For-estimation : Post-stratification to decrease variance of forest attribute estimates in the Interior West

---

Math 343 Final Project

Edwin Alvarado, Alex Lloyd-Damnjanovic, Miranda Rintoul, Mai Toyohara

May 3rd, 2019

Acknowledgements

Please be sure to thank the stakeholders in the acknowledgements!

Table of Contents

List of Tables

List of Figures

Abstract

The National Forest Inventory and Analysis (FIA) Program of the US Forest Service regularly estimates forest attributes such as biomass and trees per acre. These estimates are used in a wide variety of applications, such as policy formulation, scientific analysis, land management, and business plan development. The use of post-stratification in prediction of forest attributes can greatly improve the efficiency of these estimates. Current procedures used by the Interior West unit of the FIA involve stratifying by forest and non-forest areas. This project aims to increase the efficiency of estimates by finding a better stratification scheme. We plan to propose different stratification schemes based on satellite image data, and compare their variance estimates to that of the current stratification scheme. This research seeks to improve on the previous FIA stratification scheme by reducing the amount of variation in forest attribute estimation, which can in turn make for more useful and informative analyses. We have found that two new schemes, based on vegetation type and tree canopy cover, could be potential candidates for a high efficiency replacement of the current forest-nonforest scheme.

The National Forest Inventory and Analysis (FIA) Program of the US Forest Service regularly estimates forest attributes such as biomass and trees per acre. These estimates are used in a wide variety of applications, such as policy formulation, scientific analysis, land management, and business plan development. The use of post-stratification in prediction of forest attributes can greatly improve the efficiency of these estimates. Current procedures used by the Interior West unit of the FIA involve stratifying by forest and non-forest areas. This project aims to increase the efficiency of estimates by finding a better stratification scheme. We plan to propose different stratification schemes based on satellite image data, and compare their variance estimates to that of the current stratification scheme. This research seeks to improve on the previous FIA stratification scheme by reducing the amount of variation in forest attribute estimation, which can in turn make for more useful and informative analyses. We have found that two new schemes, based on vegetation type and tree canopy cover, could be potential candidates for a high efficiency replacement of the current forest-nonforest scheme.

Chapter 1

Introduction

The National Forest Inventory and Analysis (FIA) Program of the U.S. Forest Service is responsible for the monitoring of forest ecosystem attributes across the United States for geographic areas ranging from counties, states, and provinces. Due to natural variability between plots, as well as limitations of data collection, the FIA must estimate forest attributes using data gathered from a sample of ground plots, and supplemented using auxiliary data that is available at a finer resolution. The estimates of these attributes, such as biomass or trees per acre, are used in a number of applications. For example, information from the FIA is used by lawmakers to assess the viability of environmental policies, by land management bureaus for land use planning, and by scientists as a basis for ecological investigations. The FIA program is crucial in that it is the only program to provide consistent annual forestry data across the entire United States. The Interior West (IW) unit of the FIA is responsible for tracking forest attributes in Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming.

The FIA currently uses a three-phase system for collecting forestry data. The first phase involves using aerial photography or satellite imagery to create large-scale maps of different forest attributes and identify areas that are forest or non-forest. The next phase requires manually sampling a number of previously established field plots distributed across different states and counties. The final phase involves sampling from a subset of the plots from the second phase.

The FIA has access to two types of data to create forest attribue estimates. First, several ground plots are randomly sampled with an intensity of approximately one plot per 6,000 acres (McConville, Moisen, & Frescino, n.d.). On each of these plots, a number of forest attributes are measured by hand. The other type of data used by the FIA is pixel-level data of auxiliary predictor variables. These two levels of data are used to create state-wide estimates of forest attributes.

1.1 Plot sampling

The FIA has selected field plots in a way that produces an equal probability random sample representative of the region (McRoberts, Holden, Nelson, Liknes, & Gormanson,

2005). A map of the United States was tiled with hexagons, forming a tessallation that completely covers the region. Each FIA plot was randomly sampled from one of these hexagons, which are, approximately 2,400 hectares (~5,900 acres) in area (McRoberts et al., 2005). This hexagonal area was determined by FIA standards stipulating that standard error in estimates cannot exceed 3% per million acres of forest land area or 5% per billion cubic feet of growing stock (McRoberts and Hansen, 1999). The hexagons are subdivisions of an existing hexagonal tessellation created for the Forest Health Monitoring (FHM) program (McRobert and Hansen, 1999). A plot was sampled from each hexagon using the following method:

1. If an existing FHM or FIA plot lay within the hexagon, it was assigned as the plot for that hexagon.
2. If the hexagon contained no FHM plot, the existing permanent FIA plot closest to the center of the hexagon was assigned.
3. If there was no existing FHM or FIA plot in the hexagon, a new permanent FIA plot was established in the center of the hexagon and assigned as that hexagon's plot.

For hexagons overlapping multiple states, the plot was chosen to be in the portion of the hexagon lying in the largest state. Equal probability is assumed by the random orientation of the FHM hexagonal tessellation and the absence of a relationship between the locations of the hexagons and the locations of the existing permanent FIA plots.

Each FIA field plot consists 4 circular subplots of radius 48 feet. One subplot is the "center"; the other subplots are spaced equally (at angles of 120 degrees apart) at a distance of 120 feet from the center. Each plot effectively covers an area of 1 acre. Data from the subplots is aggregated to the center subplot, which is the designated location of the field plot (McRoberts et al., 2005).

1.2 Post-stratification

To improve the quality of forest attribute estimates, the FIA uses a statistical technique known as post-stratification to decrease variability. The process of post-statratification begins by grouping similar sample observations together into a small number of categories, called *strata*. Next, estimates are calculated separately for each stratum. Finally, these estimates are weighted based on the number of observations within the corresponding stratum, then added together to create a final post-stratified estimate. The estimate produced by this technique is unbiased, and has lower variability than a simple random sampling estimate, which is computed by simply taking the mean across all observations.

These groupings, called *stratification schemes*, are chosen such that within-group observations are as similar as possible and between-group observations are as dissimilar as possible. The high homogeneity within groups means that observations within groups will have similar attributes, which means that post-stratified estimates will be more precise.

The FIA-IW currently uses a stratification scheme that categorizes plots as either forest or non-forest using remotely sensed data by the Moderate Resolution Imaging Spectroradiometer (MODIS) (J. Blackard et al., 2008). However, the FIA has access to other sets of auxiliary data, such as LANDFIRE's existing vegetation type data and forest type group data. Previous studies have shown that using this auxiliary data to create an alternative stratification scheme could improve the efficiency of different estimations (McConville et al., n.d.).

In this paper, we will explore a number of stratification schemes that could be alternatives to the FIA's current forest-nonforest scheme. Because the post-stratified estimator is unbiased, our overall goal will be to lower the variance, or improve the efficiency, of this estimator. There are a number of important factors we will need to take into account when creating new schemes. We will need to devise informed schemes that have more within-stratum homogeneity and more between-stratum heterogeneity. Furthermore, previous research has also shown that stratification schemes that have low within-strata sample sizes have negatively biased estimates of variance (Westfall, Patterson, & Coulston, 2011). Given this, the stratification schemes we choose to study should have strata that each have at least 10 observations within them.

It is important to note that the FIA-IW's analysis and estimation of forest attributes is done at the pixel level. In the case of post-stratification, this means that stratum weights are calculated using the proportion of pixels on a map that fall within each stratum. We do not yet have access to full pixel data, so our analysis will be done at the plot level, i.e. stratum weights will be calculated using plot proportions rather than pixel proportions. This means that the estimates in this paper will have higher variance than those that would be derived from the same stratification schemes analyzed with pixel data.

The rest of the paper is organized as follows. Chapter 2 provides a description of key variables and data sources. Chapter 3 motivates strata and describes post-stratification methods. Chapter 4 presents results. Chapter 5 discusses results and concludes.

Chapter 2

Data

The data used in this analysis is a collection of national databases used by the FIA for forest inventory purposes. The data comes from a number of sources, including the Moderate Resolution Imaging Spectroradiometer (MODIS), the Landscape Fire and Resource Management Planning Tools (LANDFIRE) program, and the National Land Cover Database (NLCD). These various datasets are used together to create in-depth inventories and informative analyses on national forests. This chapter provides a description of the response variables, the predictor variables as well as a description of corresponding summary statistics, and transformed predictor variables.

2.1 Responses

The field plot locations used by the FIA are assumed to produce a random sample (McRoberts et al., 2005). Plots are spatially distributed through a hexagonal grid, with the location and orientation of the grid randomly selected, and plots within each hexagon assumed to be randomly distributed with respect to the grid. Various forest attributes are measured by hand at these field plot locations. We will use four of these forest attributes as response variables in our analysis: basal area, trees per acre, above-ground biomass, and sawlog volume. These variables are all attributes of live trees and are measured at the plot level.

2.2 Predictors

In this analysis, the predictor variables come from geospatial maps on a national level, which allows us to perform prediction on both plot and pixel levels. The predictors we use are Cleland ecological subregion, LANDFIRE existing vegetation, forest type group, forest probability, and forest biomass. In addition, we make use of transformed versions of these predictor variables in existing vegetation type bins (eight and four bins), mountain region, and forest group bins.

