stratification 4	DH Stratification 6	Comb. Ratio 2 (Clus.)	Comb. Ratio 4 (Clus.)	Comb. Ratio 6 (Clus.)	Comb. Ratio 2 (DH)	Comb. Ratio 4 (DH)	Comb. Ratio 6 (DH)
SRS	1.0	0.2	0.3	0.1	0.1	0.4	0.2	0.1	0.1	0.1	0.1	0.2	0.1	0.1
Ratio	**4.4**	1.0	**1.2**	0.6	0.4	**1.6**	0.8	0.4	0.5	0.4	0.3	0.7	0.5	0.3
Clus. Stratification 2	**3.5**	0.8	1.0	0.5	0.4	**1.3**	0.6	0.3	0.4	0.4	0.2	0.6	0.4	0.2
Clus. Stratification 4	**7.0**	**1.6**	**2.0**	1.0	0.7	**2.6**	**1.2**	0.6	0.8	0.7	0.5	**1.1**	0.8	0.4
Clus. Stratification 6	**10.0**	**2.3**	**2.8**	**1.4**	1.0	**3.7**	**1.8**	0.8	**1.2**	**1.0**	0.6	**1.6**	**1.2**	0.6
DH Stratification 2	**2.7**	0.6	0.8	0.4	0.3	1.0	0.5	0.2	0.3	0.3	0.2	0.4	0.3	0.2
DH Stratification 4	**5.6**	**1.3**	**1.6**	0.8	0.6	**2.1**	1.0	0.5	0.7	0.6	0.4	0.9	0.7	0.4
DH Stratification 6	**12.2**	**2.8**	**3.5**	**1.8**	**1.2**	**4.5**	**2.2**	1.0	**1.4**	**1.2**	0.8	**1.9**	**1.4**	0.8
Comb. Ratio 2 (Clus.)	**8.5**	**2.0**	**2.4**	**1.2**	0.9	**3.2**	**1.5**	0.7	1.0	0.9	0.6	**1.4**	1.0	0.5
Comb. Ratio 4 (Clus.)	**9.9**	**2.3**	**2.8**	**1.4**	1.0	**3.7**	**1.7**	0.8	**1.2**	1.0	0.6	**1.6**	**1.1**	**0.6**
Comb. Ratio 6 (Clus.)	**15.4**	**3.5**	**4.4**	**2.2**	**1.5**	**5.7**	**2.7**	**1.3**	**1.8**	**1.6**	1.0	**2.5**	**1.8**	1.0
Comb. Ratio 2 (DH)	**6.3**	**1.4**	**1.8**	0.9	0.6	**2.3**	**1.1**	0.5	0.7	0.6	0.4	1.0	0.7	0.4
Comb. Ratio 4 (DH)	**8.6**	**2.0**	**2.4**	**1.2**	0.9	**3.2**	**1.5**	0.7	**1.0**	0.9	0.6	**1.4**	1.0	0.5
Comb. Ratio 6 (DH)	**16.0**	**3.7**	**4.5**	**2.3**	**1.6**	**5.9**	**2.8**	**1.3**	**1.9**	**1.6**	**1.0**	**2.6**	**1.9**	1.0

4.3.5 Confidence Intervals for Totals.

The 95% confidence intervals for the totals of the various methods range from a lower limit of 1.07 billion ft^3 to a high of 2.10 billion ft^3 (Figure 4.5). Confidence interval ranges are narrowest for the combined ratio estimators, while the cluster and DH stratified totals are slightly broader. The widest interval (1.26, 2.10) billion ft^3 is that of the SRS.

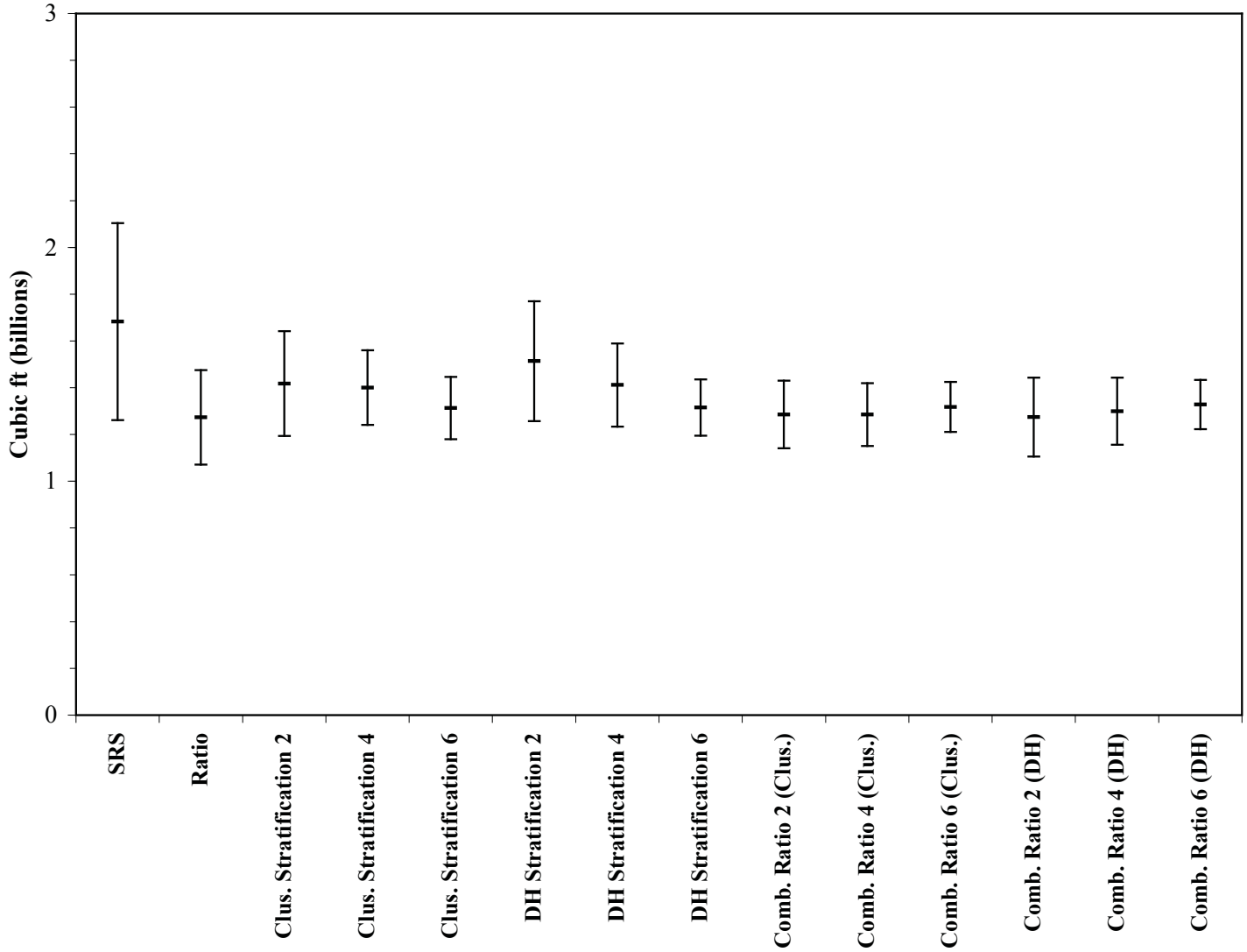

Figure 4.5: Estimates for total receipts by all methods.

4.4 DISCUSSION

4.4.1 Classification Methods Group Ranges

There were clear differences between the employee ranges for each group obtained by the cluster method and by those from the DH method. There were many groups with narrow ranges for the DH method, whereas there is more uniformity of employee ranges for the cluster solutions within specific group size solutions, e.g., Clus2 Figure 4.3. This is unexpected, as it would seem like the DH method might lead to more uniform groupings due to the initial binning of observations.

This difference in employee range sizes within the DH solution also leads to more groups for mills with fewer employees. That is not unexpected, as the method naturally concentrates high frequencies into narrow range categories. Stated differently, the CSQRTF is divided into a set number of intervals, with each interval a fixed fraction of the total CSQRTF. Increasing the number of similar observations for the variable of interest will force the determined range to be narrower. That stands to be a useful technique if the variable of interest is strongly correlated with the auxiliary variable used to classify into strata.

4.4.2 Group Sizes

There is a clear increase in estimator efficiency with increasing group size within specific stratification and means methods (Table 4.6, blue cells). The stratifications by both the cluster and DH methods allow for a decrease in variability within the groups formed by the methods. The RE(6 groups/2 groups) was greatest in all instances, with a high of 4.5 calculated for the DH stratification. For the state of Georgia, stratification into four or six groups appears to be a reasonable approach to reducing estimator variance.

4.4.3 Stratification Methods

Comparison of the stratification methods for similar group sizes can be considered within a means method, i.e., stratified or ratio estimator. For these data, neither the cluster nor the DH method appears to be more efficient (Table 4.6, diagonals of yellow and red cells). The relative efficiencies for these comparisons are close to one. Therefore, neither method is recommended over the other. However, the stratification itself is an improvement in efficiency over the SRS estimator and in most cases over the simple ratio estimator.

4.4.4 Means Methods

By holding the stratification method and group size constant, the effect of the means method can be examined. Here, there appears to be a clear advantage to using a combined ratio estimator versus a regular stratified mean (Table 4.6, diagonals, green and purple cells). Relative efficiencies range upwards from 1.3, thereby indicating that the additional information contained within the employee numbers is providing smaller variances on the estimate of the total. In

addition, the much simpler comparison for the RE(Ratio/SRS) yields an efficiency of 4.4 (Table 4.6), further evidence that the number of employees is a useful auxiliary variable.

