A Hierarchical Bayesian Approach to Small Area Estimation of Forest Attributes

---

A Thesis
Presented to
The Division of Mathematics and Natural Sciences
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Grayson White

May 2021

Approved for the Division
(Mathematics)

_____

Kelly McConville

Acknowledgements

This is where acknowledgements will go.

Preface

This is an example of a thesis setup to use the reed thesis document class (for LaTeX) and the R bookdown package, in general.

Table of Contents

List of Tables

List of Figures

Abstract

The preface pretty much says it all.

Second paragraph of abstract starts here.

Dedication

To my family.

Chapter 1

Context

National forest inventories such as the United States Forest Inventory and Analysis Program (FIA) monitor the status of a nation's forests by collecting data and estimating forest attributes such as basal area, above-ground biomass, tree count per acre, and net volume. Due to the sheer amount of forests in the United States, the FIA cannot collect the population data for these variables across the United States. Instead, they use a sampling design intended and well-suited for estimation over large geographic regions such as states. This sampling design works very well for estimation in large regions and maintains a reasonable cost of employing foresters to collect the samples. While this method works sufficiently for large areas, it has become an interest of national forest inventories such as the FIA to be able to provide reliable and efficient estimates of forest attributes in small domains such as ecological subsections (often referred to as eco-subsections) and counties. In particular, the FIA would like to have estimates with low variance in eco-subsections, however, the FIA only samples a small numbers of plots in these small areas. Currently, the FIA's standard approach to this problem is by using post-stratification. Post-stratification uses a weighted average of the forest attribute of interest and corrects for over- or under-sampling of forested land in the small area. While this estimator is unbiased and has lower variance than the mean, it does not reduce the variance enough for precise estimation at the eco-subsection level. The research goal of this thesis is to address this problem by using techniques which seek to minimize estimate variance while only introducing a small amount of bias. Having precise estimates of forest attributes at the eco-subsection level is crucial for educational programs and implementation of programs which seek to maintain the health of our forests.

In order to produce these estimates we must perform small area estimation. Small area estimation is a branch of survey statistics which includes techniques that allow us to estimate the value of parameters at a sub-population level. Typically in survey estimation, we are interested in doing inference at a population level, however we are sometimes interested in attaining estimates for sub-populations or "small areas." We can visualize the process by considering an ecological province divided into three eco-subsections, each of which have sampled areas:

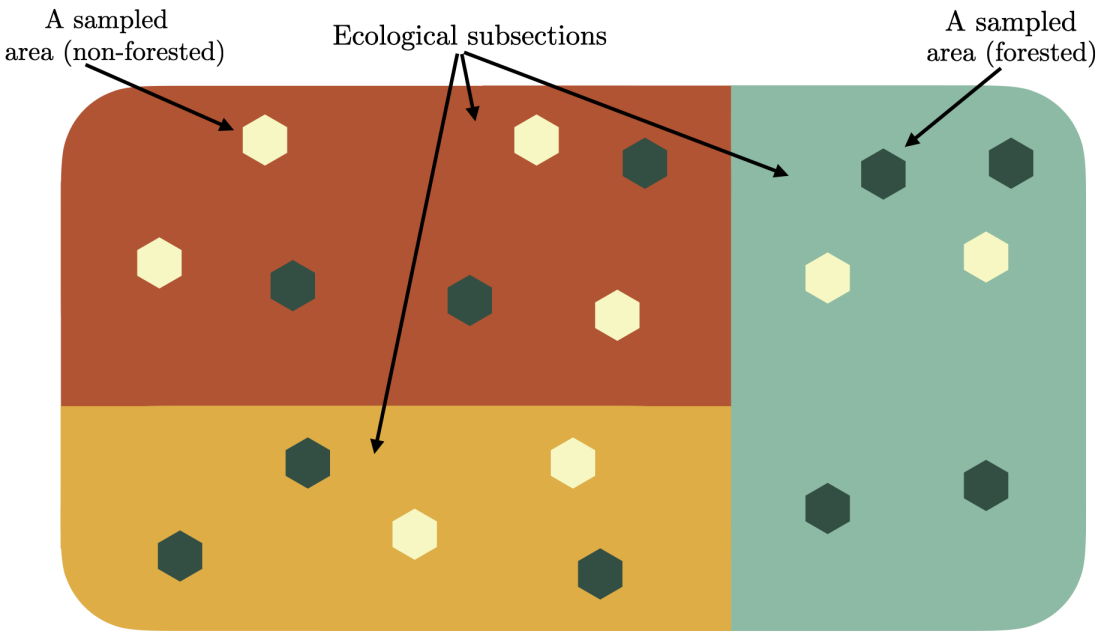

Figure 1.1: An ecological province containing three eco-subsections. The red, yellow, and seafoam areas represent eco-subsections. The green hexagons represent forested areas sampled by the FIA. The beige hexagons represent non-forested areas sampled by the FIA.

We are interested in performing inference at the sub-population level, and in Figure 1.1 these sub-populations are represented by the red, yellow, and seafoam green areas. Importantly, we want to attain estimates of forest attributes in each of these sub-populations. There are wide range of techniques that can be used to carry out this small area estimation. Broadly, these methods fall into three categories: direct estimators, indirect estimators with implicit models, and indirect estimators with explicit models. We will often refer to indirect models with explicit models as "model-based estimators." Each of these methods attempts to do inference at the sub-population level, however, they are quite different from each other.

Direct estimators are defined as those that rely only on the samples within the small area which we would like to measure. Some examples of a direct estimator are the sample mean or the post-stratified estimator. The post-stratified estimator is similar to the mean, however it accounts for under- and over-sampling of forested areas in a given sub-population. These estimators do not rely on information outside of the small area being estimated, however, the post-stratified estimator uses auxiliary information such as the true proportion of forested land within the small area to produce estimates. Direct estimation is the simplest kind of small area estimator as it only relies on samples within the sub-population of interest to produce its estimates. We can visualize these two estimators to get a better sense of their estimation process by considering how they would estimate some forest attribute y in our seafoam green sub-population j from Figure 1.1.

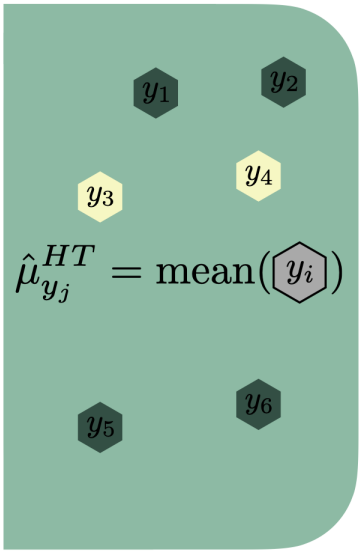

Figure 1.2: The mean as a direct estimator in the seafoam green eco-subsection. This estimator only relies on sampled areas (hexagons) within the eco-subsection. This estimator does not take into account whether the sampled areas are forested (green) or non-forested (beige) areas. This estimator does not use a model to produce estimates and hence uses the y variable collected by FIA foresters to produce the needed estimate, as shown inside the hexagons.