Additional predictor layers were used, including predictors from the Moderate Resolution Imaging Spectro-radiometer (MODIS) and NLCD92 datasets. Any data that were not at a 250 meter pixel resolution were rescaled to this resolution. The

data were split into 65 ecologically distinct mapping zones, which were labeled forest/nonforest using a classification tree model at the plot level. Following this, a regression tree algorithm was used to predict biomass in areas labeled forest, using the same predictors as the forest/nonforest model. This was extrapolated to the pixel level using the MODIS and NLCD92 layers, and another dataset of per-pixel uncertainty.

Current strata

The IW unit of the FIA currently uses a stratification scheme that divides the entire region into forest and nonforest categories, based on a 250m resolution map created by MODIS (J. Blackard et al., 2008). In the dataset we recieved, there was a separate category for "water" observations, which we decided to collapse into the nonforest category for the purpose of simplifying the creation of new stratification schemes. The majority of plots in the IW are considered to be nonforest.

Ecological subregion

Eco-regions are regions that differ based on various ecological characteristics, such as climate, physiography, and geological substrate, determined by the U.S. Department of Agriculture Forest Service (Cleland et al., 2007). These regions can be further separated into provinces, which are either montainous or non-mountainous.

The Interior West (IW) is made up of 14 separate eco-provinces, half of which are mountainous and half of which are not mountainous. Because there were so many unique eco-provinces, we decided to focus on the mountain/non-mountain aspects of the variable for the purposes of creating strata. The majority of the plots in our dataset are not mountainous, with approximately 68% of plots belonging to a non-mountainous eco-region.

Vegetation type

The Landscape Fire and Resource Management Planning Tools (LANDFIRE) program provides nationally consistent, landscape-scale geospatial layers, databases, and ecological models including a layer for existing vegetation type (EVT) (Nelson, McRoberts, Liknes, & Holden, 2002). The EVT layer categorizations represents the dominant vegetation type using NatureServe's Ecological systems vegetation classification for natural vegetation (Nelson et al., 2016). The EVT layer is mapped using a combination of different geospatial map data and stratified by landscape and species composition. These specific geological descriptions can be grouped into broad classifications of vegetation type within the NatureServe classification scheme. We will use these broad vegetation types in our analysis instead of the more specific classifications.

Figure 2.1 depicts the vegetation type distribution over FIA plots. There are eight vegetation type categories: Agriculture, Barren, Developed, Herb, Shrub, Sparse, Tree, and Water. Of these types, Herb, Shrub, and Tree are all used to describe different types of forest foliage. These three types are by far the most common in the dataset, with about 38% of plots having Shrub vegetation type.

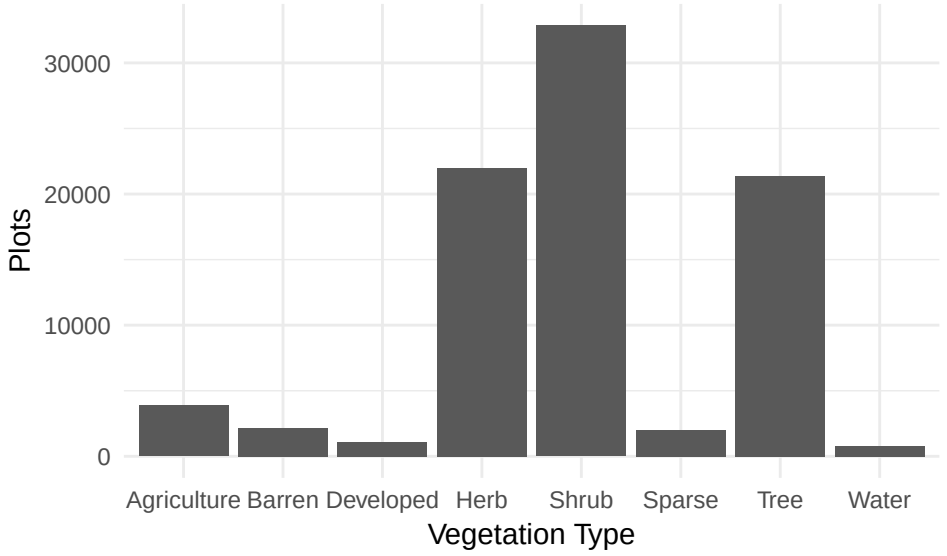


Figure 2.1: Vegetation type distribution over FIA plots in the IW region.

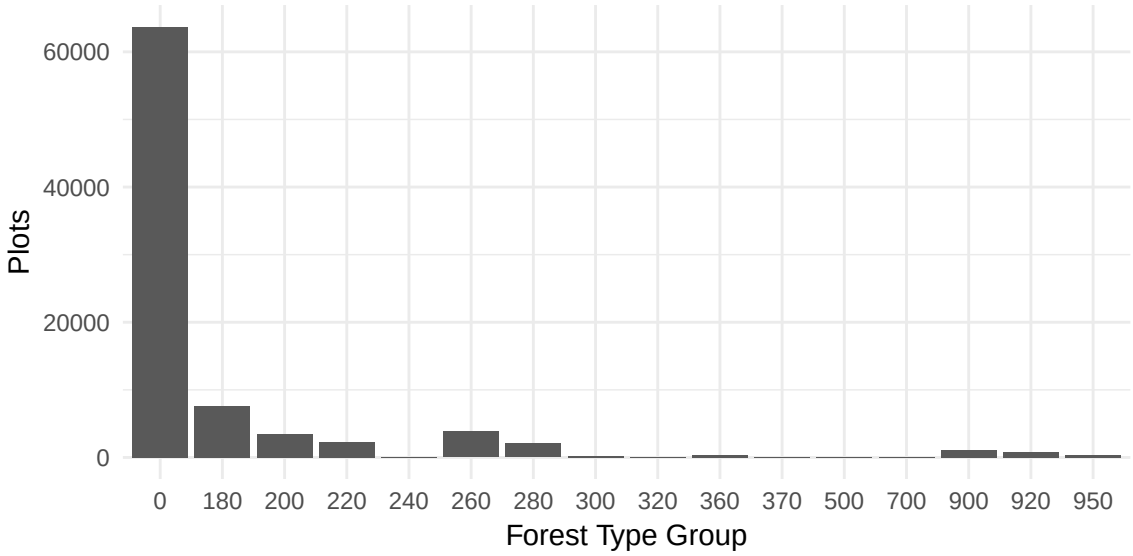


Figure 2.2: Forest type group distribution over FIA plots in the IW region. 0 refers to nonforest, while 180 refers to the Pinyon/Juniper group.

Forest type group

The FIA uses a forest type classification scheme defined by Eyer (1980) and the USFS Renewable Resources Evaluation Group (Ruefenacht et al., 2008). U.S. forests are divided into 142 specific forest type categories, which are then grouped into 28 forest type groups. The national forest type group map used by the FIA was created by intersecting 250-meter resolution geospatial data with FIA plot data, and using a decision-tree model to assign a forest type to each plot.

Figure 2.2 depicts the distribution of forest type groups. Of the 28 forest type groups, 15 of them appear in our dataset of IW plots, with the Pinyon/Juniper group being the most common. Other forest type groups are present but extremely rare, having fewer than 10 observations. The non-forest category is far more common than any of the forest types, however, with about 74% of plots being classified in this way.

Forest probability

Forest probability was obtained from the forest/non-forest classification, along with the per-pixel uncertainty measures (J. Blackard et al., 2008). The forest/non-forest predictions generated by the classification tree have at each branch a confidence value, which was used to determine forest probability. Non-forest observations have a range of forest probability from 0–0.5, while forest observations have a range of 0.5–1.0. Forest probability data has a resolution of 250 meters.

Figure 2.3 describes the distribution of forest probability. The majority of plots in the IW region are non-forest, and thus, have forest probability below 0.5. Among this group, a large number of plots have forest probability 0. Likewise, among the forested plots, a large proportion have forest probability 1.0, though this effect is not as striking as with the non-forest plots. This results in an uneven "bowl" shaped distribution of forest probability.

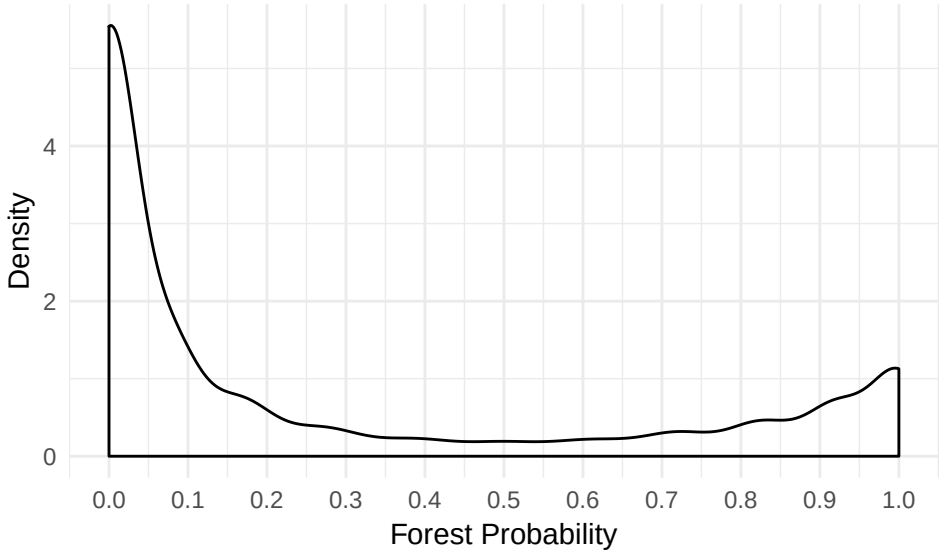


Figure 2.3: Density plot of IW forest probability.