4.5 CONCLUSION

New approaches to conducting TPO studies demonstrate that statistical estimation procedures can be of benefit to generating estimates for a state's mill receipts. Results from this analysis indicate that stratification coupled with ratio estimators based on employee numbers at mills can greatly reduce the variance of estimated totals. These reductions will allow for an expected reduction in the number of samples needed to achieve desired bounds on a state's receipts total. These reductions will be demonstrated in the next chapter.

CHAPTER 5: SAMPLE SIZE ESTIMATES FOR TOTAL STATE RECEIPTS

5.1 INTRODUCTION

Well designed studies make efforts to utilize previously collected data in order to gauge sample size requirements. Future TPO studies will benefit from utilizing prior years' data to provide guidelines on the sample sizes needed to achieve desired bounds on a state's receipts total. Additionally, firm estimates of the number of samples needed can aid in the budgeting of time and personnel. The objective of this chapter is to explore the effects that method of stratification, means methods, number of groups, and bound sizes have on the estimated sample sizes needed to calculate the state's receipt total.

5.2 MATERIALS AND METHODS

5.2.1 General Assumptions

All sample size estimates will be calculated from the standpoint of a 95% confidence interval. An assumption will be made that the sample sizes generated are sufficiently "large" such that the use of a z-value of two is appropriate. The bounds employed are: 1, 5, 10, 15, 20, & 25 percent of the estimated mean total (of all previously calculated totals, Figure 2.3).

5.2.2 Data

Data sources are described previously Sections 2.2 and 2.3. Estimated variances have been retained from calculations in Section 4.3.4 and are illustrated in Figure 4.4.

5.2.3 Sample Size Estimates for SRS Totals

Calculating sample sizes for a SRS total is relatively straightforward. Needed are: the number of sample units, an estimate for the population variance, and a bound.

Sample size to estimate τ using SRS

$$n = \frac{N\sigma^2}{(N-1)\frac{B^2}{4N^2} + \sigma^2}. \qquad (1)$$

N=total number of mills in the state.
σ^2=population variance
B=bound desired.
This method uses s^2 (see 2.2.2) as an estimate of $\sigma^{2.}$

5.2.4 Sample Size Estimates for Stratified Sample Totals

Calculating sample sizes for stratified totals is more involved than for a SRS. In order to determine the total needed sample size, an allocation strategy is needed. As the costs of obtaining a mill's information are expected to be nearly equal, yet differences in strata variances are evident, Neyman Allocation is employed to calculate the strata weights. These weights are then incorporated into the formula for the total sample size. Strata samples sizes are obtained from multiplying out the weights and the total size.

Neyman Allocation for determining stratum allocation weight

$$w_i = \left(\frac{N_i \sigma_i}{\sum_{k=i}^{L} N_k \sigma_k} \right). \qquad (2)$$

where

i=1…L.

k=1…L.

w_i=sample weight for stratum i.

N_i (N_k)=number of sampling units in stratum i (k).

σ_i (σ_k) =population variance for stratum i (k).

This method uses s_i (see 2.2.5) as an estimate of σ_i.

Sample size to estimate τ using stratified sampling

$$n = \frac{\sum_{i=1}^{L} N_i^2 \sigma_i^2 / w_i}{\frac{B^2}{4} + \sum_{i=1}^{L} N_i \sigma_i^2}. \qquad (3)$$

where

N=number of sampling units in the population.

i=1…L.

N_i=number of sampling units in stratum i.

σ_i=population variance for stratum i.

w_i=fraction of observations allocated to stratum i (from Neyman allocation above).

B=bound desired.

This method uses s_i^2 (see 2.2.5) as an estimate of σ_i^2.

5.2.5 Sample Size Estimates for Ratio and Combined Ratio Totals

Ordinary ratio sample sizes are as straightforward as the SRS.

Sample size to estimate τ using a Ratio estimator

$$n = \frac{N\sigma^2}{\frac{B^2}{4N} + \sigma^2}. \qquad (4)$$

N=total number of mills in the state.

σ^2=population variance.

B=bound desired.

Using s_r^2 (see 2.2.5) as an estimate of σ^2.

Sample sizes for the combined ratio estimator are calculated similarly to that of the stratified estimator (see 5.2.3) with the exception that s_{ri}^2 is used to estimate σ_i^2.

5.3 RESULTS

5.3.1 Estimated Sample Sizes for SRS and Ratio Totals

SRS sample sizes ranged downwards from 170 down to 82 to achieve bounds of 1 percent to 25 percent of the total receipts (Figure 5.1). Ratio sample sizes ranged from 169 down to 44 for the same bounds. Ratio sample sizes were less than SRS sample sizes for all considered bounds.

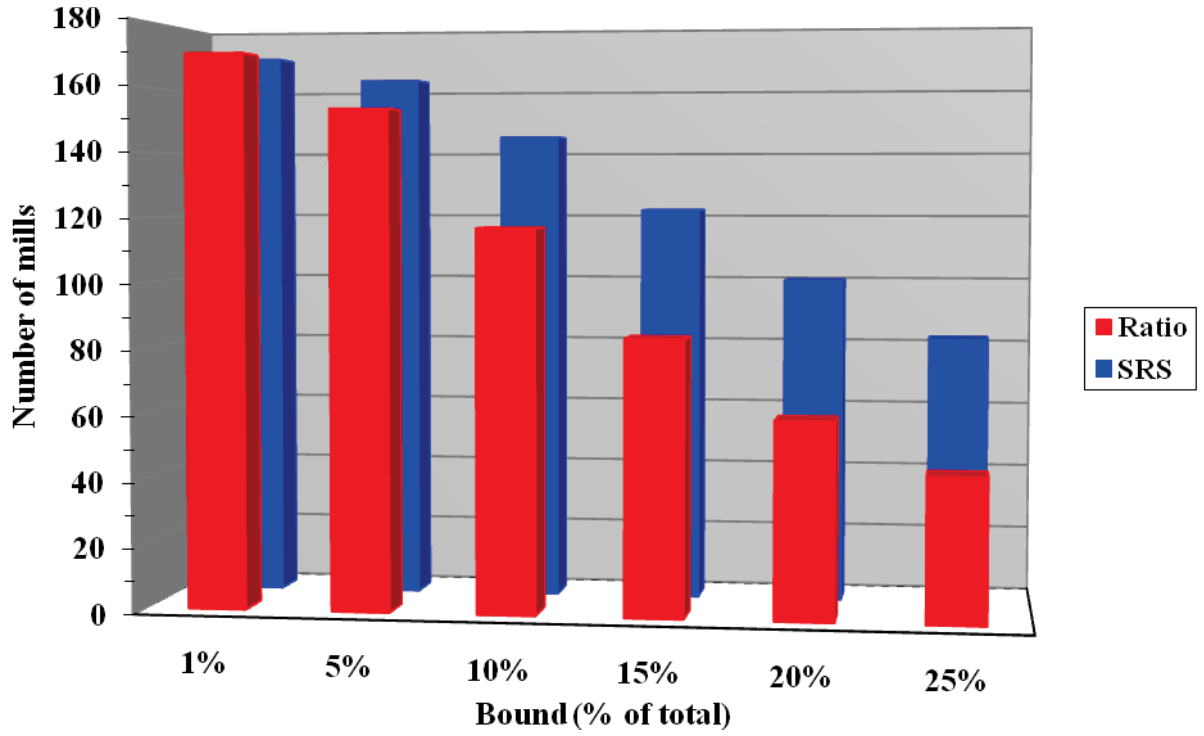

Figure 5.1: SRS and ratio estimated sample sizes for state totals from a population of 170 mills.

5.3.2 Estimated Sample Sizes for Stratified Totals

Stratum weights were calculated for all strata in each of the stratification methods using the class sample standard deviation estimates and the number of sampling units available for each stratum (Table 5.1). One important factor here that is the breakdown of stratum sampling units into the classes generated from the two stratification methods. The cluster analysis solutions have in all cases one stratum with 10 or fewer mills. For each of the 4 and 6/5 classifications, there is a stratum with only four sampling units available.

Table 5.1: Stratum sample sizes for state stratified totals.

					Sample size for several bounds					
	Stratum	**N_i**	**Est. σ_i**	**Weight**	**1%**	**5%**	**10%**	**15%**	**20%**	**25%**
Cluster 2	1	160	8.28E+06	0.86	128	120	98	76	58	44
	2	10	2.24E+07	0.14	22	20	17	13	10	8
Cluster 4	1	91	1.08E+06	0.11	8	8	6	4	3	2
	2	63	7.90E+06	0.55	43	39	30	21	15	11
	3	12	1.70E+07	0.22	17	16	12	9	6	5
	4	4	2.78E+07	0.12	10	9	7	5	3	3
Cluster 6/5	1	101	1.07E+06	0.13	9	8	6	4	3	2
	2	46	7.56E+06	0.43	30	27	20	14	10	7
	3	11	8.69E+06	0.12	8	7	5	4	3	2
	4	8	1.82E+07	0.18	13	11	8	6	4	3
	5	4	2.78E+07	0.14	10	9	6	5	3	2
DH 2	1	123	3.84E+06	0.33	31	29	25	21	17	13
	2	47	2.08E+07	0.67	64	61	53	43	34	27
DH 4	1	91	1.08E+06	0.10	7	7	5	4	3	2
	2	49	7.26E+06	0.36	27	24	19	15	11	8
	3	11	8.64E+06	0.10	7	7	5	4	3	2
	4	19	2.25E+07	0.44	32	29	23	18	13	10
DH 6	1	66	5.53E+05	0.05	3	3	2	1	1	1
	2	50	3.28E+06	0.22	16	14	10	7	5	3
	3	16	5.76E+06	0.12	9	8	5	4	3	2
	4	15	8.00E+06	0.16	11	10	7	5	3	2
	5	16	1.12E+07	0.24	17	15	11	7	5	4
	6	7	2.34E+07	0.22	16	14	10	7	5	3

Values in red indicate stratum sample size exceeds stratum size.