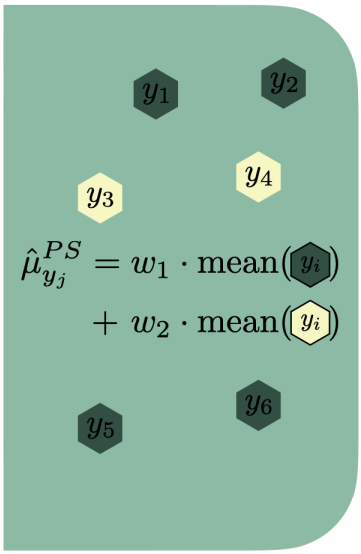

Figure 1.3: The post-stratified direct estimator in the seafoam green eco-subsection. This estimator only relies on sampled areas (hexagons) within the eco-subsection. This estimator takes into account whether the sampled areas are forested (green) or non-forested (beige) areas. It then weights our estimate based on the true population's proportion of forested area in the eco-subsection. This estimator does not use a model to produce estimates and hence uses the y variable collected by FIA foresters to produce the needed estimate, as shown inside the hexagons.

The second kind of estimator, indirect estimators with implicit models, rely on data outside of the area of interest to produce their estimate and can rely on auxiliary data, but implement a model implicitly. With implicit, model-based, indirect estimators we can use information (or "borrow strength") from nearby small areas to help improve our estimate (i.e. reduce variance) in our area of interest through implicit use of a model. These indirect estimators are quite a bit more complicated than direct estimators due to the fact that they borrow strength from the variable of interest across small areas, however, they often significantly reduce variance in estimates due to the added information from other sub-populations. According to Rao (2014), while indirect estimators with implicit models reduce variance, they are often design biased due to the inability to specify between-area variation. This is a large cost of implementing implicit, model-based, indirect estimators. Further, these estimators do not reduce variance as significantly as explicit model-based estimators. Thus, we will not implement indirect estimators with implicit models in this thesis.

Finally, explicit model-based estimators are those which both borrow strength from other small areas, use auxiliary information, and explicitly use a model to compute the estimate of interest. These estimators are still within the family of indirect

estimators. However they make explicit use of a model. Similarly to the indirect estimators with implicit models discussed previously, these models can further reduce the variance of our estimates because they allow for more information to be used in the estimate. We can further categorize these explicit model-based estimators into two classes: unit-level and area-level models. Unit-level models consider information at the level of which the data was collected. Area-level models consider information that has been aggregated to the level of a small area before the model is fit to the data. Commonly, the empirical best linear unbiased prediction model (EBLUP) is used in small area estimation as the model-based estimator of choice. This is because both the area- and unit-level EBLUP models reduce variance further than direct and indirect estimators with implicit models, and they are design unbiased given the modeling assumptions are met. This thesis primarily investigates the usefulness of the hierarchical Bayesian unit-level model (HBU) and hierarchical Bayesian area-level model (HBA). We can visualize HBU and HBA estimators to give a better sense of the differences between the two.

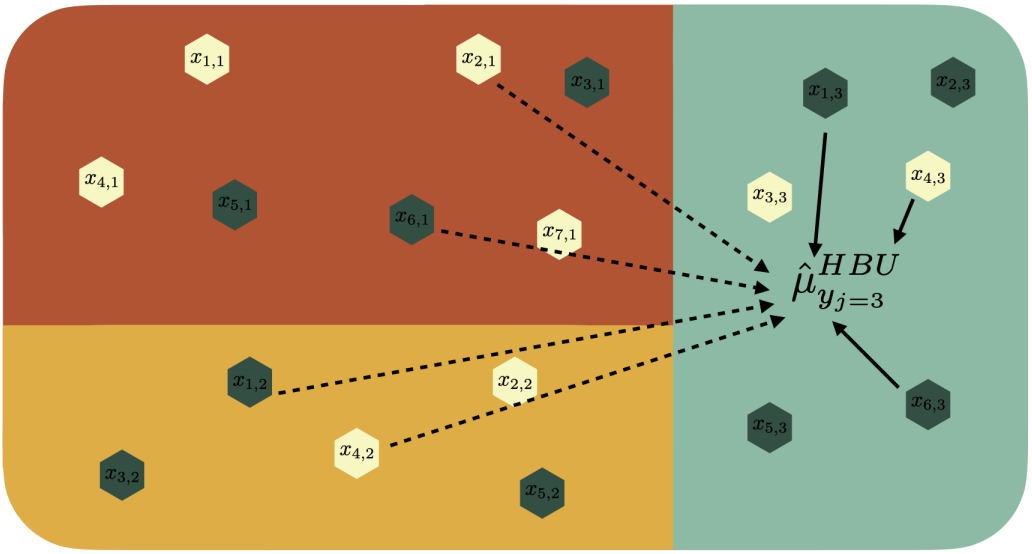

Figure 1.4: The unit-level hierarchical Bayesian model producing an estimate in our seafoam green eco-subsection. This estimator relies on auxiliary information from all eco-subsections within a given eco-province to produce estimates as shown by the arrows. Notably, more information is used in the eco-subsection of interest to produce the estimate, denoted by the solid arrows. The dashed arrows tell us that less information is being used from outside eco-subsections. This estimator also produces estimates based on remotely sensed data, as denoted by the x variables in each hexagon. This estimator uses information at the unit-level, meaning that it produces estimates from the plot level of granularity.

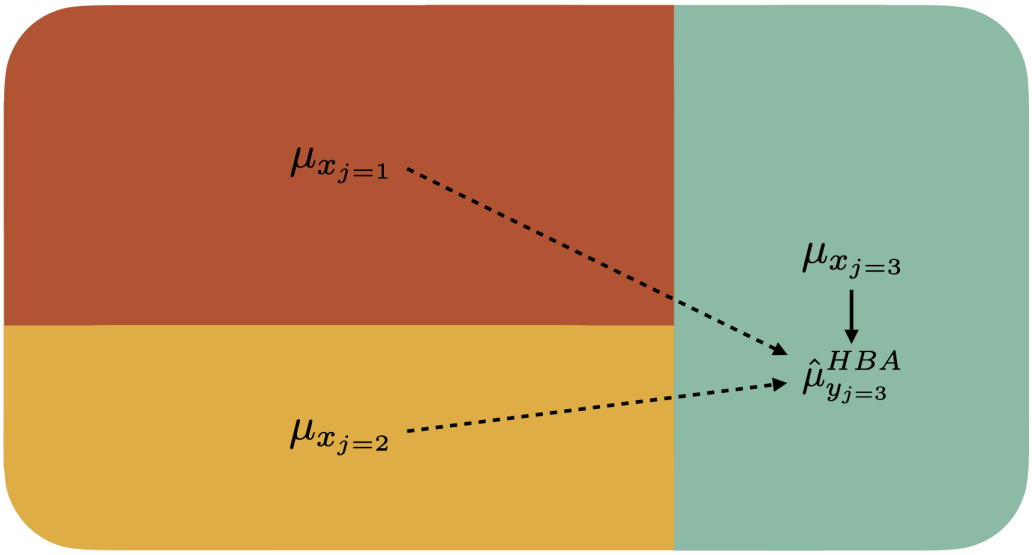

Figure 1.5: The area-level hierarchical Bayesian model producing an estimate in our seafoam green eco-subsection. This estimator relies on auxiliary information from all eco-subsections within a given eco-province to produce estimates as shown by the arrows. Notably, more information is used in the eco-subsection of interest to produce the estimate, denoted by the solid arrows. The dashed arrows tell us that less information is being used from outside eco-subsections. This estimator also produces estimates based on remotely sensed data, as denoted by the x variables in each hexagon. This estimator uses information at the area-level, meaning that it produces estimates based on the post-stratified estimate in each eco-subsection.

It is important to note that while Figures 1.4 and 1.5 describe the hierarchical Bayesian estimators, the diagrams would be the same for the unit- and area-level EBLUP estimators. This is due to the fact that the EBLUP estimators assume a frequentist mixed model rather than a Bayesian one. These estimators are of course different, we just do not explicitly show how the magnitude of strength borrowed is decided in Figures 1.4 and 1.5.