Forest biomass

Forest biomass is defined to be all above-ground, live biomass, measured in Mg/ha. The FIA uses a spatial biomass dataset created using MODIS imagery and NLCD data (J. Blackard et al., 2008). This dataset, which has biomass estimates for all national regions, has a resolution of 250 meters.

Figure 2.4 shows the distribution of the biomass across FIA plots. In the IW region, most plots are non-forest and thus have biomass 0. The data is heavily right-skewed, but there is also a small peak at 10 Mg/ha. Though the maximum forest biomass is 118 Mg/ha, less than 2% of plots have biomass greater than 60 Mg/ha.

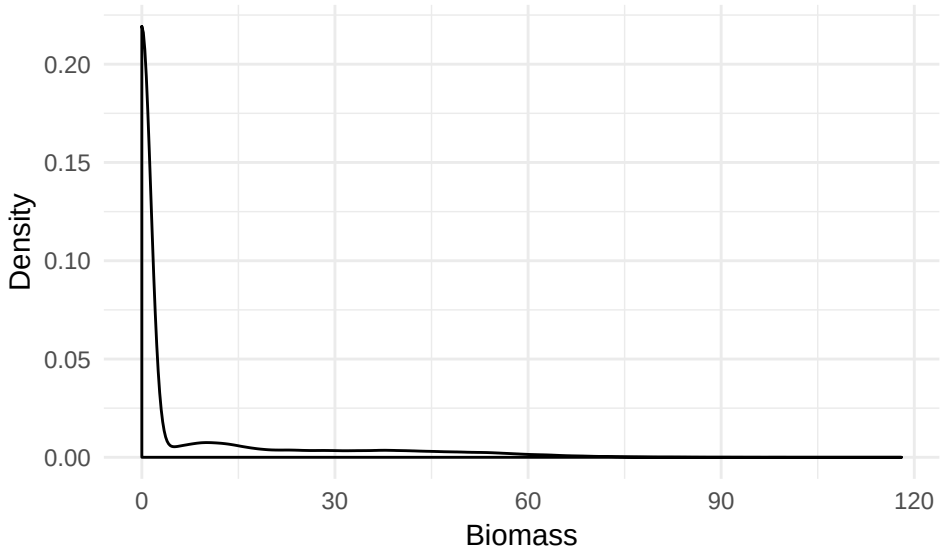


Figure 2.4: Density plot of IW biomass.

Tree canopy cover

Tree canopy cover data comes from the National Land Cover Database (NLCD) canopy cover map (Homer et al., 2015). The NLCD utilizes imagery data from the National Agriculture Imagery Program (NAIP), Landsat 5 Thematic Mapper (and its derivatives), as well as previous NLCD data. NLCD data is available at every pixel of a 30-by-30 meter grid. Percent tree canopy cover was modeled using a random forest model for each NLCD mapping zone.

The distribution of tree canopy cover is depicted in figure 2.5. Similar to forest biomass, IW tree canopy cover data is heavily right-skewed as over half of all plots have 0% canopy cover.

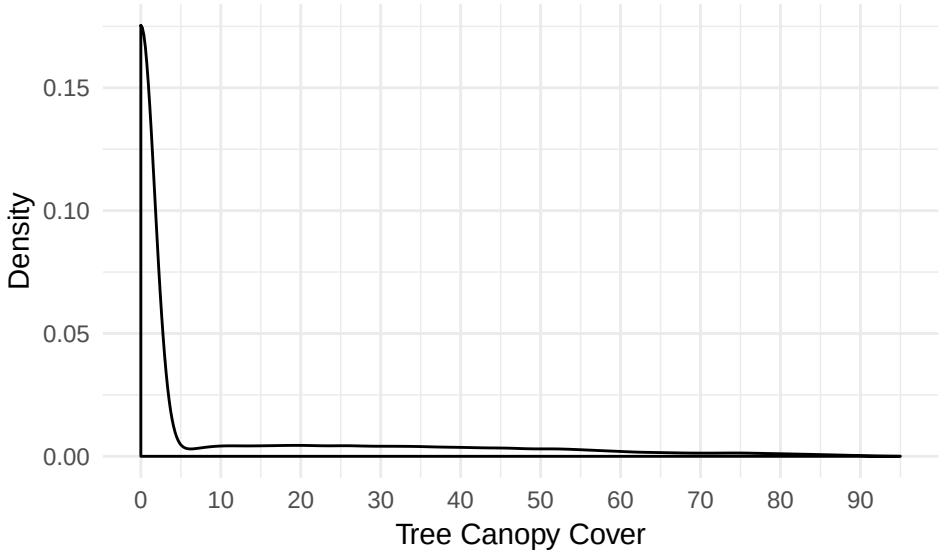


Figure 2.5: Density plot of IW tree canopy cover.

Transformed Predictors

We made adjustments to certain predictors in order to simplify our analysis. First, in order to create stratification schemes, we binned the three numeric predictors: tree canopy cover, biomass, and forest probability. Percent tree canopy cover was split into groups 0–5, 6–50, 51–65, and 66–100, according to the Northern station of the FIA. Biomass was split into groups 0–6, 7–20, 21–75, and 76–118 based on a visual analysis of the overall biomass density. Then forest probability, which is a continuous variable, was split along the cutpoints .07, .27, and .57, again based on overall density.

We also decided to bin certain categorical variables to reduce the overall number of categories and complexity of the analysis. We binned the forest type group variable to have eight total categories: non-forest, the six most common forest types, and a separate category for all other forest types. We also created bins for LANDFIRE's vegetation type by grouping all categories except Tree, Shrub, and Herb into a general "non-forest" category.

Chapter 3

Methods

3.1 Poststratification estimates

The goal of this work is to estimate the measure of our given response variables, in the case of our smallest area, a county, and to improve the estimate of its variance. In order to evaluate the performance of the post stratification estimations of population mean and variance estimate, we compare them to the simple random sampling case, in which no strata are used. In the case of simple random sampling, the estimation of the mean of the response variable is essentially the average of the population, and can be determined by the following formula, which is aggregated at the county level:

$$\bar{Y_s} = \frac{1}{n} \sum_{i \epsilon s}^{s} y_i$$

Where the variables signify:

- $\bar{Y}$ - estimation of the mean of the response variable for the sampled plots in a county
- n - the number of plots
- s - set of indices for sampled plots in a county
- y_i - response variable of the i-th plot

Similarly, the variance estimator in the simple random sampling case can be determined by the following:

$$v(\bar{Y}) = \frac{\sum_i^n (Y_i - \bar{Y})^2}{n(n-1)}$$

These equations amount to a simple average of all the subplots of each subsample area, such as a county or state, and the variance of that simple average. This can be improved upon by grouping plots into different strata and finding estimations of variance within the strata before aggregating to the larger subsample, i.e. county or state. These groupings improve the efficiency of estimates of mean, or reduce variance estimates. When the groupings of plots are more similar, the estimate of the responses

will become more similar, and the variance will become lower. An approximation of the mean of a population area in the post-stratification case is adapted from Equation 4.13 in Bechtold and Patterson's (2015) stratification framework. As our smallest unit of interest is at the county level, we estimate the mean of the response variable at the county level using the following (Bechtold & Patterson, 2015):

$$\bar{Y_{ps}} = \sum_{h}^{H} \frac{n_h}{n} (\frac{\sum_{i}^{n_h} y_{h_i}}{n_h})$$

Where:

- $\bar{Y_{ps}}$ - estimation of the mean of the response variable with post-stratification
- h - stratum h where $h = 1, 2, \ldots, H$
- n - number of plots in a given county
- n_h - number of plots in a given county in strata h
- N - number of pixels in a given county
- N_h - number of pixels in a given county in strata h
- y_{h_i} - ith-response variable for stratum h

To determine the estimations of the variances of the post-stratified estimator, we begin by evaluating the performance of each stratification scheme at the county level. Given the independent nature of each county, the variance of a larger sub-sample, such as at the state level, is additive. An approximation of the variance of the mean estimate given a population area in the post-stratification case is adapted from Equation 4.6 in the stratification framework given by Bechtold and Patterson (2015) for a county here (Bechtold & Patterson, 2015):

$$v(\bar{Y}) = \frac{1}{n} \left[\sum_{h}^{H} \frac{n_h}{n} v(\bar{Y_h}) + \sum_{h}^{H} (1 - \frac{n_h}{n}) \frac{n_h}{n} v(\bar{Y_h}) \right]$$

Where the variables previously mentioned are the same and the rest signifiy:

- $v(\bar{Y})$ - the estimate of the variance of the mean of the response variable
- H - the total number of strata in a given stratification scheme

Here the variances are calculated at the county level and we aggregate these variances to determine performance for a larger sub-sample, such as state.