The estimated strata standard deviations obtained from the two stratification methods differ among classes almost exclusively by no more than one order of magnitude. The DH 6 solution is the only one where this is not true and sample standard deviations differ by two orders of magnitude

Weights for both stratification methods range from 0.05 to 0.86, with most weights ranging from 0.10-0.25. The Cluster 2 solution-stratum 1 weight is the greatest and the the DH 6-stratum 1 weight is the smallest. The Cluster 4-stratum2 and DH4-stratum 4 weights, at 0.55 and 0.43 are the most dissimilar within all the solutions.

These weights are then used to calculate the stratum sample sizes needed for each method from the calculated total sample size necessary for each bound (Figure 5.2). Due to a finite population, several stratum sample size estimates exceeded the total number of sampling units in the stratum, and are noted in red in Table 5.1 and as stippled columns in Figure 5.2.

Sample size estimates are highest at 150 for the one percent bound of the cluster stratification two group solution. The lowest sample size needed is 14 for the 25 percent bound on the DH stratification six group solution. Sample sizes decrease as expected as bound size increases and also decrease as the number of groups used to estimate the total increases (Figure 5.2).

5.3.3 Estimated Sample Sizes for Combined Ratio Totals

The division of sampling units is the same for each stratum here as was noted previously in section 5.3.2 (Table 5.1).

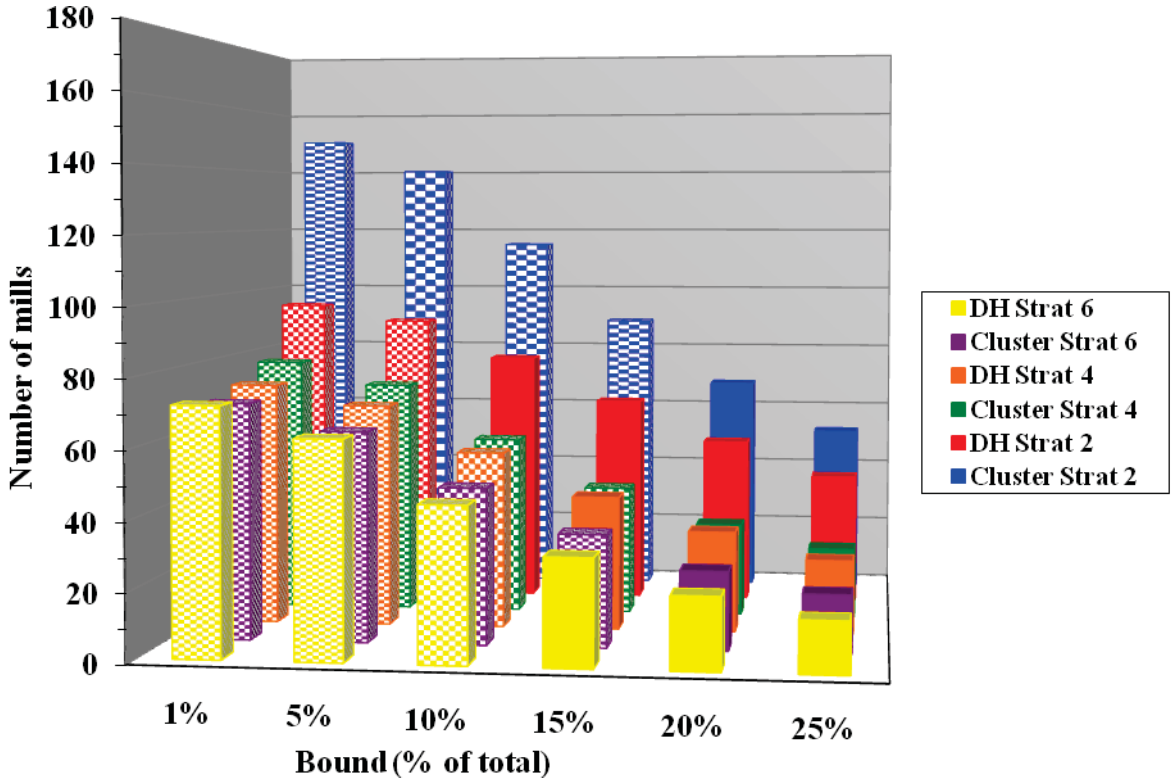

Figure 5.2: Stratified estimated sample sizes for state totals from a population of 170 mills. Stippled columns indicate one or more stratum sample sizes exceed the stratum for that particular method.

The estimated stratum standard deviations vary from one to two orders of magnitude within stratification solutions. Here, both the Cluster 4 and DH 6 solutions have the widest ranges in estimated standard deviations (Table 5.2).

Table 5.2: Stratum sample sizes for state combined ratio totals.

					Sample size for several bounds					
	Stratum	**N_i**	**Est. σ_i**	**Weight**	**1%**	**5%**	**10%**	**15%**	**20%**	**25%**
Cluster 2	1	160	4.95E+06	0.78	92	81	60	41	29	20
	2	10	2.19E+07	0.22	25	22	16	11	8	6
Cluster 4	1	91	9.78E+05	0.11	8	7	5	4	3	2
	2	63	7.03E+06	0.55	41	36	26	18	13	9
	3	12	1.27E+07	0.19	14	13	9	6	4	3
	4	4	2.98E+07	0.15	11	10	7	5	3	3
Cluster 6/5	1	101	1.02E+06	0.15	10	8	6	4	3	2
	2	46	6.09E+06	0.40	26	23	16	11	7	5
	3	11	7.31E+06	0.12	7	7	5	3	2	2
	4	8	1.37E+07	0.16	10	9	6	4	3	2
	5	4	2.98E+07	0.17	11	10	7	4	3	2
DH 2	1	123	2.45E+06	0.32	30	26	20	14	10	7
	2	47	1.37E+07	0.68	63	57	42	30	21	16
DH 4	1	91	9.81E+05	0.11	8	7	5	4	3	2
	2	49	5.56E+06	0.34	25	22	16	11	8	6
	3	11	1.09E+07	0.15	11	10	7	5	4	3
	4	19	1.74E+07	0.41	31	27	20	14	10	7
DH 6	1	66	4.88E+05	0.05	3	3	2	1	1	1
	2	50	2.51E+06	0.19	12	10	7	5	3	2
	3	16	5.59E+06	0.14	9	7	5	3	2	2
	4	15	7.97E+06	0.18	12	10	7	5	3	2
	5	16	8.02E+06	0.19	13	11	7	5	3	2
	6	7	2.37E+07	0.25	16	14	10	6	4	3

Values in red indicate stratum sample size exceeds stratum size.

Weights range from a low of 0.05 to 0.72, for the DH 6-stratum1 and Cluster 2-stratum 1 strata respectively. Weights are again in the 0.10-0.25 range, with some minor deviations from those weights as happened in the stratifications in section 5.3.2.

Sample sizes were again highest at 117 for the cluster stratification into two groups (Figure 5.3). The lowest sample size of 12 was for the DH stratification into six groups. The weights and stratum sample sizes calculated to estimate the combined ratio totals also result in several strata sample sizes that are unattainable (red text, Table 5.2).