We can see that both the hierarchical Bayesian unit- and area-level models borrow strength from surrounding areas and explicitly model the y variable outcome as a function of remotely sensed x variable(s). The notable difference between the two models is that the hierarchical Bayesian unit-level models borrows strength from the unit-level data while the area-level model borrows strength from data aggregated by the post-stratified direct estimator.

Explicit model-based estimation has been increasing in popularity in the realm of applications to the FIA and forestry data in general. As the FIA requires a reduction in variance for their estimates of increasingly smaller areas, it becomes inevitable that

borrowing strength from surrounding areas, the use of auxiliary data, and the explicit use of a model is needed to maintain a satisfactory amount of variance. Commonly, frequentist model-based estimators are used for model-based small area estimation, such as the EBLUP estimator. Models such as the EBLUP have some very nice properties. Most notably, they are "unbiased" if the assumed model is correct. To understand what it means to have an "unbiased" estimator, we must first define bias of some estimator $\hat{\mu}_{y_j}$ of a parameter μ_{y_j}:

$$\text{Bias}(\hat{\mu}_{y_j}) = E[\hat{\mu}_{y_j}] - \mu_{y_j} \tag{1.1}$$

It intuitively follows that if the modeling assumptions are met and our estimator $\hat{\mu}_{y_j}$ is unbiased that we will have the following property:

$$\text{Bias}(\hat{\mu}_{y_j}) = E[\hat{\mu}_{y_j}] - \mu_{y_j} = 0 \tag{1.2}$$

That is, the expected value of the estimator, $\hat{\mu}_{y_j}$, is in fact the true value of the forest attribute of interest. It is clear as to why this is a trait we would want in our model and to why it is so commonly used, however, what is not clear is the cost of this trait. By only focusing on reducing the bias in our estimates, we must ignore the second piece of the mean squared error, the variance. While it is important for bias to be low, we can often reduce our mean squared error by a large amount by increasing bias slightly, as bias and variance are inversely related. We can see by the representation of the mean squared error (MSE) as the sum of the variance and squared bias of our estimator:

$$\text{MSE}(\hat{\mu}_{y_j}) = \text{Var}(\hat{\mu}_{y_j}) + \text{Bias}(\hat{\mu}_{y_j},\ \mu_{y_j})^2 \tag{1.3}$$

This thesis explores this trade-off between bias and variance in depth. We implement hierarchical Bayesian unit- and area-level models which allow for the estimates to be slightly biased while reducing variance. Throughout this thesis, we compare these techniques to small area estimations methods such as the EBLUP and the post-stratified direct estimator. By applying these models on four response variables across the entire Interior West at the eco-subsection level, we can add a great deal of understanding to the usefulness of hierarchical Bayesian models in a small area estimation context, especially when considering its usefulness to the FIA and other forestry organizations. We only have been able to source one paper which uses hierarchical Bayesian modeling for small area estimation with a forestry application, and they only consider the area-level model with a particular response variable in particular forest (Ver Planck, Finley, & Huff, 2017). This thesis thus adds significantly to our understanding of the usefulness of hierarchical Bayesian small area estimation in a forestry setting due to the introduction of the unit-level model, the vast number of response variables studied, and the vast range of area where we test the usefulness of this model.

Chapter 2

Data

2.1 The Forest Inventory & Analysis Program

The FIA is a program within the United States Forest Service which aims to collect information and data in order to assess the country's forests. The FIA has been continuously operating since 1930 and their official mission is 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 US" (FIA, 2020).

The FIA collects data all throughout the United States by completing a survey each year of many plots of land. The units measured by the FIA and their ground crews are approximately 30 meter by 30 meter hexagonal units. Due to the vast size of the United States and immense amount of forested land, it would be nearly impossible for the FIA to attain population data for the country, so they use sampling instead. The FIA samples from the population of 30 meter by 30 meter hexagonal units by using a geographically-based systematic sampling design (K. S. McConville, Moisen, & Frescino, 2020). The FIA chooses these samples by first overlaying a hexagonal grid over the United States where each hexagon contains approximately 6000 acres of land. Then, they fill these hexagons with much smaller hexagons and randomly sample from the population of small hexagons. Then, ground crews go to these sampled small hexagons and collect variables such as basal area, trees per acre, etc. Along with this hand-collected data from FIA ground crews, the FIA also uses remotely sensed data to gain more information about the areas which they collect data. For example, the `nlcd11` variable, which measures total percent tree canopy cover of a plot, is collected via remote sensing by the Multi-Resolution Land Characteristics Consortium (Homer, 2015). Throughout the duration of the thesis, we will be working to predict ground-collected data with remotely sensed variables, such as `nlcd11`. Having remotely sensed variables like `nlcd11` is useful to us and the FIA because if our models can predict ground-collected variables well, the FIA can collect less data and have a larger effective sample size.

2.2 The Interior West

While the FIA collects data in all regions of the United States, the analyses done in this thesis uses data from the Interior West Forest Inventory and Analysis Unit (IW-FIA). Data from this unit will henceforth be referred to as data from "the Interior West." The Interior West is defined as a broad region of the United States, covering the states of Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming. For reference we have provided the Interior West colored green on a map of the continental United States:

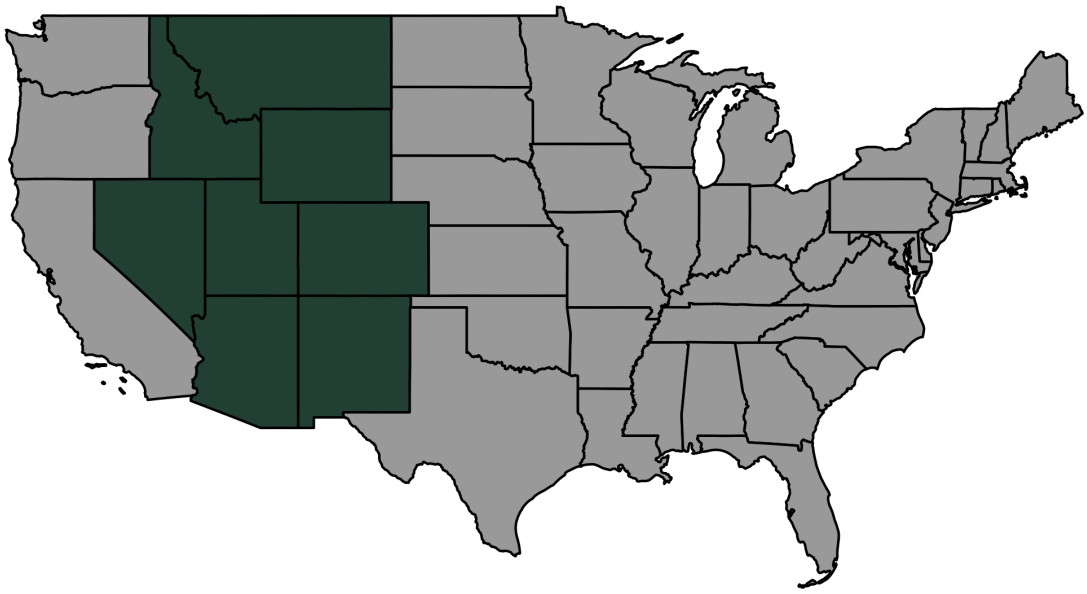

Figure 2.1: The Interior West region of the United States

The IW-FIA collects annual inventories of the Interior West, with the goal of covering 10% of the region each year, so every decade the IW-FIA should have measurement of 100% of each Interior West state's forests.