3.2 Stratification scheme motivations

In order to come up with new stratification schemes, we had three main lines of exploration: we looked at ways of refining the old FIA scheme, we investigated relationships between other predictors in the dataset, and we looked at single predictors as potential stratification schemes. Our first step was to explore the data to try and see if we could find any significant features in the distributions of predictors. A good stratification scheme should exhibit a certain degree of homogeneity *within* categories, but heterogeneity *between* categories. Ideally, having many categories per strata would more precisely cluster similar observation units together, but due to complexity concerns we had to keep the number of categories under 10. When there are many categories per strata, estimation may be more efficient, but we risk the problem of overfitting. We also risk the problem of ... This meant that our continuous predictors– biomass, forest probability, and tree canopy cover– had to be binned. For tree canopy cover, we used the FIA Northern station's binning scheme: the tree canopy variable was partitioned into the intervals 0–5%, 6–50%, 51–65%, and 66–100%.[1] In order to bin biomass and forest probability, we created density plots of each variable, and then looked for natural cutpoints in the plots. Similarly, the vegetation type and forest type group predictors had to be binned to reduce the number of categories. To determine the appropriate bins, we created histograms of each predictor to see which categories contained significantly different proportions of plots.

This visual approach guided much of our exploratory data analysis. To identify potentially useful interactions, we plotted the distributions of predictors after splitting by another predictor. To illustrate our process, consider the interaction between the the old stratification scheme and the mountain variable. It seemed most natural place to start looking at interactions was with the mountain variable, since it has the fewest categories. Additionally, mountains tend to be heavily forested, and so we expected that adding a mountain/nonmountain classification would refine the old strata. We created two different versions of this interaction; one that was simply old strata crossed with mountain, and one that was the old strata crossed with mountain, but with all nonmountain categories grouped together. We also considered the interactions between the old strata and the vegetation type and tree canopy variables.

For our second strategy, we began by considering which combinations of variables would result in a manageable (< 10) number of categories. Since the mountain variable has only two categories, we began by looking at interactions between mountain and other variables. For example, consider the interaction between our binned forest probability variable and the mountain variable (Scheme 15). Intuitively, we expected that a split of the nonmountain category into the different forest probability bins would not be helpful, as there probably are not many nonmountain plots with high forest probability. So in this case, we split only the mountain category by forest probability, but grouped all nonmountain observations together, regardless of forest

[1] Although our research area of interest is the Interior West, where tree canopy cover is significantly different from Northern station observations, due to time constraints, we chose not to partition using a "Interior West" partitioning scheme. Further stratification efforts using tree canopy cover should investigate an Interior West-specific binning scheme.

probability. We also expected there to be a strong positive relationship between forest probability and tree canopy cover, based on Figure 3.1:

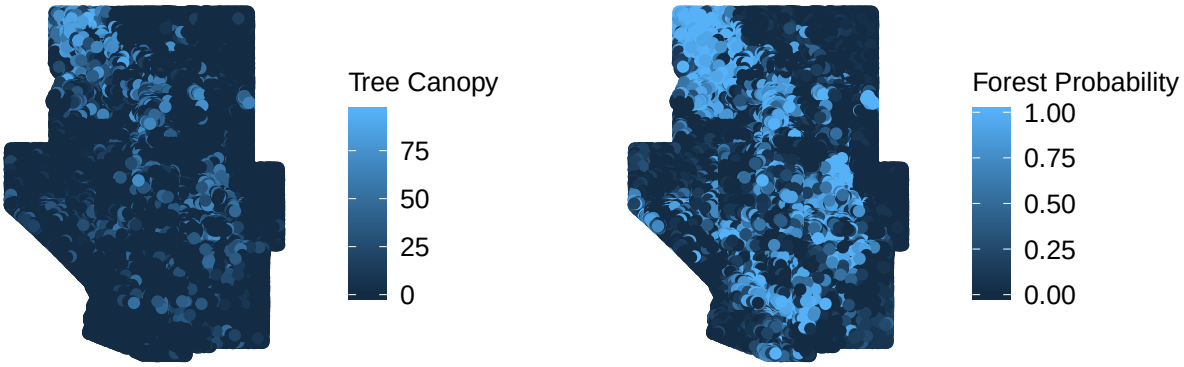


Figure 3.1: Maps comparing the distributions of forest probability and tree canopy cover across the Interior West.

This relationship motivated us to also look at a tree canopy-mountain split, with nonmountain as a separate category again (Scheme 10). We then decided to look at interactions between mountain and the vegetation type and forest group variables, as these are pretty different from the variables listed above.

We decided that tree canopy, biomass, forest probability, and mountain were good proxies for whether a plot was forested or not, and so we treated these as stratification schemes on their own (Schemes 5, 6, 13, and 1). We also decided that making forest group and vegetation type strata could be useful (Schemes 4 and 7), as these would try to classify a plot as forest or nonforest not by looking at the types of lifeforms present. This is a different approach, as we are not looking at forests as a whole (like tree canopy cover and biomass do), but rather, looking at if the types of plants usually present in forest environments are present in a plot. These are the considerations that motivated our single-variable stratifications.

3.3 Stratification schemes

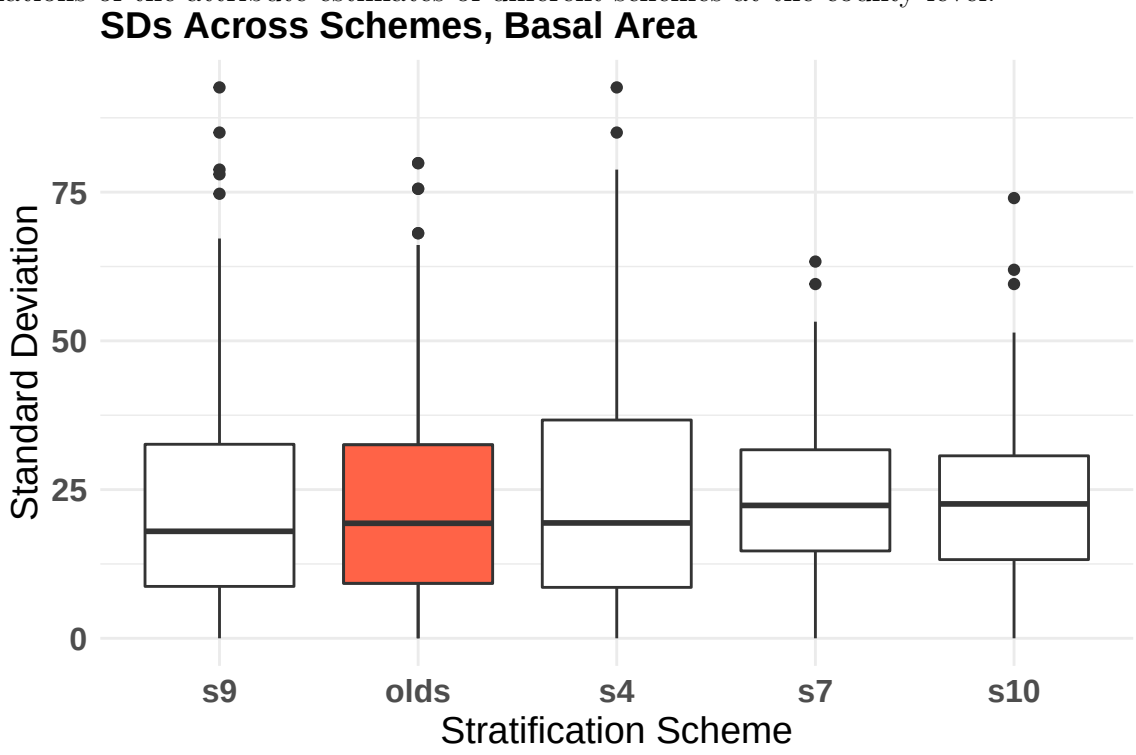
We devised 15 new stratification schemes to compare with the forest-nonforest scheme currently used by the FIA-IW. They each use different interactions of the predictor variables to categorize the plots into different strata. For example, the third scheme, which has three strata, is based on an interaction between the current strata variable and the Ecoregion mountain variable. One strata in that scheme contains all the plots that are nonmountainous and forested, the other has all the plots that are mountainous and forested, and the final strata contains all the plots that are nonforest. Table 3.1 describes the variables used and the number of strata found in each scheme. For a full list of all tested stratification schemes, see Appendix Table B.1.

Table 3.1: Overview of stratification schemes.

Scheme	Variables	Strata
1	Mountain	2
2	Mountain, current strata	3
3	Mountain, current strata	4
4	Veg type	8
5	Biomass	4
6	Tree canopy	4
7	Forest group bins	8
8	Veg type bins	4
9	Mountain, veg bins	8
10	Mountain, tree canopy	8
11	Current strata, tree canopy	5
12	Current strata, veg bins	8
13	Forest probability	4
14	Mountain, biomass	8
15	Mountain, forest probability	5

3.3.1 Scheme performance evaluation

Evaluating the efficiency of the schemes was a multi-step process. Initially, we used boxplots such as the one in Figure **??** to visually compare the median standard deviations of the attribute estimates of different schemes at the county level.