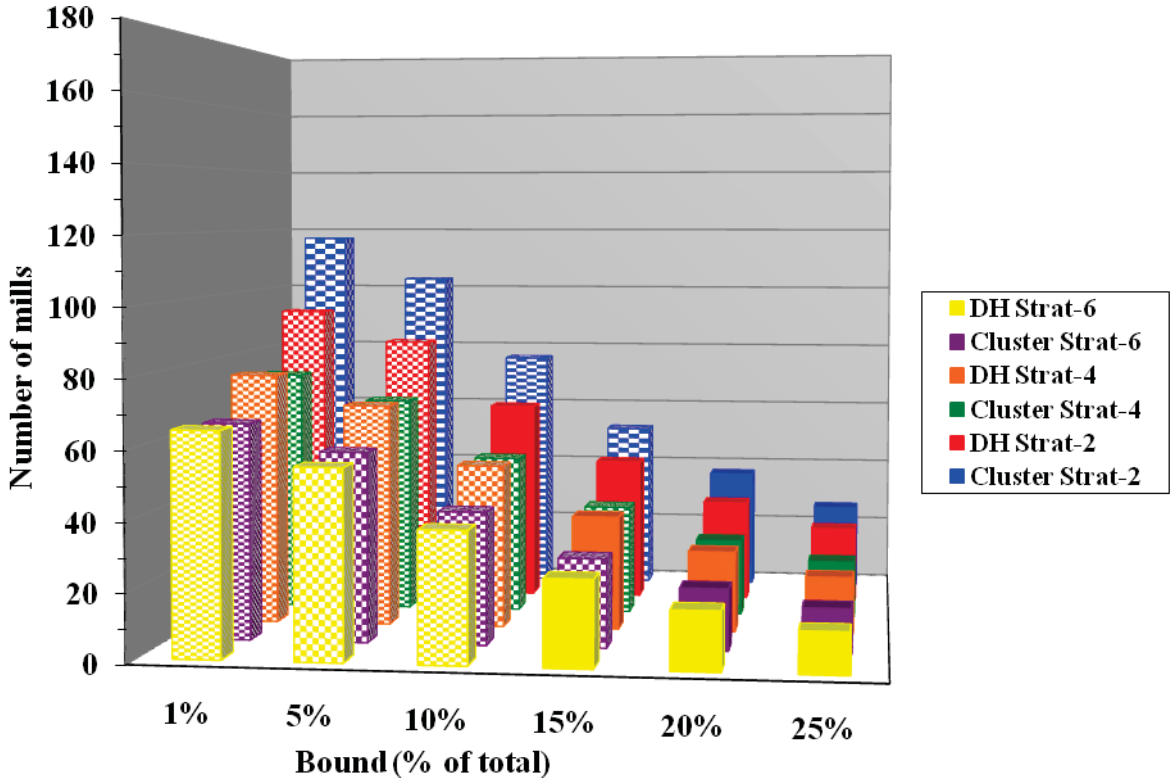

Figure 5.3: Combined ratio estimated sample sizes for state totals from a population of 170 mills. Stippled columns indicate one or more stratum sample sizes exceed the stratum for that particular method.

5.4 DISCUSSION

5.4.1 Group Sizes

Group size is an important factor in reducing the number of samples needed to achieve the several bounds examined (Figures 5.2 and 5.3). All of the cluster method by means method combinations show a decrease in the necessary sample sizes as the number of groups employed in the stratification increases.

5.4.2 Stratification Methods

There were few differences in sample size estimates for the two stratification methods (Cluster and DH). The differences were nearly immeasurable for the four and six group solutions within any means method. However, there were somewhat larger differences when comparing sample size estimates for the DH two group solutions to the cluster solutions for the stratified total (Figure 5.1). Differences for these comparisons occurred again for the combined ratio sample size estimates (Figure 5.2). These differences in the two group solution may be a result of only having two groups and that the cluster solution had an unbalanced distribution of sampling units within its two strata—N_1=170 and N_2=10.

5.4.3 Means Methods

There is a marked reduction in the number of samples needed for a fixed group size and stratification method when switching from a stratified total to a combined ratio total in nearly all cases (Figures 5.1 and 5.2). Giving specific consideration to the fact that some bound sizes

produce unattainable stratum sample sizes, at least 20 percent more samples are needed by the ordinary stratification total to achieve the same bound as the combined ratio total.

The DH 2 method using combined ratio estimators produces the only sampling strategy where a 10 percent bound on the total is attainable. Further, the DH method for the combined ratio estimator is the only method by which 15% bounds on the total are attainable (all group size solutions). All methods can be used if a 20% bound is an acceptable range for the total.

5.5 CONCLUSION

Stratification and the use of auxiliary information are demonstrably useful techniques in achieving reasonable bounds on a state's total receipts data. Slightly better results are produced from using the DH stratification method, allowing for bounds of 15 percent of the estimated total for both means methods, as opposed to the cluster method attaining only the 20 percent level. The use of a combined ratio total estimator allows for a significant reduction in the necessary sample size. For these data, the easily obtained value for the total number of employees provides a much greater benefit over using SRS estimators.

CHAPTER 6: SPATIAL ALLOCATION OF ROUNDWOOD RECEIPTS

6.1 INTRODUCTION

A further need for completing TPO studies are models which relate roundwood receipts spatially to the resource base. The mandate for reporting utilization is to record removals at the county level. All mills are contacted and information is gathered as to how much is removed and in which county. When a mill fails to respond, that translates to an underestimate for one or more counties in the region. Just a few missing mills can easily impact quite a substantial number of counties in the TPO study, particularly if the mill is a large sawmill or pulpmill. It is critical that sound methods be employed to estimate a missing mill's information. Therefore, this analysis is a first step in developing methodology for allocating mill roundwood receipts to counties both within the target state and its regional neighbors.

6.2 MATERIALS AND METHODS

6.2.1 Data

Basic geographic data was previously described in section 2.2.3. This analysis focuses on distance models which require further manipulation of the data. Two variations for a county center were used in the analysis. The first center generated was the geographic center (Geographic Center) for each county (Figure 6.1). The ten state county polygon layer was used as the input layer to the Mean Center Tool in ArcToolbox. The second county center generated

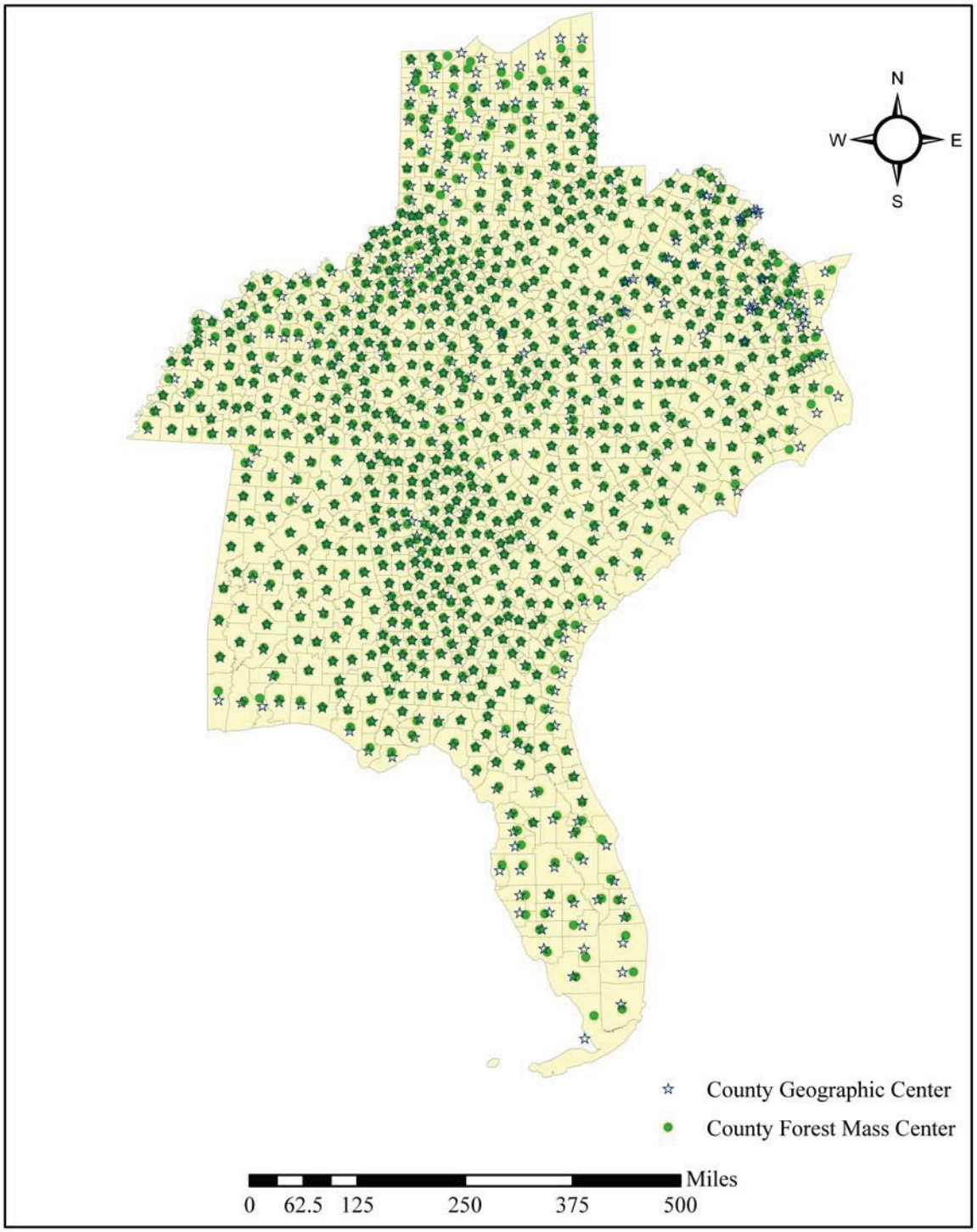

Figure 6.1: Ten state region showing County Geographic Centers and County Forest Mass Centers. Map created by author.

was the weighted forest center (Forest Mass Center) for the county. Here, FIA inventory plots were obtained from the inventory previous to the TPO collection year. If a plot has at least 10% stocking, that plot is classified as forested in the TPO data base. These forested plots were added to ArcMap as a points layer. The forested plots within a county represent a recorded number of acres. The forest center is then calculated as the center of the plot locations weighted by the number of acres (forested) represented by the plots. This was done using the Mean Center Tool as well using the acre weights from the FIA database as the weight field. The final result was two point layers with each point representing either the Geographic Center or the Forest Mass Center for the county.

The next step after obtaining the center layers was to calculate the distances between the mills and the two centers. Straight line distance was the first distance computed. The Point Distance tool in ArcToolbox was invoked using first the Geographic Center layer and the mill layer. This was repeated for the Forest Mass Center layer. This created two tables with distances from each mill to all county centers.