The Interior West region itself contains the states which encompass the Rocky Mountains along with some other smaller mountain ranges. The Interior West contains 855,767 square miles of land which has an extremely diverse landscape ranging from the high mountain peaks of the Rockies to flat desert plains in Nevada and other Interior West states. Along with desert and mountains, the Interior West also includes parts of the Great Plains. Throughout this diverse landscape, there is a similarly diverse range of forested areas. The forested areas range from areas that are humid and temperate to areas like the Northern Rocky Mountain Forest which is dry and considered a temperate desert.

2.3 Our Data: Specifics

The data used in this thesis was collected by the FIA in the span of 10 years from 2007 to 2017. While this data was collected over this 10 year period, the analyses done throughout this thesis are under the assumption that this is a "snapshot" of the Interior West at some moment in time. Thus we do not consider any temporal features of this dataset, however the inventory year information is available to us. The data we have is plot-level (sometimes referred to as "unit-level") data for the Interior West region of the United States, where the data for each plot consists of ground data collected by FIA and remotely sensed data.

The dataframe used in this thesis is a joined dataframe derived from two FIA datasets of the Interior West, `spatial` and `response`. The `spatial` dataframe contains 89444 observations and 70 variables, most notably our remotely sensed predictor variable (`nlcd11`), location information, and eco-subsection. The `nlcd11` variable was collected by the Multi-Resolution Land Characteristics Consortium (Homer, 2015). This variable measures percent tree canopy cover in a given plot.

The `response` dataframe contains 86085 observations and 67 variables, most notably four response variables collected by FIA crew members (`BALIVE_TPA`, `CNTLIVE_TPA`, `BIOLIVE_TPA`, and `VOLNLIVE_TPA`), location information, and eco-subsection. The response variables noted above measure basal area, tree count, biomass, and volume, respectively. We join these dataframes by their unique plot number, and subset the number of variables significantly to 19 variables which contain plot information, longitude & latitude, elevation, predictor variables, response variables, eco-subsection, eco-section, and eco-province. The resulting joined dataframe has 86085 rows as these are the rows which share the same plots between the `response` and `spatial` dataframes. We can see the first few rows of the dataframe with relevant columns selected and values rounded to the second decimal place:

Table 2.1: Relevant Glimpse of Data

Plot	Latitude	Longitude	nlcd11	BIOLIVE_TPA	subsection
83574	-109.71	32.85	21	0.00	321Af
84904	-109.88	32.99	0	0.00	321Af
83021	-109.88	32.81	0	0.00	321Aj
82635	-109.89	32.65	26	14.74	321Am
90381	-109.83	32.62	41	31.50	321Am
81801	-109.79	32.35	0	0.00	321Aj

While the data covers the Interior West as a whole, we have very granular information, as each row represents a plot sampled by the FIA. The data also includes variables that subset the Interior West into provinces which contain eco-sections, and these eco-sections contain eco-subsections. In our data, on average, each eco-section contains approximately 7.06 eco-subsections, and each province contains an average of

4.86 eco-sections. So, an average province then contains just over 34 eco-subsections. We can take a look at the Northern Rocky Forest province, colored by eco-section, with lines dividing each eco-subsection to see this structure in action:

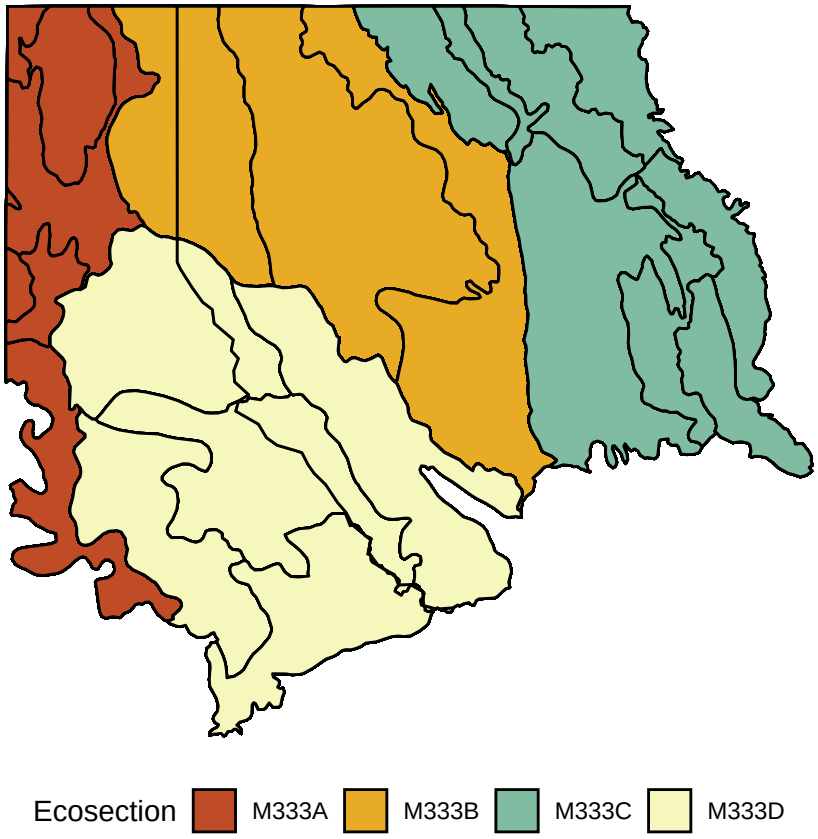

Figure 2.2: The Northern Rocky Forest colored by eco-section

The data we have covers a total of 14 provinces, 68 eco-sections, and 480 eco-subsections. The hierarchical structure of the data and nested nature of the eco-subsections within eco-sections within eco-provinces lends itself to be able to create hierarchical models which borrow strength from surrounding areas.

While this data contains a multitude of variables, the analyses done in this thesis focus on four key response variables and one explanatory variable. The response variables used are basal area (square-foot), trees per acre, above-ground biomass (lbs), and net volume (ft^3). These variables are coded as `BALIVE_TPA`, `CNTLIVE_TPA`, `BIOLIVE_TPA`, and `VOLNLIVE_TPA`, respectively. We can look at the average of these variables across the Interior West region by eco-subsection in the four following maps of the interior west.