At first glance, it appears that a few schemes, such as S7, have extremely low variance and are thus extremely efficient. However, this result was driven by the presence of a large number of zero variance (and therefore zero standard devation) mean estimates within the schemes. Zero variance estimates can occur when the forest attribute of interest has mean 0, but most often occur when a county contains only one plot belonging to a certain strata. A single estimate cannot have any variance, so the variance estimate for the entire county under that scheme becomes 0. If there are many such counties for a given scheme, this effect will lower that scheme's overall variance, creating a misleading impression that it is highly efficient.

To remedy this, we collapsed any strata with less than 10 plots in a county into the county's most popular stratum.[2] For example, assume we have a stratification scheme with 4 strata. Suppose a county has 432 plots in stratum A, 3 plots in stratum B, 34 plots in stratum C, and 0 plots in stratum D. The three plots in stratum B would be collapsed into A, for 435 total plots in A and 0 in B. Strata C and D would remain the same. This process was applied to all the counties in our best stratification schemes.

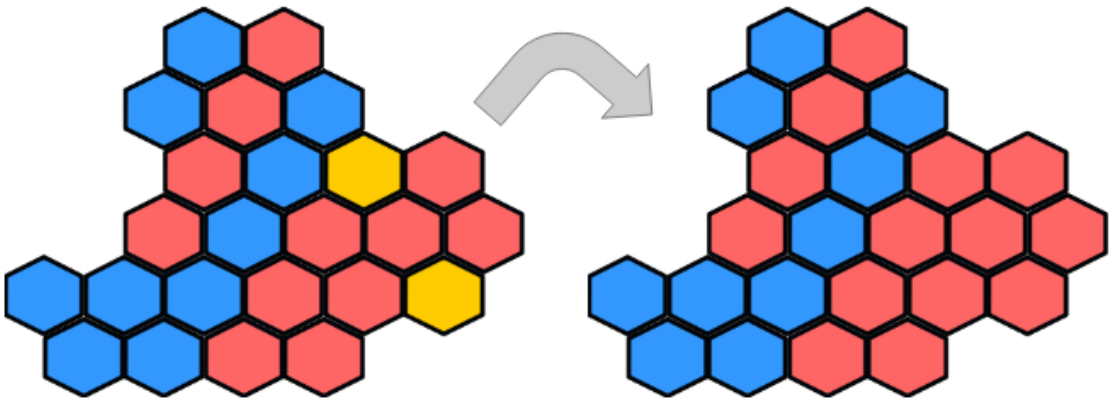


Figure 3.2: Collapsing strata

This effectively raised the variance estimates for each of the schemes, which lowered the estimated efficiency. However, these estimates are a much more accurate representation of the efficiency for each of the new candidate schemes.

To evaluate these collapsed schemes, we derived normalized mean and median standard deviation measures of all of the schemes, which produced an unbiased scheme evaluation metric. We added these scores together for each of the four responses, such that each scheme had a mean and median score from 0 to 4, with 0 being the best score. Then, the two scores were added so that each scheme could be scored on a scale of 0 to 8, with mean and median SD measures having equal weight.

[2]Another way to collapse strata could be to collapse ecologically similar categories together (e.g., plots in the Ponderosa Pine group into the Lodgepole Pine group). However, due to a lack of domain knowledge, we were unable to pursue this strategy. Future work should make use of a domain expert to collapse categories in a ecologically consistent way.

3.3.2 The best stratification schemes

We used the above process to evaluate the schemes on different subsets of the full plot dataset, and found that the best schemes overall were S4, S7, S9, and S10.

S4

S4 is a stratification scheme based solely on LANDFIRE's categorical existing vegetation data. The eight levels of that variable correspond to the eight strata in the scheme:

- Agriculture (4.5%)
- Barren (2.5%)
- Developed (1.3%)
- Herb (25.5%)
- Shrub (38.2%)
- Sparse (2.3%)
- Tree (24.8%)
- Water (.9%)

S7

Scheme S7 is based on the FIA's forest type group variable. The strata are bins of the 15 forest type groups that appear in the IW region, which include forest type groups that have over 1000 observations, a separate "other" category for all other forest type groups, and a "nonforest" category. In total, the plots are divided into eight strata:

- Aspen/Birch (1.3%)
- Douglas Fir (4%)
- Fir/Spruce/Hemlock (4.5%)
- Lodgepole Pine (2.4%)
- Nonforest (74%)
- Other (2.2%)
- Pinyon/Juniper (8.8%)
- Ponderosa Pine (2.7%)

S9

Scheme S9 is based on two variables: the binned version of LANDFIRE's existing vegetation type, and mountain/nonmountain ecological region. There are four levels of the binned vegetation type, and two levels of the moutain variable, which means all plots are divided into eight strata:

- Herb, Mountain (3.9%)
- Herb, Nonmountain (2.2%)
- Other, Mountain (1.8%)
- Other, Nonmountain (9.6%)

- Shrub, Mountain (7.9%)
- Shrub, Nonmountain (30.3%)
- Tree, Mountain (17.6%)
- Tree, Nonmountain (7.3%)

S10

Scheme S10 is based on mountain/nonmountain ecological region and binned percent tree canopy cover. The four tree canopy bins result in the plots being divided among eight strata:

- 0-5% Canopy cover, Mountain (13.9%)
- 0-5% Canopy cover, Nonmountain (62.7%)
- 6-50% Canopy cover, Mountain (11.9%)
- 6-50% Canopy cover, Nonmountain (5.9%)
- 51-65% Canopy cover, Mountain (3%)
- 51-65% Canopy cover, Nonmountain (.2%)
- 66-100% Canopy cover, Mountain (2.4%)
- 66-100% Canopy cover, Nonmountain (.1%)

Chapter 4

Results

This chapter presents results for the current scheme, stratification schemes 4, 7, 9, and 10.

4.1 Basal Area

The stratification scheme with the lowest mean standard deviation of basal area estimate was the current stratification scheme with an average standard deviation of 21.97 (Table 4.1). However, the stratification scheme with the lowest median standard deviation of basal area estimate was S9, with a value of 17.98. When we evaluate the scaled average and median, the current stratification scheme performed the best in terms of mean, whereas S9 performed the best in terms of median.

Table 4.1: Basal Area variance estimates for stratification schemes 4, 7, 9, and 10.

Stratification Scheme	Mean	Median	Scaled Mean	Scaled Median
Current Scheme	21.97	19.34	0.00	0.29
S4- Existing Vegetation	24.06	19.39	1.00	0.31
S7- Forest group bins	23.23	22.34	0.60	0.95
S9- Vegetation bins * mountain	22.68	17.98	0.34	0.00
S10- Tree canopy * mountain	22.20	22.59	0.11	1.00

4.2 Trees per Acre

The stratification scheme with the lowest mean standard deviation of trees per acre estimate was the S10 scheme, with an average standard deviation of 142.26 (Table 4.2). However, the stratification scheme with the lowest median standard deviation of trees per acre estimate was S9, with a value of 88.81. When we evaluate the scaled average and median, the S10 scheme performed the best in terms of mean, and S9 performed the best in terms of median.

Table 4.2: Trees per acre variance estimates for stratification schemes 4, 7, 9, and 10.

Stratification Scheme	Mean	Median	Scaled Mean	Scaled Median
Current Scheme	151.99	98.08	0.64	0.29
S4- Existing Vegetation	157.48	98.18	1.00	0.29
S7- Forest group bins	143.56	120.97	0.09	1.00
S9- Vegetation bins * mountain	150.36	88.81	0.53	0.00
S10- Tree canopy * mountain	142.26	115.79	0.00	0.84

4.3 Aboveground Biomass

The stratification scheme with the lowest mean standard deviation of aboveground biomass estimate was the S10 scheme with an average standard deviation of 7.26 (4.3). However, the stratification scheme with the lowest median standard deviation of biomass estimate was the current stratification scheme, with a value of 5.73. When we evaluate the scaled average and median, the old stratification scheme performed the best in terms of mean, and the S9 scheme performed the best in terms of median.

Table 4.3: Aboveground biomass variance estimation for stratification schemes 4, 7, 9, and 10.

Stratification Scheme	Mean	Median	Scaled Mean	Scaled Median
Current Scheme	7.62	5.06	0.30	0.00
S4- Existing Vegetation	8.46	5.73	1.00	0.59
S7- Forest group bins	7.42	6.19	0.14	1.00
S9- Vegetation bins * mountain	8.00	5.18	0.62	0.11
S10- Tree canopy * mountain	7.26	5.95	0.00	0.79

4.4 Sawlog Volume

The stratification scheme with the lowest mean standard deviation of sawlog volume estimate was the S10 scheme with an average standard deviation of 425.49 (Table 4.4). However, the stratification scheme with the lowest median standard deviation of sawlog volume estimate was the current stratification scheme, with a value of 322.18. When we evaluate the scaled average and median, the S10 scheme performed the best in terms of mean, and the current scheme performed the best in terms of median.

Table 4.4: Sawlog volume variance estimation for stratification schemes 4, 7, 9, and 10.