The second type of distance generated was the shortest road route from the mill to the county center. Roads for each county were integrated into a ten state road network built with Network Analyst (ArcMap) from the previously obtained TIGER roads layers (Figure 6.2). Mills and county centers were designated as the origin and destination layers, and the Network Analyst tools calculated the shortest road distances. Any mills or

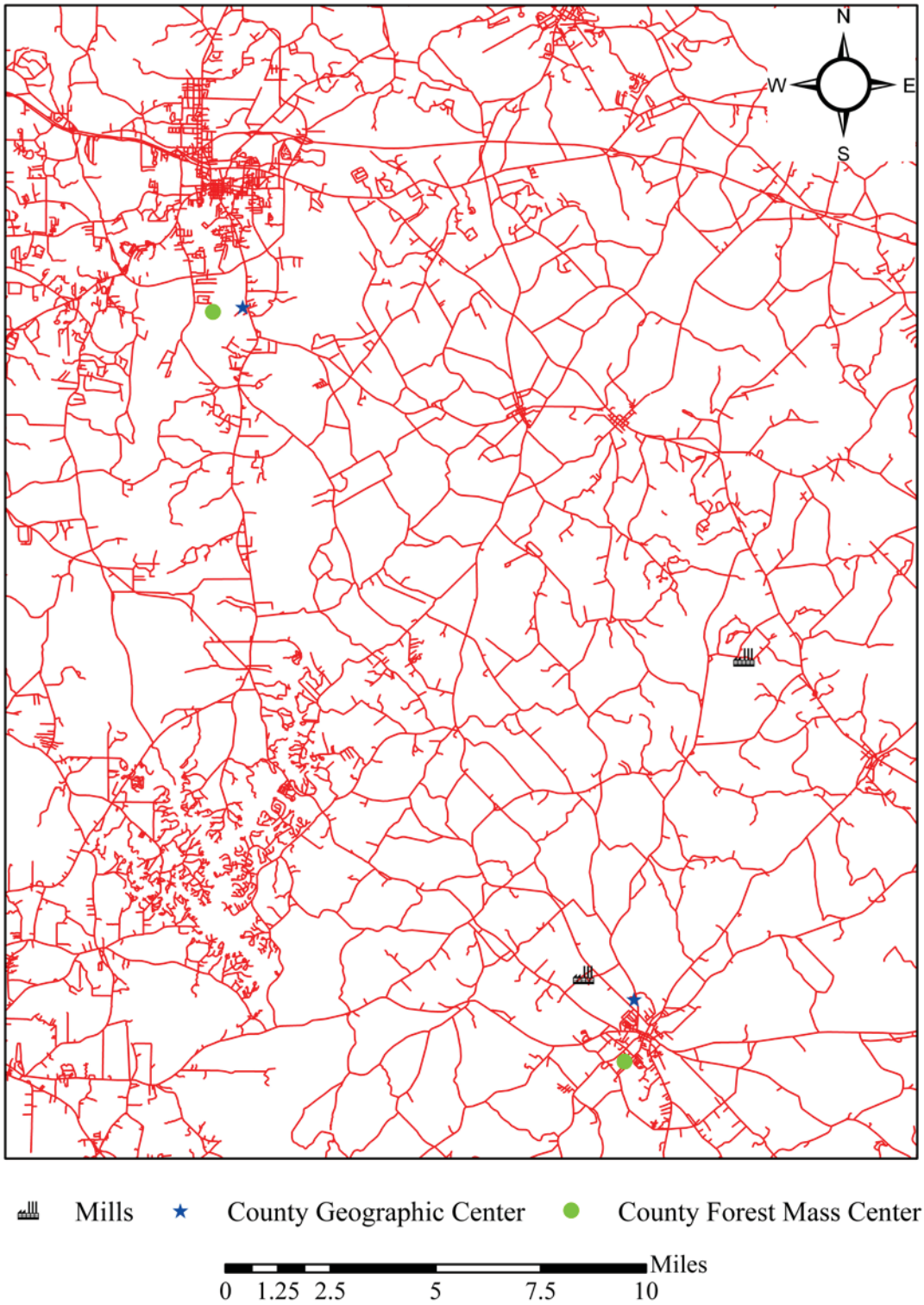

Figure 6.2: Road network example. Map created by author.

centers not directly on a road are considered to be on the closest road. Two tables of distances were again created for every mill to every geographic and forest mass center.

Four tables were thus created: straight line distance between mill and geographic center, straight line distance between mill and forest mass center, road distance between mill and geographic center, and road distance between mill and forest mass center. These distance tables were then joined to the table containing mill receipts by county, with this table also containing the employee numbers for each mill.

Last, the cumulative frequency of mill receipts was calculated for each mill. Receipts data for each mill were arranged in distance order and the cumulative frequency for each county recorded was computed. Each mill therefore has an empirical distribution of mill receipts by county.

6.2.2 Statistical Models

The models developed related the distances from the mill to the county centers to the cumulative frequency of the receipts received by the mill for the year 2001. Initial exploratory analysis suggested that the mill distributions might best be modeled by sigmoid functions, as there are two asymptotes present in the cumulative frequency, zero and one. Two types of sigmoid functions were selected to analyze the data: logistic and Gompertz functions.

Sigmoid functions are nonlinear in the parameters, and since each mill had repeated measures, this analysis is a nonlinear repeated measures design and nonlinear mixed model procedures were employed (SAS macro NLinMix). As there were no like studies detailing a known covariance structure, the following spatial covariance models were considered: spherical, exponential, Gaussian, linear, linear log, and power (Little, Milliken, Stroup, and Wolfinger 1996, p305). Models were evaluated using the corrected Aikake's information criterion (AICC) in SAS® (smaller AICC is better), as well as an examination of the residuals using the 2003 data as a validation set.

Nonlinear Model

$$\mathbf{y} = f(\mathbf{X}, \boldsymbol{\beta}) + \mathbf{e} \qquad (1)$$

where

y-vector of mill cumulative frequency receipts values.

X- matrix of observed independent variable values.

β-vector of unknown fixed effects parameters.

e-vector of random error.

f-is a function, here either a logistic or Gompertz function.

Logistic function

$$y=\frac{1}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_1X_2}} \qquad (2)$$

Gompertz function

$$y=e^{\beta_0 e^{\beta 1X_1+\beta_2X_2+\beta_3X_1X_2}} \qquad (3)$$

where

e-base natural log.

$\beta_0 \ldots \beta_3$- unknown fixed effects parameters.

X_1-distance.

X_2-number of employees

6.2.3 Procurement Radius

An estimate of a mill's procurement radius can be generated from any of the estimated models. Three procurement levels have been selected: 25, 50, and 75 percent. The levels selected represent procurement distances that encompass 25, 50, and 75 percent of each mill's receipts.

These values are substituted as values of the cumulative frequency into the final models and solved algebraically for the estimated distance given a fixed number of employees.

6.3 RESULTS

6.3.1 Models

AICC values for the multiple distance measures under each nonlinear model form are presented in Table 6.1. There was a wide range of AICC values within a model form for most distance measures. For instance, for forest centers using road distances to mills the AICC values ranged from -1972.7 to 122.7 using the logistic model and from -1964.8 to 90.8 using the Gompertz model. The geographic center with road distances exhibited very similar ranges, -1916.5 to 130.8 and -1905.3 to 98.8 respectively for the logistic and Gompertz models. The range on the geographic centers using straight line distance were smaller, ranging from -1217.2 to 113.4 for the logistic models and from -1287.5 to 82.4 for the Gompertz models. The forest center straight line distance models had much narrower ranges, 522 to 561.9 and 525.9 to 565.6 for the logistic and Gompertz models respectively.

The overall "best" models in terms of AICC were road distance models, which had AICC values less than -1900. The straight line distance models did not have as low values, with -1287 being the least. Models marked with asterisks in the table failed to converge.

Table 6.1: AICC Values

	Model	Geographic Straight Line	Forest Straight Line	Geographic Road	Forest Road
Logistic	Identity	105.4	553.9	121.3	114.7
	Sp(Exp)	-1034.7	553.9	-1916.5	-1972.7
	Sp(Expa)	113.4	561.9	130.8	122.7
	Sp(Gau)	107.4	555.9	124.7	116.7
	Sp(Pow)	-1034.7	522	-1916.5	-1972.7
	Sp(Powa)	-943.4	526	-1879.9	-1932.6
	Sp(Lin)	-434.9	*	-832.6	-889
	Sp(Linl)	-1217.2	*	-1830.3	-1862.4
	Sp(Sph)	-966.9	553.9	-1904.3	-1962.7
Gompertz	Identity	74.4	*	89.3	82.8
	Sp(Exp)	-1092.8	557.6	-1905.3	-1964.8
	Sp(Expa)	82.4	565.6	98.8	90.8
	Sp(Gau)	76.4	559.6	92.8	84.8
	Sp(Pow)	-1092.8	525.9	-1905.3	-1964.8
	Sp(Powa)	-994.8	529.8	-1854.4	-1913.9
	Sp(Lin)	-479.7	*	-853.1	-909.4
	Sp(Linl)	-1287.5	*	-1852.7	-1888.9
	Sp(Sph)	-1037.4	557.6	-1895.9	-1956.5

*indicates model did not converge.
Values in blue represent models with the smallest AICC values for that specific distance model.

Only models which produced unbiased residuals for the 2003 data were included for further consideration (Table 6.2). Models which had confidence intervals containing zero were considered unbiased. The mean residuals for the cumulative frequencies for all models using the 2003 data were very close to zero, ranging from -0.0082 to 0.0094. Standard errors were all very similar, ranging from 0.0050 to 0.0055.