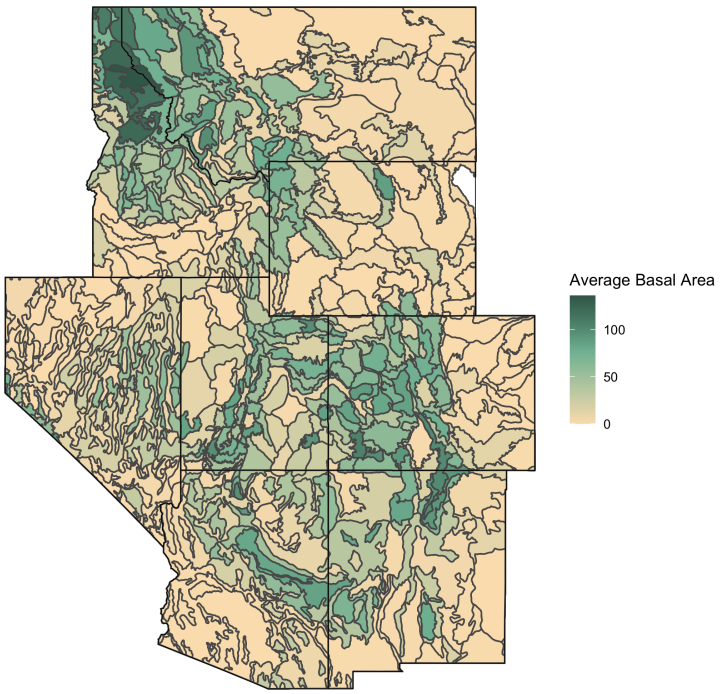

Figure 2.3: Mean basal area in Interior West eco-subsections

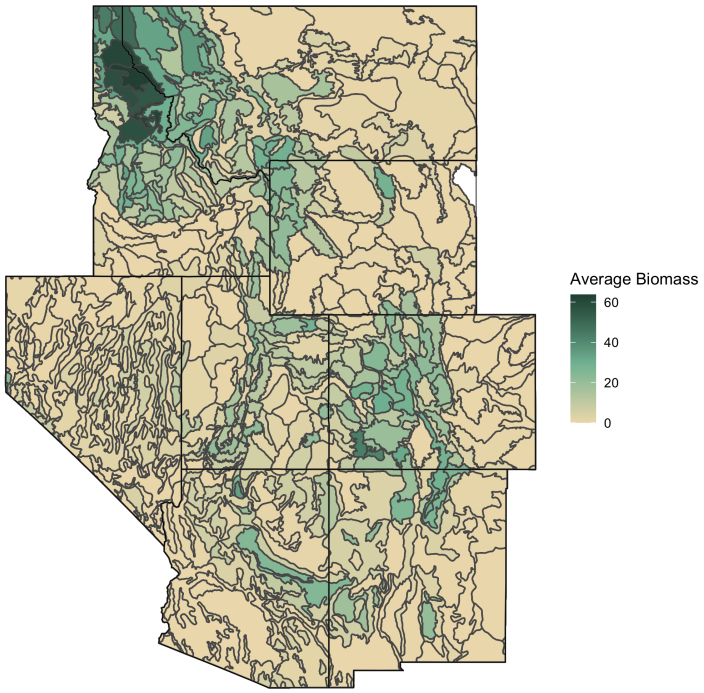

Figure 2.4: Mean biomass in Interior West eco-subsections

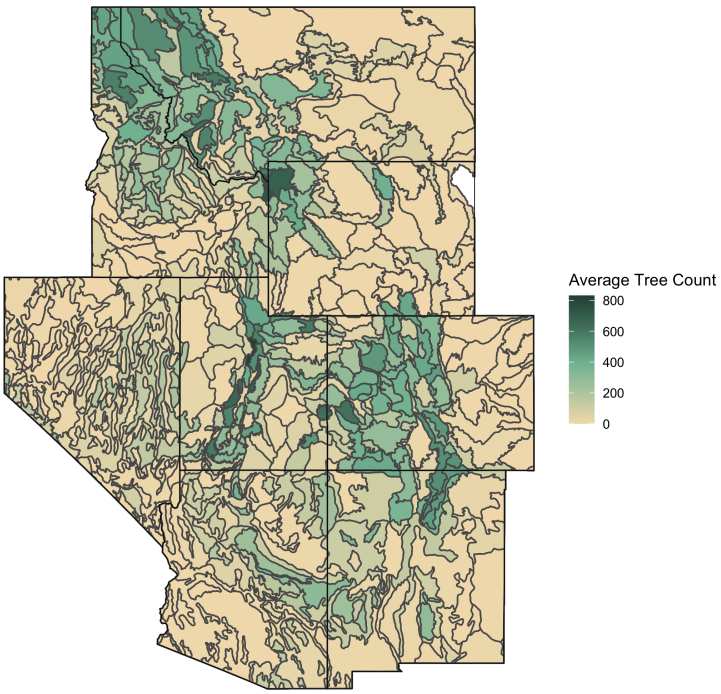

Figure 2.5: Mean tree count per acre in Interior West eco-subsections

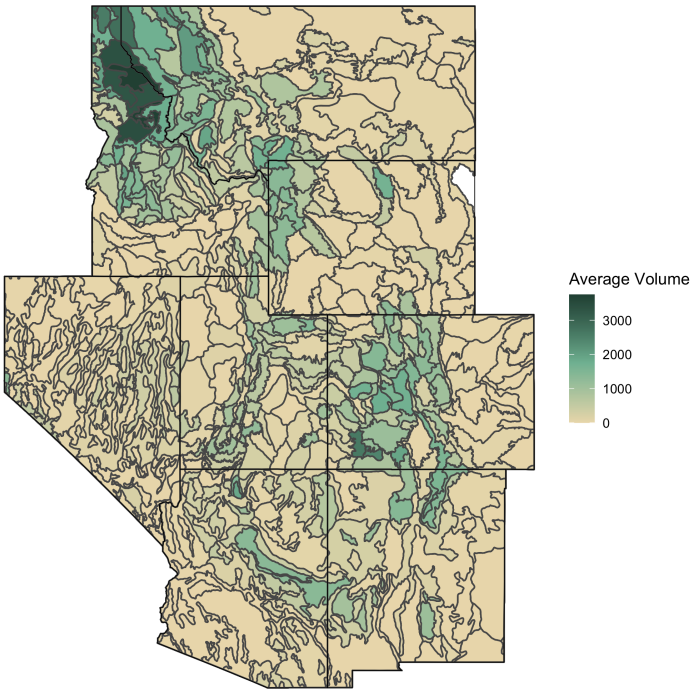

Figure 2.6: Mean net volume in Interior West eco-subsections

While we have four variables which we will model as response variables throughout the analyses, we also have one predictor variables which will be of much use to us. In particular, total tree canopy cover (coded as `nlcd11`.) This variable is remotely sensed, meaning that they were not collected by FIA crew members, but rather with aerial photography and/or satellite imagery. However, we will be using these variables to attempt to predict our response variables in order to understand how good of estimates we can make with this remote data that does not require as much effort to collect. Notably, we can consider this variable to contain the population totals for total canopy cover as the data is collected for the entire Interior West region.

To get a sense of a our predictor variable `nlcd11`, we will look at its distribution in the Northern Rocky Forest subset of our data compared to its distribution across the entire Interior West:

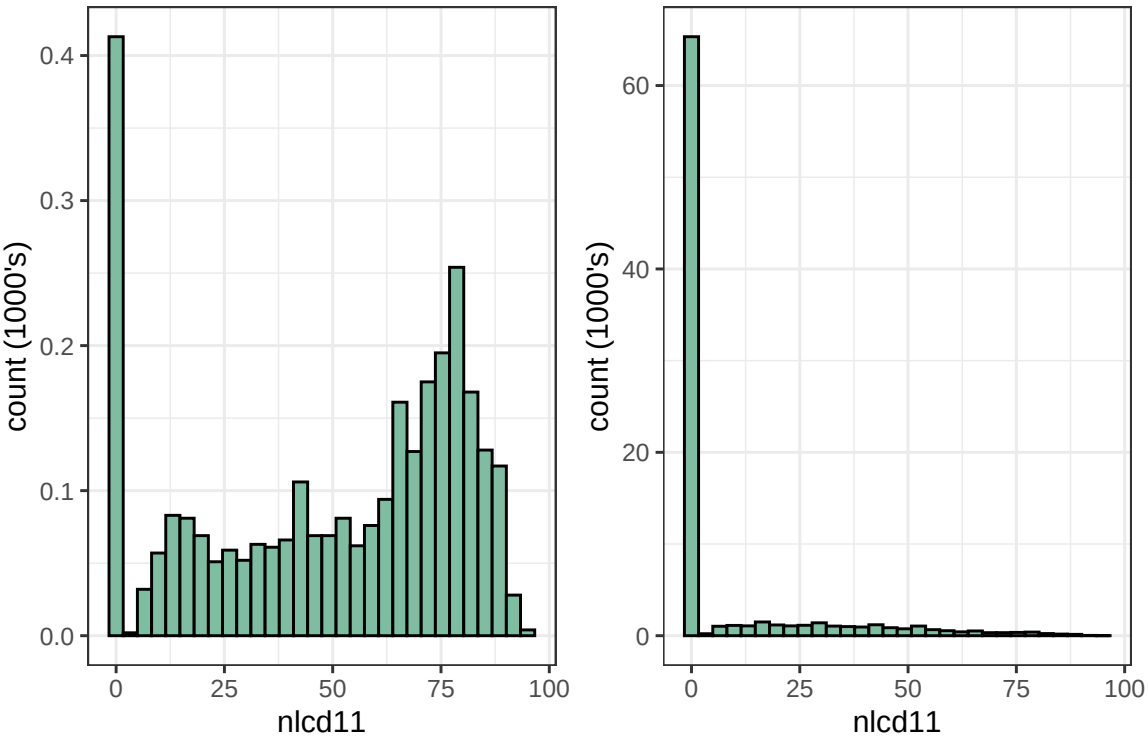

Figure 2.7: Distribution of total canopy cover in the M333 eco-province (left) and the entire Interior West (right)

Notably, the Northern Rocky Forest Province (M333) is much more forested than the Interior West, so we see much different distributions of total canopy cover in this subset of the data. Apart from making these histograms, we can also summarize the entire, unit-level data and see some summary statistics of our five key variables:

Table 2.2: Summary Statistics of Relevant Variables

Variable	Mean	SD	Median	75th Percentile	Min	Max
nlcd11	8.73	18.57	0	0.00	0	95.00

BIOLIVE_TPA	6.23	16.84	0	1.98	0	244.35
BALIVE_TPA	22.75	48.06	0	14.75	0	469.39
CNTLIVE_TPA	98.60	283.09	0	30.09	0	6677.93
VOLNLIVE_TPA	342.32	972.78	0	74.69	0	16435.55

From this table, we can see how heavily skewed these key variables are, with all the variables having median of zero. This does not stop us from doing meaningful analyses though, as the sample size of this dataset is so large ($n = 86085$) and thus we have plenty of data to create models with.

Finally, we also have population data showing the proportion of each eco-subsection that is forested. This data allows us to create our post-stratified estimates which are discussed in detail in the following chapter.

2.4 Data Structure & Hierarchy

As hinted at throughout earlier parts of the chapter, the data used in this thesis has a hierarchical structure, where eco-subsections are nested within eco-sections which are in turn nested within eco-provinces. Every plot has each level of granularity of location data recorded and this is what allows us to choose how far to borrow strength from other plots. We can see this structure of nested data by looking at a diagram depicting this data structure:

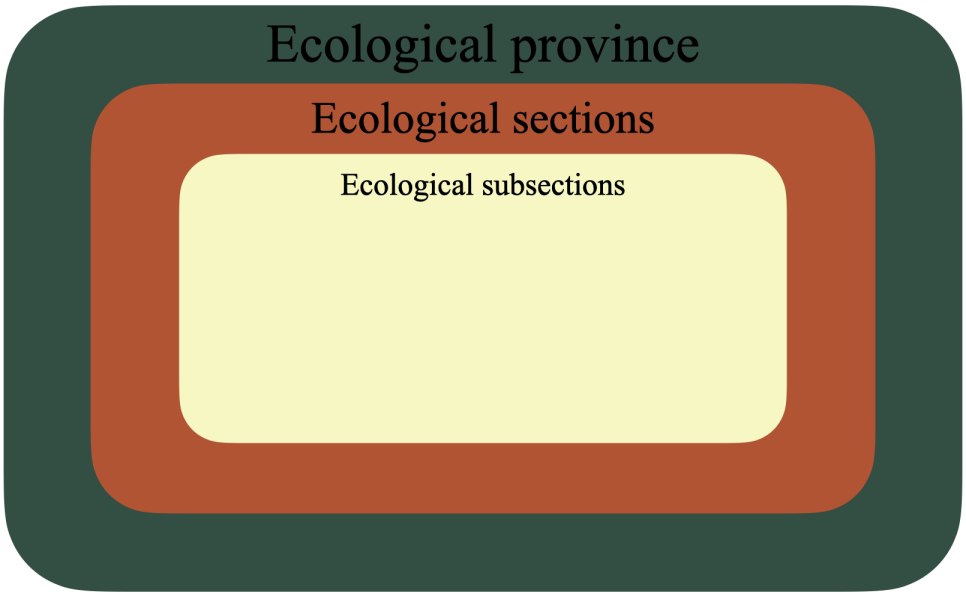

Figure 2.8: The nested data structure of the Interior West. The green area is the eco-province which is divided into eco-sections (red area) which is in turn divided into eco-subsections (beige area).

The largest motivation for hierarchical modeling in this particular application is

that observations are more similar within the hierarchies which we split them into. To understand if this is true, we can do a preliminary analysis on the data by performing three-way ANOVAs for each key variable with predictors `province`, `section`, and `subsection`. For succinctness, we can look at the ANOVA results for one of the response variables, `BIOLIVE_TPA`, but the other variables tell a very similar story in terms of homogeneity. By just looking at the MSE of the ANOVA results, we can see that we should expect more homogeneity within eco-subsections:

Table 2.3: Analysis of Variance Model (Biomass Response)

term	df	sumsq	meansq	statistic	p.value
province	13	6512457.0	500958.2335	2921.45793	0
section	54	967169.1	17910.5394	104.44960	0
subsection	412	2247964.7	5456.2249	31.81928	0
Residuals	85605	14679153.5	171.4754	NA	NA

These results allow us to conclude that it is reasonable to believe that observations within a given eco-province are more homogeneous than observations throughout the Interior West. Thus, if we want eco-subsection level estimates of variables, it makes sense to borrow information from other eco-subsections within the same province as each other. This data structure and homogeneity within provinces is what drives the analyses done henceforth in this thesis.

Chapter 3

Methods

Currently, there are three main types of estimators used to estimate the value of forest attributes: direct, indirect with implicit models, and indirect with explicit models. Direct estimators are commonly thought of as the simplest estimators as they do not borrow strength across small areas for estimation. Direct estimators are hence easy to use and interpret, but we often do not get precise enough estimates with these estimators, in other words, they have high variance. Indirect estimators with implicit models borrow strength implicitly across small areas to produce estimates. These estimators decrease variance by providing a link to related small areas through supplementary data (Rao, 2014). Finally, indirect estimators that make explicit use of a model, or "model-based estimators," aim to reduce variance in estimates by using auxiliary data and making specific allowance for between area variation. As explained in Rao (2014), these model-based estimators have significant advantages over direct estimators and implicit indirect estimators. Notably, model diagnostics can be used, small area-specific measures of precision can be attained (in our case, we compare the coefficient of variation between estimators), and we can used mixed or hierarchical models to capture the dependence structure in the data.

This thesis explores the application of two model-based estimators, the hierarchical Bayesian unit-level model and the hierarchical Bayesian area-level model. These hierarchical Bayesian models are not commonly used in forest inventory research, however these models are often applied in a variety of other application areas ranging from ophthalmology to economic growth to tracking the spread of invasive species, to name a few (Hooten & Wikle, 2008; Iriawan & Yasmirullah, 2019; Tojtovska, Ribarski, & Ljubic, 2019). We compare these novel Bayesian models to the frequentist EBLUP unit- and area-level models and two common direct estimators, the sample mean and the post-stratified estimator. To compare these estimators, we will apply them over many ecological provinces in the Interior West and study their performance when considering four key response variables with one explanatory variable.