Stratification Scheme	Mean	Median	Scaled Mean	Scaled Median
Current Scheme	453.81	291.01	0.36	0.00
S4- Existing Vegetation	503.18	322.18	1.00	0.52
S7- Forest group bins	430.64	348.60	0.07	0.96
S9- Vegetation bins * mountain	474.92	292.03	0.64	0.02
S10- Tree canopy * mountain	425.49	351.12	0.00	1.00

4.5 Overall Performance

Performance for stratification schemes vary depending on the measure used. For example, S10 performed best under scaled mean, and S9 performed best using scaled median estimates (Figure 4.1). To consider how scheme performed across all schemes, we added the scores together across response variables to obtain a single mean- and median-based score for each scheme. In Table 4.5, for all schemes, we observe that S10 performed best under mean scaling, with S7, S9, S4, then the current scheme following. Under median scaling, S9 performed best, with the current scheme, S4, S10, then S7 following. If the lowest average standard deviation is desired, a scheme like S10 might be preferred. However, if a low standard deviation is preferred, but we are averse to extreme maximum values, S9 might be preferred.

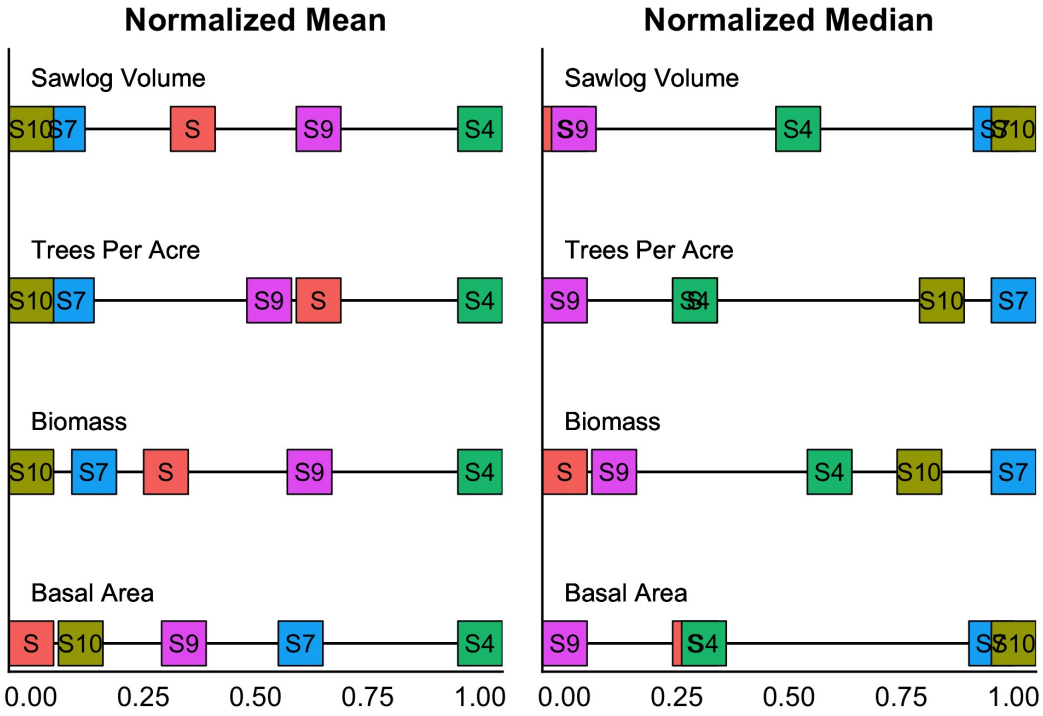


Figure 4.1: Normalized Standard Deviation Scores across schemes.

To account for varying performance across measures, we added the two scores together across all response variables to have a single normalized standard deviation score for each stratification scheme. In considering this measure, the current scheme performed the best, with S9, S10, S7, then S4 following (Table 4.5).

Table 4.5: Aggregated normalized standard deviation scores across schemes.

Stratification Scheme	Mean Sum	Median Sum	Total Sum
Current Scheme	1.30	0.58	1.89
S4- Existing Vegetation	4.00	1.71	5.71
S7- Forest group bins	0.89	3.90	4.79
S9- Vegetation bins * mountain	2.12	0.13	2.25
S10- Tree canopy * mountain	0.11	3.63	3.74

Chapter 5

Discussion

Our experimental results demonstrate that there are viable options for a new stratification scheme that could be used to improve the efficiency of forest attribute estimates. In particular, schemes 9 and 10, based on the variables for existing vegetation type and tree canopy percent respectively, often had the lowest mean and median standard deviation estimates across the four response variables. Scheme 10 had the smallest mean standard deviation for estimates of sawlog volume, aboveground biomass, and trees per acre. Scheme 9 had excellent overall performance under the median standard deviation metric.

Schemes 9 and 10 both use the Cleland Ecoregion mountain variable to stratify observations, but otherwise have little in common. So, it is possible that further exploration of this mountain variable as a tool for stratification, used in conjunction with other predictor variables, could help improve the efficiency of different forest attributes.

However, the FIA-IW's current statratification scheme also had relatively efficient predictive power. That scheme had the smallest mean standard deviation of basal area estimates, and the smallest median standard deviation of sawlog volume and aboveground biomass estimates. The current scheme performed fairly well under both mean and median metrics, while schemes 9 and 10 only performed well under a single efficiency measure.

5.1 Future Research

5.1.1 Pixel level data

Our analysis of the different stratification schemes was done at the plot level. That is, we calculated the post-stratification weights using the proportion of plots that fell into each stratum. This resulted in an overestimate of the variance of the different schemes. A sensible follow-up to this experiment would be to repeat the scheme analysis, but at the pixel level instead. This finer data would allow us to calculate more accurate post-stratification weights, which would give us a better picture of the actual variance of each stratification scheme.

5.1.2 Tree canopy cover

One of the most efficient stratification schemes in our analysis, scheme 10, used percent tree canopy cover as a predictor variable. This variable was binned into groups 0–5, 6–50, 51–65, and 66–100, which is the system used by the Northern station of the FIA. However, the trees and forests of the Interior West are ecologically very different from those in the Northern region. It is possible that a different binning scheme for tree canopy cover, to reflect the IW region, would lead to more efficient prediction. Specifically, most regions in the IW have less than 50% canopy cover, so a better binning scheme might have three or four canopy cover bins between 0 and 50% cover.

5.1.3 Nonresponse

FIA sampling nonresponse occurs when field plot locations cannot be accessed for some reason (Patterson, Coulston, Roesch, Westfall, & Hill, 2012). This is most often caused by hazardous environmental conditions, or public or private landowners denying the FIA access. Nonresponse can affect the accuracy of poststratified estimates if missing samples are not evenly distributed across strata, e.g. if the vast majority of nonsampled plots fall within a single stratum. A nonresponse adjustment factor can be used in poststratified estimates to remove any bias caused by nonresponse (Patterson et al., 2012).

Appendix A

R code

Appendix A includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readibility and/or setup.

In the main Rmd file

```
# This chunk ensures that the thesisdown package is
# installed and loaded. This thesisdown package includes
# the template files for the thesis.
if(!require(devtools))
  install.packages("devtools", repos = "http://cran.rstudio.com")
if(!require(thesisdown))
  devtools::install_github("ismayc/thesisdown")
library(thesisdown)
```

In Chapter 2:

```
stratadf %>%
  ggplot(aes(x = veg_type)) + geom_bar() +
  labs(x = "Vegetation Type", y = "Plots") +
  theme_minimal()
```

```
stratadf %>%
  ggplot(aes(x = forest_type_group)) + geom_bar() +
  labs(x = "Forest Type Group", y = "Plots") +
  theme_minimal()
```

```
stratadf %>%
  ggplot(aes(x = forest_prob)) + geom_density() +
  labs(x = "Forest Probability", y = "Density") +
  theme_minimal() +
  scale_x_continuous(breaks = round(seq(min(stratadf$tree_canopy),
```

```
                                        max(stratadf$tree_canopy),
                                        by = .1),1))
```

```
stratadf %>%
  ggplot(aes(x = biomass)) + geom_density() +
  labs(x = "Biomass", y = "Density") +
  theme_minimal()
```

```
stratadf %>%
  ggplot(aes(x = tree_canopy)) + geom_density() +
  labs(x = "Tree Canopy Cover", y = "Density") +
  theme_minimal() +
  scale_x_continuous(breaks = round(seq(min(stratadf$tree_canopy),
                                        max(stratadf$tree_canopy),
                                        by = 10),1))
```