Table 6.2: Confidence limits for cumulative frequency of mill receipts for models with unbiased residuals.

Model Type	Mean	LCL	UCL	s.e.
	Straight Line Distance-Geographic Center-Gompertz Model			
Exponential Anisotropic	0.0060	-0.0038	0.0158	0.0050
Power Anisotropic	-0.0030	-0.0135	0.0074	0.0053
	Road Distance-Geographic Center-Gompertz Model			
Identity	0.0084	-0.0014	0.0181	0.0050
Exponential Anisotropic	0.0082	-0.0016	0.0179	0.0050
Gaussian	0.0082	-0.0016	0.0179	0.0050
Power Anisotropic	-0.0043	-0.0149	0.0063	0.0054
Exponential	-0.0049	-0.0157	0.0058	0.0055
Power	-0.0050	-0.0157	0.0058	0.0055
Spherical	-0.0064	-0.0172	0.0044	0.0055
	Road Distance-Forest Mass Center-Gompertz Model			
Identity	0.0094	-0.0004	0.0192	0.0050
Exponential Anisotropic	0.0094	-0.0004	0.0192	0.0050
Gaussian	0.0094	-0.0004	0.0192	0.0050
Power Anisotropic	-0.0043	-0.0149	0.0063	0.0054
Exponential	-0.0070	-0.0177	0.0037	0.0055
Power	-0.0070	-0.0177	0.0037	0.0055
Spherical	-0.0082	-0.0189	0.0026	0.0055

Values in blue represent models with the smallest AICC values for that specific distance model.

6.3.2 Procurement radius

Two models were used to examine procurement radius at the three levels chosen (25, 50, and 75 percent). The models used were Gompertz models for the road distance to both the geographic and forest mass centers. Both models are very close in terms of coefficients. Considering both simultaneously, 25 percent of all receipts are obtained within 20 to 50 miles, and 50 percent of all receipts are obtained from 30 to 75 miles of the mill, for all employee numbers in mills in this study (Figures 6.3A & 6.3B. Seventy-five percent of all receipts fall within 110 miles of all mills, using the most conservative estimate (Figure 6.3B).

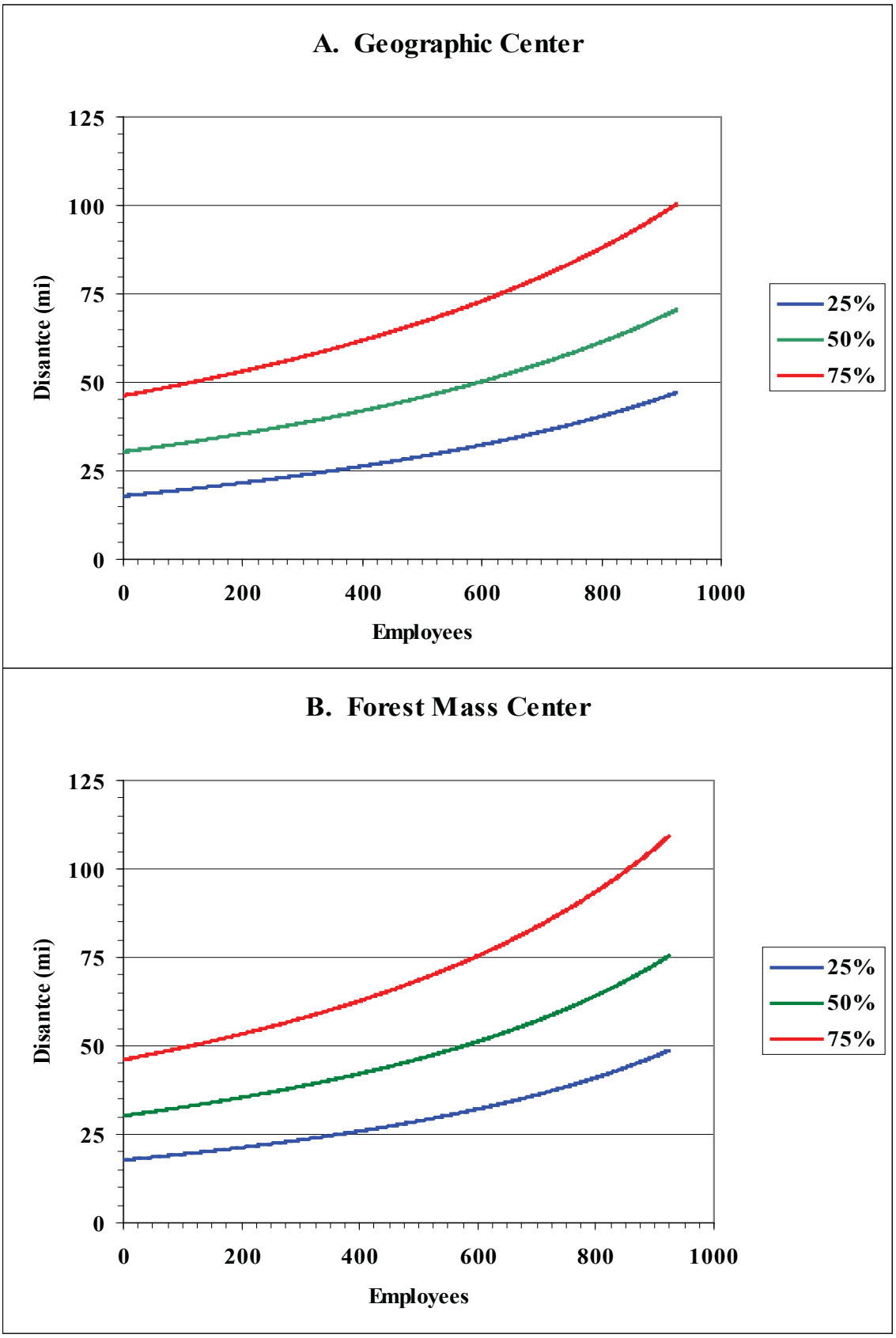

Figure 6.3: Mill procurement radius for cumulative receipts under two road distances types modeled under the spatial power covariance structure.

6.4 DISCUSSION

For this data set, the spatial exponential and power covariance types were chosen as best covariance structures. These two types performed equally well within the Gompertz functional form for road distances based on either the geographic center or forest mass centers. The means for the 2003 residuals were unbiased and the standard errors were very reasonable at 0.0055. Given that the two covariance structures produced nearly identical results, and that the power type has a slightly simpler model form than the exponential, the power type is better suited for future use.

In selecting a distance type to use, both the geographic center and forest mass center road models performed nearly identically. The road forest mass center had a lower AICC value, while the residual mean for the geographic center was lower than the forest mass (yet the standard errors were identical). In choosing a method to calculate cumulative frequencies, it appears equally useful to utilize either a forest mass center or a geographic center. Calculating the forest mass however, requires a lot of additional data computation. It is much simpler to calculate a geographic center. Therefore, for future modeling work, it is recommended that the spatial covariance power structure be used and that distances be calculated as road distances to the geographic center of the county.

6.5 CONCLUSION

The models developed using this methodology are an important beginning step to creating distance driven receipts models for mills in the state of Georgia. Future studies which contain simple total receipts will allow their implementation. An important second step will be to collect total receipts regardless of county data. In addition, a much needed second stage of research will be needed to determine what factors determine harvesting in a particular county. This will prevent estimates being made for counties were no harvesting occurs, e.g. cities. It is hoped that these models may also be transferable to other states, with models developed from past TPO studies.

CHAPTER 7: CONCLUSIONS AND RECOMMENDATIONS

7.1 SUMMARY

Data from two TPO surveys indicate a wide range of response rates from primary processing mills in Georgia, with response rates as low as 50.0 percent for the year 2001 followed by an improvement to 96.8 percent in 2003. Conversations with TPO specialists and review of the literature indicated that non-response rates in other state surveys is equally varied and that methods to address to address the problem are needed.

This research first considered multiple imputation as a method to adjust for non-response. Mean roundwood volume receipts per mill for the year 2003 were calculated using the methods developed by Rubin (1987). A mean of 9.3 million cubic feet was arrived at through the use of multiple imputation. While this value was higher than three reference means, NONMISS, SINGLE1, and SINGLE2, it was not significantly different from each of them.

Sampling methods were explored as an approach to reduce the number of mills included in future surveys. A variety of means and totals estimators were considered for possible use in a future sampling program. The means and totals estimators included: simple random sample, stratified sample, ratio, and combined ratio. Means and totals estimators incorporating stratification utilized two methods for stratification, a cluster analysis technique and the Dalenius-Hodges cumulative square root of the frequency method. There were two major factors to consider with these estimators, how efficient were they and could sufficient sample sizes be achieved given the size of the Georgia mill population.

Relative efficiency was used as a gauge for whether one estimator was better than the other. It was found that a ratio estimator using the number of mill employees as an auxiliary variable was an improvement over SRS. In regard to relative efficiency, neither method of stratification, the DH method or the cluster analysis method, was better than the other. However, the number of groups used in the stratifications lead to increased efficiency when the number of groups was increased. Also, using a combined ratio estimator was a more efficient choice than just a stratified mean (or total).

Six bound sizes (1, 5, 10, 15, 20, and 25 percent) were considered for deriving samples sizes for the total volume of roundwood from Georgia mills' receipts. The minimum achievable bound size was found to be 10 percent of the total for the DH-method using a two group stratification. This was true for both the stratified and combined ratio estimators. In addition, for the stratified and combined ratio estimators, only the DH method stratifications were able to reach a 10 percent bound on the total. All 14 estimators were able to achieve a 20 percent bound of the total.