In order to explore these estimators in depth, we must introduce notation relevant to them. First of all, our indices will work as follows: i indexes over units sampled; j indexes over eco-subsections or "small areas"; and k indexes over strata. Now, we are interested in estimating the mean of some response variable y, such as trees per acre or biomass, in a small area. So, let μ_{y_j} be the population mean of the study variable

in eco-subsection j in the Interior West. To denote the estimate produced of μ_{y_j} we will use $\hat{\mu}_{y_j}$ with a superscript denoting which estimator is being used. We will also use $V(\hat{\mu}_{y_j})$ to denote the variance of $\hat{\mu}_{y_j}$. In summary, each of our estimators aims to estimate μ_{y_j}, the population mean of some study variable in the jth eco-subsection and our estimated value is denoted by $\hat{\mu}_{y_j}$ and its variance is denoted by $V(\hat{\mu}_{y_j})$. We also must introduce s_j, which is a set. The set s_j includes all units sampled within eco-subsection j. We also introduce U_j, the set of all the area in subsection j. The "U" is chosen as it stands for "universe." Also, we introduce n_j, this denotes the number of sampled units within an eco-subsection j, i.e. the cardinality of s_j.

3.1 Direct Estimation

There are two direct estimators that we implement throughout this thesis: the sample mean (a.k.a. the Horvitz-Thompson estimator) and the post-stratified estimator. While the sample mean is an intuitive choice for estimating the population mean of a variable of interest y, the post-stratified estimator helps correct for over- and under-sampling of forested areas. We also introduce a direct estimator which helps create intuition for the unit-level frequentist model, however we do not fit this model to our data. We will now explore the mathematics behind these estimators in depth.

3.1.1 The Horvitz-Thompson Estimator

What might be the most intuitive approach to estimating the population mean μ_{y_j} is taking the sample mean, i.e. using the "Horvitz-Thompson estimator" as survey samplers like to say (Horvitz & Thompson, 1952). This estimator can be expressed as follows:

$$\hat{\mu}_{y_j}^{HT} = \frac{1}{n_j} \sum_{i \in s_j} y_i \tag{3.1}$$

Ignoring the finite population correction, the variance estimate for Horvitz-Thompson estimator is calculated by:

$$\hat{V}\left(\hat{\mu}_{y_j}^{HT}\right) = \frac{1}{n_j - 1} \sum_{i \in s_j} \left(y_{ij} - \hat{\mu}_{y_j}^{HT}\right)^2 \tag{3.2}$$

Recall that the Horvitz-Thompson estimator is just taking the mean of the study variable of interest, y. Note that y_i represents the value of the study variable in the ith unit sampled of eco-subsection j. This estimator is useful as it is easy to compute and does not require any auxiliary information. However, the Horvitz-Thompson estimator high variance relative to other estimators we will discuss. The post-stratified estimator begins to address the variance of the Horvitz-Thompson estimator.

3.1.2 The Post-Stratified Estimator

The post-stratified estimator is very similar to the Horvitz-Thompson estimator however, as stated above, it addresses variance that occurred from using the Horvitz-Thompson estimator. While decreasing variance seems like a no cost solution to some of our problems, the post-stratified estimator requires auxiliary information in order to be used. The post-stratified estimator is a weighted sum of two Horvitz-Thompson estimators: one Horvitz-Thompson estimator giving the estimate of the mean in sampled units which are forested and the other Horvitz-Thompson estimator in non-forested areas. We then weight these estimates by the true proportion of area in the eco-subsection of interest that is forested. Note that while we are using auxiliary information such as the true proportion of forested area in the eco-subsection of interest and whether or not the sampled units were forested areas or not, both of these pieces of information only consider the eco-subsection of interest. Therefore, since information is not used from outside of the eco-subsection of interest, the post-stratified estimator is still within the family of direct estimators. We can represent the post-stratified estimator as follows:

$$\hat{\mu}_{y_j}^{PS} = \sum_{k=1}^{2} w_{jk} \cdot \hat{\mu}_{y_{jk}}^{HT} \tag{3.3}$$

Ignoring the finite population correction, the variance estimate for post-stratified estimator is calculated by:

$$\hat{V}\left(\hat{\mu}_{y_j}^{PS}\right) = \frac{1}{n_j}\left(\sum_{k=1}^{2} w_{jk} n_{jk} \hat{V}\left(\hat{\mu}_{y_{jk}}^{HT}\right) + \sum_{k=1}^{2}(1 - w_{jk})\frac{n_{jk}}{n_j}\hat{V}\left(\hat{\mu}_{y_{jk}}^{HT}\right)\right)$$

Recall that k indexes over our strata, which in this case is forested and non-forested sampled units. We also have w_k, which is a survey weight entirely decided by the true proportion of eco-subsection j which is forested. For example, if eco-subsection j was 80% forested, we would have $w_1 = 0.8$ and $w_2 = 0.2$. Therefore if we had under-sampled forests at only 60% of our samples, we correct this under-sampling with our survey weights and this results in a better estimate of μ_{y_j} that is less variable.

The post-stratified estimator is a great alternative to the Horvitz-Thompson estimator when auxiliary information related to the response variable of interest is available, and in that situation there is not a justifiable reason to pick the Horvitz-Thompson estimator over the post-stratified estimator as the direct estimator of choice. We have access to the information needed to compute the post-stratified estimate in the Interior West so we will primarily be comparing our indirect estimators to the post-stratified estimator in our results. Also, our area-level indirect model-based estimators will be based on the post-stratified estimate.

3.1.3 The Survey Regression Estimator

Another direct estimator that is often used for small area estimation is the survey regression estimator. While we do not implement in this estimator, the unit-level

frequentist EBLUP estimator can be written as a weighted average of the survey regression estimator and a regression-synthetic estimator. Thus, it is important to understand the intuition behind this estimator in order to fully understand how the unit-level frequentist EBLUP estimator works. We can represent the survey regression estimator as follows:

$$\hat{\mu}_{y_j}^{SR} = \frac{1}{N_j} \sum_{i \in U_j} \hat{y}_{ji} + \frac{1}{n_j} \sum_{i \in s_j} (y_{ji} - \hat{y}_{ji}) \tag{3.4}$$

where N_j is the population size in the jth area. The survey regression estimator has a particular subtlety which makes it arguably not a direct estimator. Particularly, the $\hat{y}_{j,i}$ is calculated through linear regression with β's fit by ordinary least squares across the study region, rather than the small area. This is the case for this particular survey regression estimator, as it is the estimator which we can partially represent the unit-level EBLUP estimator by. One may make a quite legitimate argument that this estimator then should in fact be considered an implicit model-based estimator, however Rao (2014) sticks to the convention that this estimator is in fact still a direct estimator. He argues that we should consider this estimator to be a direct estimator as the strength we are borrowing does not increase the "effective" sample size. Making a final decision on this estimator's class is left as an exercise to the reader.

3.2 Implicit Model-Based Indirect Estimation

Often times it is the case that direct estimation techniques do not provide sufficiently small standard errors to be able to make informative inferences at the small area level. That is the case in our situation, as the FIA's survey design was based around making inferences at larger levels than eco-subsections. One relatively simple approach to decreasing variance of estimates at a small area level is to use auxiliary data from surrounding areas to help attain the estimate in the small area of interest. This process is called borrowing strength. It is very similar to the borrowing of strength outlined in Figures 1.4 and 1.5, however, in the case of implicit model-based indirect estimation we do not allow for between-area variation of explanatory variables. That is, we may borrow strength across small areas by using auxiliary data, but this auxiliary data must be aggregated and have the same effect on each small area within a designated area (in our case, this is the ecological province).