In Chapter 3:

```
library(cowplot)
p1 <- ggplot(stratadf) +
  geom_point(aes(x = longitude, y = latitude,
                 group = eco_province, color = tree_canopy)) +
  coord_fixed(1.3) +
  labs(x = "Longitude", y = "Latitude", color = "Tree Canopy") +
  theme(axis.text=element_blank(),
        axis.line=element_blank(),
        axis.ticks=element_blank(),
        axis.title=element_blank(),
        legend.title=element_text(size=10),
        legend.text=element_text(size=10))
p2 <- ggplot(stratadf) +
  geom_point(aes(x = longitude, y = latitude,
                 group = eco_province, color = forest_prob)) +
  coord_fixed(1.3) +
  labs(x = "Longitude", y = "Latitude", color = "Forest Probability") +
  theme(axis.text=element_blank(),
        axis.line=element_blank(),
        axis.ticks=element_blank(),
        axis.title=element_blank(),
        legend.title=element_text(size=10),
        legend.text=element_text(size=10))

plot_grid(p1, p2, ncol = 2, align = "hv")
```

```
Scheme <- seq(1, 15, by=1)
Variables <- c("Mountain", "Mountain, current strata", "Mountain,
               current strata", "Veg type", "Biomass", "Tree canopy",
               "Forest group bins", "Veg type bins",
               "Mountain, veg bins", "Mountain, tree canopy",
               "Current strata, tree canopy",
               "Current strata, veg bins", "Forest probability",
               "Mountain, biomass", "Mountain, forest probability")
Strata <- c(2, 3, 4, 8, 4, 4, 8, 4, 8, 8, 5, 8, 4, 8, 5)
schemedf <- as.data.frame(cbind(Scheme, Variables, Strata))

knitr::kable(schemedf,
  caption = "\\label{tab:schemetab} Overview of stratification schemes.",
  format="latex", booktabs = T)  %>%
  kableExtra::kable_styling(latex_options = c("HOLD_position"))
```

```
area_df %>%
  ggplot() + geom_boxplot(aes(x = reorder(scheme, sd, FUN = median),
                              y = sd)) +
  geom_boxplot(data = area_df[area_df$scheme == "olds",],
               aes(x = scheme, y = sd), fill = "tomato") +
  theme_minimal() +
  labs(x = "Stratification Scheme", y = "Standard Deviation",
       title = "SDs Across Schemes, Basal Area") +
  theme(legend.position = "none",
        axis.text.x = element_text(face = "bold", size = 12),
        axis.text.y = element_text(face = "bold", size = 12),
        axis.title.x = element_text(size = 14),
        axis.title.y = element_text(size = 14),
        plot.title = element_text(face = "bold", size = 14))
```

In Chapter 4:

```
#Set up, loading packages
library(tidyverse)
library(kableExtra)
library(ggrepel)
library(gridExtra)
library(grid)

area_df <- readRDS("figure/area_df.rds")
biomass_df <- readRDS("figure/biomass_df.rds")
tpa_df <- readRDS("figure/tpa_df.rds")
sawlog_df <- readRDS("figure/sawlog_df.rds")
```

```
# Summary Tables
# Basal Area
summary_area <- area_df %>%
  group_by(scheme) %>%
  summarize(mean_sd = mean(sd),
            median_sd = median(sd)) %>%
  ungroup() %>%
  mutate(scale_mean =
           (mean_sd - min(mean_sd))/(max(mean_sd)- min(mean_sd)),
         scale_median =
           (median_sd - min(median_sd))/(max(median_sd)-
                                           min(median_sd)))

#tpa
summary_tpa <- tpa_df %>%
  group_by(scheme) %>%
  summarize(mean_sd = mean(sd),
            median_sd = median(sd)) %>%
  ungroup() %>%
  mutate(scale_mean =
           (mean_sd - min(mean_sd))/(max(mean_sd)- min(mean_sd)),
         scale_median =
           (median_sd - min(median_sd))/(max(median_sd)-
                                           min(median_sd)))

#sawlog
summary_sawlog <- sawlog_df %>%
  group_by(scheme) %>%
  summarize(mean_sd = mean(sd),
            median_sd = median(sd)) %>%
  ungroup() %>%
  mutate(scale_mean =
           (mean_sd - min(mean_sd))/(max(mean_sd)- min(mean_sd)),
         scale_median =
           (median_sd - min(median_sd))/(max(median_sd)-
                                           min(median_sd)))

#biomass
summary_biomass <- biomass_df %>%
  group_by(scheme) %>%
  summarize(mean_sd = mean(sd),
            median_sd = median(sd)) %>%
  ungroup() %>%
  mutate(scale_mean =
```

```
          (mean_sd - min(mean_sd))/(max(mean_sd)-
                                      min(mean_sd)),
        scale_median =
          (median_sd - min(median_sd))/(max(median_sd)-
                                        min(median_sd)))

#Total
total_sum <- rbind(summary_area, summary_biomass, summary_sawlog,
                   summary_tpa)

results <- total_sum %>%
  group_by(scheme) %>%
  summarise(sum_mean = sum(scale_mean),
            sum_median = sum(scale_median)) %>%
  ungroup() %>%
  mutate(total_results = sum_mean + sum_median)

#Rename
summary_area <- summary_area %>%
    arrange(match(scheme, c("olds", "s4", "s7", "s9", "s10")))

names <- c("Current Scheme", "S4- Existing Vegetation",
           "S7- Forest group bins", "S9- Vegetation bins * mountain",
           "S10- Tree canopy * mountain")
summary_area$scheme <- names

summary_area <- summary_area %>%
  rename("Stratification Scheme" = scheme,
    "Mean" = mean_sd,
         "Median" = median_sd,
         "Scaled Mean" = scale_mean,
         "Scaled Median" = scale_median)

summary_tpa <- summary_tpa %>%
    arrange(match(scheme, c("olds", "s4", "s7", "s9", "s10")))

summary_tpa$scheme <- names

summary_tpa <- summary_tpa %>%
  rename("Stratification Scheme" = scheme,
    "Mean" = mean_sd,
         "Median" = median_sd,
         "Scaled Mean" = scale_mean,
         "Scaled Median" = scale_median)
```

```
summary_sawlog <- summary_sawlog %>%
    arrange(match(scheme, c("olds", "s4", "s7", "s9", "s10")))

summary_sawlog$scheme <- names

summary_sawlog <- summary_sawlog %>%
  rename("Stratification Scheme" = scheme,
    "Mean" = mean_sd,
         "Median" = median_sd,
         "Scaled Mean" = scale_mean,
         "Scaled Median" = scale_median)

summary_biomass <- summary_biomass %>%
    arrange(match(scheme, c("olds", "s4", "s7", "s9", "s10")))

summary_biomass$scheme <- names

summary_biomass <- summary_biomass %>%
  rename("Stratification Scheme" = scheme,
    "Mean" = mean_sd,
         "Median" = median_sd,
         "Scaled Mean" = scale_mean,
         "Scaled Median" = scale_median)

results <- results %>%
    arrange(match(scheme, c("olds", "s4", "s7", "s9", "s10")))

results$scheme <- names

results <- results %>%
  rename("Stratification Scheme" = scheme,
    "Mean Sum" = sum_mean,
         "Median Sum" = sum_median,
         "Total Sum" = total_results)

#Rounding values
summary_sawlog[,2:5] <- round(summary_sawlog[,2:5], digits = 2)
summary_biomass[,2:5] <- round(summary_biomass[,2:5], digits = 2)
summary_area[,2:5] <- round(summary_area[,2:5], digits = 2)
summary_tpa[,2:5] <- round(summary_tpa[,2:5], digits = 2)
results[,2:4] <- round(results[,2:4], digits = 2)
```

```
#Basal area table
summary_area %>%
  kable(caption =
          "\\label{tab:basaltab} Basal Area variance estimates for stratification
        format= "latex", booktabs = T) %>%
  kableExtra::kable_styling(latex_options = c("HOLD_position"))
```

```
#TPA table
knitr::kable(summary_tpa,
             caption =
               "\\label{tab:tpatab} Trees per acre variance estimates for stratifi
             format= "latex", booktabs = T) %>%
  kableExtra::kable_styling(latex_options = c("HOLD_position"))
```

```
#Biomass table
knitr::kable(summary_biomass,
             caption =
               "\\label{tab:biomasstab} Aboveground biomass variance estimation fo
             format="latex", booktabs = T) %>%
  kableExtra::kable_styling(latex_options = c("HOLD_position"))
```

```
#Sawlog table
knitr::kable(summary_sawlog, caption =
               "\\label{tab:sawlogtab} Sawlog volume variance estimation for strat
             format="latex", booktabs = T) %>%
  kableExtra::kable_styling(latex_options = c("HOLD_position"))
```

```
#Setting up overall results graphic
summary_overall <- summary_area[, 1]
summary_overall <- summary_overall %>%
  rename(scheme = "Stratification Scheme")

names1 <- c("S", "S4", "S7", "S9", "S10")
summary_overall$scheme <- names1

summary_overall <- summary_overall %>%
  mutate(basal_mean = summary_area$`Scaled Mean`,
         basal_placeholder = 1,
         basal_median = summary_area$`Scaled Median`,
         biomass_mean = summary_biomass$`Scaled Mean`,
         biomass_placeholder = 1.5,
         biomass_median = summary_biomass$`Scaled Median`,
         tpa_mean = summary_tpa$`Scaled Mean`,
```

```
          tpa_placeholder = 2,
          tpa_median = summary_tpa$`Scaled Median`,
          sawlog_mean = summary_sawlog$`Scaled Mean`,
          sawlog_placeholder = 2.5,
          sawlog_median = summary_sawlog$`Scaled Median`)


scaled_mean <- summary_overall %>%
  ggplot() + geom_segment(y = 1, yend = 1, x = 0, xend = 1) +
  geom_point(shape = 22, size=10,
             aes(x = basal_mean,
                 y = basal_placeholder,
                 fill = factor(scheme))) +
  geom_text(aes(label = scheme,
                x = basal_mean,
                y = basal_placeholder)) +
  geom_segment(y = 1.5, yend = 1.5, x = 0, xend = 1) +
  geom_point(shape = 22, size=10,
             aes(x = biomass_mean,
                 y = biomass_placeholder,
                 fill = factor(scheme))) +
  geom_text(aes(label = scheme,
                x = biomass_mean,
                y = biomass_placeholder)) +
  geom_segment(y = 2, yend = 2, x = 0, xend = 1) +
  geom_point(shape = 22, size=10,
             aes(x = tpa_mean,
                 y = tpa_placeholder,
                 fill = factor(scheme))) +
  geom_text(aes(label = scheme,
                x = tpa_mean,
                y = tpa_placeholder)) +
  geom_segment(y = 2.5, yend = 2.5, x = 0, xend = 1) +
  geom_point(shape = 22, size=10,
             aes(x = sawlog_mean,
                 y = sawlog_placeholder,
                 fill = factor(scheme))) +
  geom_text(aes(label = scheme,
                x = sawlog_mean,
                y = sawlog_placeholder)) +
  ggtitle("Normalized Mean")

scaled_mean <- scaled_mean +
   theme(legend.title = element_blank(),
```

```
        axis.ticks.y = element_blank(),
        axis.text.y = element_blank(),
        axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        axis.ticks.x = element_blank(),
        axis.line.x = element_line(color = "black",
                                   linetype = "solid"),
        panel.background =
          element_rect(fill = "transparent"),
        plot.background =
          element_rect(fill = "transparent",
                       color = NA),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.background =
          element_rect(fill = "transparent"),
        legend.box.background =
          element_rect(fill = "transparent"),
        legend.position = "none",
        plot.title = element_text(face = "bold")) +
  geom_text(aes(label = "Basal Area",
                x = 0,
                y = 1.15,
                hjust = 0)) +
  geom_text(aes(label = "Biomass",
                x = 0,
                y = 1.65,
                hjust = 0)) +
  geom_text(aes(label = "Trees Per Acre",
                x = 0,
                y = 2.15,
                hjust = 0)) +
  geom_text(aes(label = "Sawlog Volume",
                x = 0,
                y = 2.65,
                hjust = 0))


scaled_median <- summary_overall %>%
  ggplot() +
  geom_segment(y = 1, yend = 1, x = 0, xend = 1) +
  geom_point(shape = 22, size=10,
             aes(x = basal_median,
                 y = basal_placeholder,
```

```
                fill = factor(scheme))) +
  geom_text(aes(label = scheme,
                x = basal_median,
                y = basal_placeholder)) +
  geom_segment(y = 1.5, yend = 1.5, x = 0, xend = 1) +
  geom_point(shape = 22, size=10,
             aes(x = biomass_median,
                 y = biomass_placeholder,
                 fill = factor(scheme))) +
  geom_text(aes(label = scheme,
                x = biomass_median,
                y = biomass_placeholder)) +
  geom_segment(y = 2, yend = 2, x = 0, xend = 1) +
  geom_point(shape = 22, size=10,
             aes(x = tpa_median,
                 y = tpa_placeholder,
                 fill = factor(scheme))) +
  geom_text(aes(label = scheme,
                x = tpa_median,
                y = tpa_placeholder)) +
  geom_segment(y = 2.5, yend = 2.5, x = 0, xend = 1) +
  geom_point(shape = 22, size=10,
             aes(x = sawlog_median,
                 y = sawlog_placeholder,
                 fill = factor(scheme))) +
  geom_text(aes(label = scheme,
                x = sawlog_median,
                y = sawlog_placeholder)) +
  ggtitle("Normalized Median")


scaled_median <- scaled_median +
   theme(legend.title = element_blank(),
        axis.ticks.y = element_blank(),
        axis.text.y = element_blank(),
        axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        axis.ticks.x = element_blank(),
        axis.line.x = element_line(color = "black",
                                   linetype = "solid"),
        panel.background =
          element_rect(fill = "transparent"),
        plot.background =
          element_rect(fill = "transparent",
```

```
                      color = NA),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.background =
          element_rect(fill = "transparent"),
        legend.box.background =
          element_rect(fill = "transparent"),
        legend.position = "none",
        plot.title =
          element_text(face = "bold")) +
  geom_text(aes(label = "Basal Area",
                x = 0, y = 1.15, hjust = 0)) +
  geom_text(aes(label = "Biomass",
                x = 0, y = 1.65, hjust = 0)) +
  geom_text(aes(label = "Trees Per Acre",
                x = 0, y = 2.15, hjust = 0)) +
  geom_text(aes(label = "Sawlog Volume",
                x = 0, y = 2.65, hjust = 0))

norm <- gridExtra::arrangeGrob(scaled_mean, scaled_median, ncol = 2)
ggsave(file="norm.jpeg", norm)
```

```
#Overall results visual
include_graphics("figure/norm.png")
```

```
#Overall Results
knitr::kable(results, caption =
               "\\label{tab:resultstab} Aggregated normalized standard deviation s
             format="latex", booktabs = T) %>%
  kableExtra::kable_styling(latex_options = c("HOLD_position"))
```