Finally, nonlinear repeated measures models were developed to spatially allocate mill receipts to surrounding counties in the event of obtaining only a mill's total receipt volume. A Gompertz model with a power spatial covariance was found to be the best performing when using road distances from the mills to either county center type (geographic or forest mass). These models utilized the cumulative frequency of mill receipts as the response variable, with cumulative frequencies based on distance from the mill to the county.

7.2 CONCLUSION

While the multiple imputation method did not show a statistical difference between the reference means, it did have benefits not reflected in that result. Using multiple imputation for nonresponse helps to adjust for data that is perhaps missing at random and not missing completely at random. This is of benefit to avoiding potential bias in the totals arrived at for a state. The method is broad in that it accepts point estimates from a variety of sources and is therefore widely usable in situations of nonresponse. Future TPO studies could provide further evidence for its applicability.

Sampling also can be a very useful tool for determining a state's means and totals. Several of the estimators had sample sizes of less than 40 mills, with 40 being less than a quarter of the population of 170 mills. At one quarter of the population size, greater effort can either allow for reaching 100 percent response in the sample or have so few nonrespondents that multiple imputation is easily applied and readily accepted.

The spatial allocation models selected had mean residual values for the cumulative frequency of the 2003 receipts of between -0.016 and 0.006 for the Gompertz-Road distance-Geographic Center model and between -0.018 and 0.004 for the Gompertz-Road Distance-Forest Mass Center model (~95 percent confidence interval). Both models therefore predicted the new observations with near equal accuracy. In future TPO studies, if the sampled mills can at least supply their employee numbers and total receipts, models can be developed which allocate those receipts to surrounding counties.

Other research building on these methods should focus on models that better select counties that have actual withdrawals. A drawback to the spatial allocation models is that while they predict well for counties with known removals, including counties with zero harvest creates a prediction where none should occur. Therefore, models need to be developed to determine when a county has an actual harvest. These models could be built with logistic regression models, categorical and regression trees (CART), or perhaps discriminant analysis.

REFERENCES

Alderman, D. and Luppold, W. 2005. Examination of regional hardwood roundwood markets in West Virginia. For. Prod. J. 55(12):153-157.

Allison, P. D. 2002. Missing Data. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136. Thousand Oaks, CA: Sage. 93pp.

Anderson, N. and Germain, R. 2007. Variation and trends in sawmill wood procurement in the Northeastern United States. For. Prod. J. 57(10):36-44

Bechtold, W.A. and Patterson, P. L. (Eds.) 2005. The Enhanced Forest Inventory and Analysis Program—National Sampling Design and estimation Procedures. USDA Forest Service General Technical Report SRS-80. Asheville, NC. 85 pp.

Blyth, J. E., McGuire, D. H., and Smith, W. B. 1987. Indiana timber industry--an assessment of timber product output and use.. Res Bull. NC-102. St. Paul, MN: U. S. Dept. of Ag., Forest Service, North Central Forest Experiment Station. 34 pp.

Cooper, M. C. and Milligan, G. W. 1988. The effect of measurement error on determining the number of clusters in cluster analysis *in* Proceedings of the International Workshop on Data Analysis, Decision Support, and Expert Knowledge Representation in Marketing and Related Areas of Research. p319-328.

Dalenius, T. and Hodges, J. L. 1959. Minimum variance stratification. J. Amer. Stat. Assoc. 54:88-101.

Davis, J. C. 2002. Statistics and Data Analysis in Geology. New York, NY:John Wiley and Sons. 638 pp.

Gennings, C., Chinchilli, V. M., and Carter, W. H. 1989. Response surface analysis with correleated data: a nonlinear model approach. Journal of the American Statistical Association 84:805-809.

Johnson, T. G. 1994. Georgia's timber industry—an assessment of timber product output and use, 1992. Res Bull. SE-144. Asheville, NC: U.S. Dept. of Ag., Forest Service, Southeastern Forest Experiment Station. 32 pp.

Johnson, T. G., Jenkins, A., Wells, J. L. 1997. Georgia's timber industry—an assessment of timber product output and use, 1995. Res Bull. SRS-14. Asheville, NC: U.S. Dept. of Ag., Forest Service, Southeastern Forest Experiment Station. 37 pp.

Johnson, T. G., and Wells, J. L. 1999. Georgia's timber industry—an assessment of timber product output and use, 1997. Res Bull. SRS-38. Asheville, NC: U.S. Dept. of Ag., Forest Service, Southeastern Forest Experiment Station. 36 pp.

Johnson, T. G., and Wells, J. L. 2002. Georgia's timber industry—an assessment of timber product output and use, 1999. Res Bull. SRS-68. Asheville, NC: U.S. Dept. of Ag., Forest Service, Southeastern Forest Experiment Station. 40 pp.

Johnson, T. G., and Wells, J. L. 2004. Georgia's timber industry—an assessment of timber product output and use, 2001. Res Bull. SRS-92. Asheville, NC: U.S. Dept. of Ag., Forest Service, Southeastern Forest Experiment Station. 40 pp.

Johnson, T. G. and Wells, J. L. 2005. Georgia's timber industry—an assessment of timber product output and use, 2003. Res Bull. SRS-104. Asheville, NC: U.S. Dept. of Ag., Forest Service, Southeastern Forest Experiment Station. 46 pp.

Johnson, T. G. and Stratton, D. P. 1998. Historical trends of timber product output in the South. Res Bull. SRS-33. Asheville, NC: U.S. Dept. of Ag., Forest Service, Southeastern Forest Experiment Station. 56 pp.

Jongman, R. H. G., C. J. F. Ter Braak, and Van Tongeren O. F. R.. 1995. Data analysis in community and landscape ecology. New York, NY: Cambridge University Press. 299 pp.

Keegan III, C. E.; Chase, A. L.; Morgan, T. A.; Bodmer, S. E.; Van Hooser, D. D.; Mortimer, M.. 2001. Arizona's forest products industry: A descriptive analysis 1998. 20 pp.

Larson, R. W., Spad, B. 1963. Georgia's timber. Res. Bull. SE-1. Asheville, NC: U.S. Dept. of Ag., Forest Service, Southeastern Forest Experiment Station. 39 pp.

Little, F. J. and Rubin, D. B. 1987. Statistical analysis with missing data. New York, NY: John Wiley and Sons 278 pp.

Little, F. J. 2004. Multiple Imputation for nonresponse in surveys. Hoboken, NJ: John Wiley and Sons. 287 pp.

Little, R.C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. 1996. SAS® System for Mixed Models. Cary, NC:SAS Institute Inc. 633 pp.

McCormick, J. F., Cruikshank, J. F. 1954. Forest statistics for Georgia, 1951-1953. For. Sur. Release 44. Asheville, NC: U.S. Dept. of Ag., Forest Service, Southeastern Forest Experiment Station. 77 pp.

McGarigal, K., Cushman, S. and Stafford, S. 2000. Multivariate statistics for wildlife and ecology research. New York, New York: Springer-Verlag. 283 pp.

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

McRoberts, R. E. 2003. Compensating for missing plot observations in forest inventory estimation. Can. J. For. Res. 33(10):1990-1997.

Milligan, G. W. and Cooper, M. C. 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50:159-179.

Reams, G. A. and McCollum, J. M. 2000. The use of multiple imputation in the southern annual forest inventory system *in* Hansen, M.; Burk, T., eds. Integrated tools for natural resources inventories in the 21st century: an international conference on the inventory and monitoring of forested ecosystems; 1998 August 16-19; Boise, ID. Gen. Tech. Rep. NCRS-212. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station: 228-233.

Rubin, D. B. 1987. Multiple imputation for nonresponse in surveys. John Wiley and Sons, Inc. New York. 288 pp.

Sarle, W. 1983. Cubic Clustering Criterion. SAS® Technical Report A-108, SAS® Institute, Inc., Cary, NC.

Schaeffer, R.L., Mendenhall, W, and Ott, R. L. 2006. Elementary Survey Sampling 6th edition. Belmont, CA:Thomson Brooks/Cole. 464 pp.

Scheuren, F. 2005. Multiple imputation: How it began and continues. The American Statistician 59(4):315-319.

Schwab, O., Bull, G., and Maness, T.. 2005. A mill-specific roundwood demand equation for southern and central Finland. J. of For. Econ. 11:95-106.

Spillers, A. R. 1943. Georgia forest resources and industries. Misc Publ. 501 Washington, D.C.: U.S. Government Printing Office, 70 pp.

Tabachnick, B. G. and Fidell, L. S. 2001. Using Multivariate Statistics 4th edition. Boston, MA: Allyn and Bacon. 966 pp.

Tansey, J. B. and Steppleton, C. D. 1991. Georgia's timber industry—an assessment of timber product output and use, 1989. Res Bull. SE-126. Asheville, NC: U.S. Dept. of Ag., Forest Service, Southeastern Forest Experiment Station. 23 pp.

Van Deusen, P. C. 1997. Annual forest inventory statistical concepts with emphasis on multiple imputation. Can. J. For. Res. 27(3):379-384.

Wagner, J. E., Smalley, B., and Luppold, W. 2004. Factors affecting merchandising of hardwood logs in the southern tier of New York. For. Prod. J. 54(11):98-102.