Appendix B

Stratification schemes

Table B.1: Descriptions of all tested strata.

Predictor variables used	Number of Strata	Group Description	Naming Scheme
mountain	2	Ecoregion mountain, nonmountain	S1_mountain
old_strata, mountain	3	Ecoregion montain* forest + nonforest	S2_oldstrata_mountain_3
old_strata, mountain	4	Ecoregion montain * old strata forest	S3_oldstrata_mountain_4
veg_type	8	Base 8 vegetation types	S4_veg_type
biomass	4	0-6, 7-20, 21-75, and 76-118	S5_biomass
tree_canopy	4	0-5, 6-50, 51-65, and 66-100	S6_ns_tree_canopy
fgroup_bins	8	binned forest groups	S7_fgroup_bins
veg_bins	4	Four vegetation bins	S8
veg_bins, mountain	8	veg_bins x mountain	S9_veg_bins_mountain
tree_canopy, mountain	8	tree_canopy bins * mountain	S10_tree_canopy_mountain
old_strata, tree_canopy	5	tree canopy bins * forest + nonforest	S11_old_strata_tree_canopy
old_strata, veg_bins	8	FIA forest * veg_bins	S12_old_strata_veg_bins
forest_prob	4	cutpoints: .07, .27, .57	S13_forest_prob
mountain, biomass	8	mountain * biomass	S14_mountain_biomass
mountain, forest_prob	5	mountain * forest_prob + nonmountain	S15_forest_prob_mountain
old_strata	2		old_strata

References

Bechtold, W. A., & Patterson, P. L. (2015). *The Enhanced Forest Inventory and Analysis Program — National Sampling Design and Estimation Procedures* (No. SRS-GTR-80). Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. Retrieved from `https://www.fs.usda.gov/treesearch/pubs/20371`

Blackard, J., Finco, M., Helmer, E., Holden, G., Hoppus, M., Jacobs, D., ... Riemann, R. (2008). Mapping U.S. forest biomass using nationwide forest inventory data and moderate resolution information. *Remote Sensing of Environment*, *112*(4), 1658–1677. `http://doi.org/10.1016/j.rse.2007.08.021`

Cleland, D., Freeouf, J., Keys, J., Nowacki, G., Carpenter, C., & McNab, W. (2007). Ecological subregions: Sections and subsections for the conterminous united states. *General Technical Report WO-76d*, *76*.

Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Coulston, J., ... Megown, K. (2015). Completion of the 2011 National Land Cover Database for the Conterminous United States – Representing a Decade of Land Cover Change Information. *PHOTOGRAMMETRIC ENGINEERING*, 10.

McConville, K. S., Moisen, G. G., & Frescino, T. S. (n.d.). A tutorial in model-assisted estimation with application to forest inventory, 31.

McRoberts, R. E., Holden, G. R., Nelson, M. D., Liknes, G. C., & Gormanson, D. D. (2005). Using satellite imagery as ancillary data for increasing the precision of estimates for the Forest Inventory and Analysis program of the USDA Forest Service. *Canadian Journal of Forest Research*, *35*(12), 2968–2980. `http://doi.org/10.1139/x05-222`

Nelson, M. D., McRoberts, R. E., Liknes, G. C., & Holden, G. R. (2002). Comparing Forest/Nonforest Classifications of Landsat TM Imagery for Stratifying FIA Estimates of Forest Land Area. *Proceedings of the Fourth Annual Forest Inventory and Analysis Symposium*, 8.

Patterson, P. L., Coulston, J. W., Roesch, F. A., Westfall, J. A., & Hill, A. D. (2012). A primer of nonresponse in the US Forest Inventory and Analysis program. *Environmental Monitoring and Assessment. 184(3): 1423-1433.*, 1423–1433. Retrieved

from https://www.fs.usda.gov/treesearch/pubs/40200

Ruefenacht, B., Finco, M., Nelson, M., Czaplewski, R., Helmer, E., Blackard, J., ... Winterberger, K. (2008). Conterminous U.S. and Alaska Forest Type Mapping Using Forest Inventory and Analysis Data. *Photogrammetric Engineering & Remote Sensing*, *74*(11), 1379–1388. http://doi.org/10.14358/PERS.74.11.1379

Service, U. F. (2016). Interior West Forest Inventory and Analysis. United States Department of Agriculture.

Westfall, J. A., Patterson, P. L., & Coulston, J. W. (2011). Post-stratified estimation: Within-strata and total sample size recommendations. *Canadian Journal of Forest Research*, *41*(5), 1130–1139. http://doi.org/10.1139/x11-031