Welch, R. L.; Bellamy, T. R. 1976. Changes in output of industrial timber products in Georgia, 1971-1974. Res. Bull. SE-36. Asheville, NC: U.S. Dept. of Ag., Forest Service, Southeastern Forest Experiment Station. 28 pp.

Zeger, S. L., Liang, K.Y., and Albert, P. S. 1988. Models for longitudinal data: a generalized estimating equation approach. Biometrics 44:1049-1060.

USDA. 2004. USDA, The Forest Inventory and Analysis (FIA) Database: Database Description and Users Guide Version 1.7. Forest Inventory and Analysis Program, United States Department of Agriculture, Newtown Square, PA (2004).

APPENDIX A

Timber Removals Survey
Georgia Forestry Commission and
Forest Service, U.S. Department of Agriculture
Southern Research Station
Phone Number: (828) 257-4848; Fax: (828) 257-4894
200 WT Weaver Boulevard
Asheville, NC 28804

Do you wish to be included in a regional and statewide directory?
Yes _____
No _____

OMB 0596-0010
Expires:12/31/2009

LOGS AND OTHER ROUNDWOOD RECEIVED, 2003

Georgia

Name of Company _____ No. Employees[1]: _____

Local Mail Address: _____
(P. O. Box, Street or Route) (City) (State) (Zip)

Person Contacted & Title: _____ Phone: _____

E—Mail: _____ Fax: _____

Web Address: _____

Sales Contact: _____ Phone: _____

Company CEO: _____ Phone: _____

County: _____ Latitude: _____ Longitude: _____

Plant Location: _____

Type Plant: _____ Products[2]: _____

_____

Instructions: This form is for reporting the quantities and types of logs received by this mill in 2003 and the disposal of plant residues resulting from the manufacturing or processing of wood products.

PLEASE COMPLETE A SEPARATE FORM FOR EACH TYPE PLANT OR OPERATION. If records are not available, please give your best estimates.

This survey is voluntary. While you are not required to respond, your cooperation is needed to make the results of the survey comprehensive, accurate, and timely.

ALL VOLUMES REPORTED WILL BE HELD CONFIDENTIAL AND WILL ONLY BE USED TO AGGREGATE TO THE COUNTY AND STATE LEVEL. Other information may be used to compile "Regional/Statewide Industry" directory.

If no logs (any length) were received in 2003. Please check box below. No other information is needed.

☐ **No logs were received in 2003. Please indicate under remarks (page 3) if mill is closed or** operated part time.

Mill Status: Active ☐ Out-of-business ☐
Idle ☐ New ☐

[1] Excluding contract loggers.
[2] Be specific. List products such as rough or dressed lumber, pallet stock, bark, shavings, etc. Begin with the most important product, and list in order of priority.

Official Use Only:
Interviewers Name: _____ Title: _____
Telephone Number: _____ Date: _____
Information obtained by: _____ mail _____ phone _____ personal contact _____ email

Section I—Logs/Roundwood received in 2003 for processing at this plant, by Type.

1. Total quantity of logs received at this mill in 2003. (Check appropriate type. Include all logs processed, whether cut by own crews, purchased, or processed on a custom basis.)

NOTE: COMPLETE A SEPARATE FORM FOR EACH TYPE OF PLANT.

Check the type of Raw Materials Received	Total Quantity Received for Product Indicated			
	Amount[1]	Unit of Measure[2]	Average Length[3]	Average Top Diameter[3]
☐ — Saw logs				
☐ — Veneer & plywood logs or bolts				
☐ — Logs for composite board				
☐ — Poles and piling				
☐ — Post				
☐ — Pulpwood				
☐ — Other (specify) _____				
☐ — Other (specify) _____				

[1] Include wood from foreign countries. If logs are received in more than one unit of measure, indicate the volume received in each unit.
[2] Specify unit of measure such as thousand board feet, standard cords (128 cubic feet), pieces, linear feet, tons (2000 lbs.), etc.
[3] Specify average length and average top diameter for all units of measure except cords.

2. If board–feet were used in Part 1 above, specify which log rule was used:

(Check one for each species group)

Soft-woods	Hard-woods	
☐	☐	Doyle
☐	☐	Scribner Decimal C.
☐	☐	Lumber Tally
☐	☐	International ¼-inch rule
☐	☐	Other (specify) _____

3. If weights or cords were used in Part 1 above, please specify the appropriate conversion below:

Softwoods	Hardwoods	
_____	_____	Pounds per MBF
_____	_____	Pounds per cord

Section II—Volume of product produced from logs received in 2003.

(Check one for each species group)

Unit of Measure						
Soft-woods	Hard-woods			Amount	Percent Dressed	(If sawn product)
☐	☐	Lumber tally (board feet)				
☐	☐	Square feet	Softwood	_____	_____	%
☐	☐	Pieces				
☐	☐	Other (specify) _____	Hardwood	_____	_____	%

Do you export? _____

Public reporting burden for this collection of information is estimated to average 50 minutes per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Agriculture, Clearance Officer, OIRM, Room 404-W. Washington, DC 20250; and to the Office of Management and Budget, Paperwork Reduction Project (OMB# 0596-0010), Washington, DC 20503.

Section III— Receipts of logs, by species group and origin, 2003.

A.	B.													
		ter county name at top of columns C-N and percent received from each county located in-state, out-of-state or foreign county if outside of U												
	Total volume	C	D	E	F	G	H	I	J	K	L	M	N	Total
Species group	or percent													(C-N)
Yellow pine														100%
E. white pine														100%
Eastern red cedar														100%
Cypress														100%
Other:														100%
Yellow poplar														100%
Sweetgum														100%
Soft maple														100%
Black/tupelo gum														100%
Other:														100%
Red oak														100%
White oak														100%
Hickory														100%
Ash														100%
Hard maple														100%
Other:														100%
Total	100%													100%

Section IV—Check type of equipment in use.

1. Debarker	_____	10. Resaw (circular gang)	_____	19.	_____
2. Pole peeler	_____	11. Resaw (band)	_____	20.	_____
3. Chipper	_____	12. Bark grinder	_____	21.	_____
4. Chip canter	_____	13. Bark hog	_____	22.	_____
5. Chip–n–saw	_____	14. Firewood processor	_____		
6. Veneer (lathe or slice)	_____	15. Dry kiln	_____		
7. Headsaw (circular)	_____	16. Planer	_____		
8. Headsaw (band)	_____	17. Treating cylinder	_____		
9. Scragg saw	_____	18. Wood fired boiler	_____		

Remarks:

Section V—Disposal of MILL RESIDUES in 2003, by type and use.

Instructions: Please enter your best estimate of the percentage **OF EACH TYPE MILL RESIDUE** that was used for the various purposes indicated. These estimates should **not** include the disposal of chipped logs from Section VI.

Percent Use of Mill Residues Only (exclude chipped logs or bolts)								
	Bark		COARSE RESIDUES (chips, slabs, edgings, trims, cores, etc.)		FINE RESIDUES			
					Planer or Lathe Shavings		Sawdust	
1. Use of Residues	Sfwd. %	Hdwd. %	Sfwd. %	Hdwd. %	Sfwd. %	Hdwd. %	Sfwd. %	Hdwd. %
a. Manufacture of fiber products (pulp, roofing felt, hardboard, medium density fiberboard)								
b. Particle board/other fiber products								
c. Charcoal or chemical wood								
d. Small dimension and other sawn products								
e. Industrial fuel at this or any other mill and domestic household fuel								
f. Mulch								
g. Miscellanous uses specify:_____								
h. Not used (including residues burned as waste)								
TOTAL	100%	100%	100%	100%	100%	100%	100%	100%

SectionVI—Material sold for fiber to the pulp industry in 2003, from the processing of logs listed in Section I.

Note: If you have a percent in Section V row(a) under coarse residues there should be a volume amount in total quantity sold (no. 1 below).

Do not include wood sold for fuel. List only material sold for fiber products.

		Softwood	Hardwood
1. Total quantity sold -	Cords	_____	_____
	or		
	Tons	_____	_____
2. Of the total sold, what percentage was from:			
a. Residue (chips)		_____ %	_____ %
b. Shavings		_____ %	_____ %
c. Sawdust		_____ %	_____ %
d. Whole logs or short sections chipped at plant		_____ %	_____ %
		100%	100%

APPENDIX B

Definition of terms (from Johnson 2004, p12.)

Exports-The volume of domestic roundwood utilized by mills outside the state where timber is cut.

Imports-The volume of domestic roundwood delivered to a mill or group of mills in a specific state but harvested outside that state.

Industrial roundwood products-Any primary use of the main stem of a tree, such as saw logs, pulpwood, veneer logs, intended to be processed into primary wood products such as lumber, wood pulp, sheathing, at primary wood-using mills.

Primary wood-using plants-Industries that convert roundwood products (saw logs, veneer logs, pulpwood, etc.) into primary wood products, such as lumber, veneer or sheething, wood pulp.

Production-The total volume of known roundwood harvested from land within a state, regardless of where it is consumed. Production is the sum of timber harvested and used within a state, and all roundwood exported to other states.

Receipts-The quantity or volume of industrial roundwood received at a mill or by a group of mills in a state, regardless of the geographic source. Volume of roundwood receipts is equal to the volume of roundwood retained in a state plus roundwood imported from other states.

Roundwood-Logs, bolts or other round sections cut from trees for industrial manufacture or consumer uses.

Timber products output-The total volume of roundwood products from all sources plus the volume of byproducts recovered from mill residues.