We introduce two implicit model-based indirect estimators commonly used when auxiliary data is available at the unit- and area-levels. These estimators are not used in the final analyses done in this thesis. However, they are crucial building blocks for understanding the frequentist EBLUP models we discuss and implement. We discuss the regression-synthetic estimator used when area-level auxiliary data is available and then the regression-synthetic estimator used for unit-level auxiliary data.

3.2.1 The Area-level Regression-Synthetic Estimator

A common method of an implicit model-based indirect estimator outlined by Rao (2014) for area-level is the regression-synthetic estimator. Here, in order to estimate

our parameter of interest, μ_{y_j}, we fit an ordinary least squares linear regression on our auxiliary variable. Notably, we use $\bar{X}_j$ to denote the auxiliary data population mean in the jth small area.

$$\hat{\mu}_{y_j}^{RSA} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_j \tag{3.5}$$

We note that the $\hat{\beta}$'s are calculated through standard methods of ordinary least squares. Further, it is important to recognize that in the case of the regression-synthetic estimator each small area receives the same $\hat{\beta}$'s as they are calculated by fitting an ordinary least squares model over all small areas within their eco-province. This is the realization of the requirement that the auxiliary information must have the same effect on each small area. It is also important to note that this estimator can be generalized to p predictors, however in this thesis we use $p = 1$.

3.2.2 The Unit-level Regression-Synthetic Estimator

The unit-level regression-synthetic estimator is very similar to the area-level regression-synthetic estimator. Rather than the ordinary least squares model being fit at the area-level, it is fit at the unit-level. We first fit this unit-level model:

$$\hat{y}_{ji} = \hat{\beta}_0 + \hat{\beta}_1 X_{ji}$$

Note that we use capital X to denote the population mean of our explanatory variable at the unit-level. We can do express the unit-level regression-synthetic estimator below:

$$\hat{\mu}_{y_j}^{RSU} = \frac{1}{n_j} \sum_{i \in s_j} \hat{y}_{ji} \tag{3.6}$$

This estimator is very similar to the area-level regression-synthetic estimator as it fits an ordinarily least squares model over the eco-province our small area is in, and the auxiliary information has the same effect on each small area.

3.3 Explicit Model-Based Indirect Estimation

Often times, implicit model-based indirect estimators still do not reduce variance enough to give us estimates with a reasonably low variance. However, we can turn to explicit model-based indirect estimation in order to combat this issue further. Explicit model-based indirect estimators are extremely useful when relevant auxiliary data is available and we would like to allow for between-area variation of these auxiliary variables. By allowing for this between-area variation our models should fit to the data better given that there is truly between-area variation in the population. In the case of forests across a large portion of the United States, this assumption is very reasonable and one could attribute this variation to a number of factors such as temperature, humidity, or even elevation. Rather than attempting the daunting and

nonsensical task of collecting data to fully explain this between-area variation, we fit a mixed model.

In small area estimation it is most common that the mixed model estimator used is one of the EBLUP estimators: either the unit- or area-level variant depending on the resolution of auxiliary data preferred. The EBLUP estimators are unbiased given the modeling assumptions are met, similar to the post-stratified estimator. We will now explore the model specifications of these EBLUP estimators at the unit- and area-level.

3.3.1 The Unit-level EBLUP

In the case of the unit-level EBLUP, we first fit a varying-intercepts linear mixed model at the unit-level:

$$y_{i,j} = x_{ij}\beta + \nu_j + e_{ij} \tag{3.7}$$

where β is calculated from data across the eco-province and will stay constant for each eco-subsection. However, ν_j, the "random effect," will be different for each eco-subsection. This random effect helps account for variation between eco-subsections that the auxiliary data does not account for. The random effects ν and sampling errors e are assumed to be distributed as follows:

$$\nu_j \overset{\text{iid}}{\sim} \text{N}(0,\ \hat{\sigma}^2_\nu), \quad e_{ij} \overset{\text{ind}}{\sim} \text{N}(0,\ \hat{\sigma}^2_e).$$

Intuitively, we can think of the variance of the random effect, $\hat{\sigma}^2_\nu$, as an estimate of the heterogeneity between our small areas j after we account for our fixed effects (Breidenbach & Astrup, 2012). We can also think of the estimate of the variance for the sampling errors as a measurement of within-area variability. From our mixed model, Rao (2014) derives the unit-level EBLUP estimator fit in the frequentist paradigm as the weighted average of the survey regression estimator and the regression-synthetic estimator:

$$\hat{\mu}^{FRU}_{y_j} = \hat{\gamma}_j \hat{\mu}^{SR}_{y_j} + (1 - \hat{\gamma}_j)\hat{\mu}^{RSU}_{y_j} \tag{3.8}$$

where

$$\hat{\gamma}_j = \frac{\hat{\sigma}^2_\nu}{\hat{\sigma}^2_\nu + \hat{\sigma}^2_e} \tag{3.9}$$

It is easy to think about $\hat{\gamma}_j$ if we break it down into two parts: it's numerator and denominator. The numerator measures the between-area variation, and the denominator measures the between-area variation plus the within-area variation, i.e. total variability. Thus, when a large proportion of the overall variation is between-area variation, the unit-level EBLUP will put more weight on the survey regression estimator.

For this thesis, in order to compute the variance of the unit-level EBLUP estimator we chose to bootstrap. Bootstrapping is the process taking repeated samples of

our data with replacement, computing our statistic on that new bootstrap sample, and then repeating this process many times (Efron, 1992). The property of the bootstrap which is useful to us is that it produces good estimates for the variance of our estimator. For our application, we bootstrap within each small area to compute the variance. For each eco-subsection j, we sample n_j plots with replacement. Next, we fit the unit-level EBLUP estimator to the bootstrapped eco-subsections and produce our mean estimate on our first bootstrap sample (we denote each bootstrap sample with b, so in this case $b = 1$).

$$\hat{\mu}_j^{FRU}(b = 1)$$

Next, we repeat this sampling B times, and in our case we choose $B = 500$ to produce 500 estimates:

$$\hat{\mu}_j^{FRU}(b), \quad b = 1, \ldots, 500$$

Finally, we compute our estimate of the variance:

$$\hat{V}(\hat{\mu}_j) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\mu}_j^{FRU}(b) - \overline{\mu}_j^B \right)^2 \tag{3.10}$$

where $\overline{\mu}_j^B$ denotes the sample mean of the bootstrap estimates of μ_j.

3.3.2 The Area-level EBLUP

We can also fit the EBLUP at the area-level. In order to do this, we fit the following linear mixed model:

$$\hat{\mu}_j^{PS} = x_j\beta + \nu_j + e_j \tag{3.11}$$

where we assume

$$\nu_j \overset{\text{iid}}{\sim} \text{N}(0,\ \hat{\sigma}_\nu^2), \quad e_j \overset{\text{ind}}{\sim} \text{N}(0, \sigma_j^2).$$

In the case of the area-level EBLUP, the within-area variation, σ_j^2 is assumed to be known and is not a quantity we estimate. However, we still estimate the between-area variation, $\hat{\sigma}_\nu^2$. The area-level EBLUP estimator is expressed by Rao (2014) as a weighted average of the direct estimator and the regression-synthetic estimator:

$$\hat{\mu}_j^{FRA} = \hat{\gamma}_j \hat{\mu}_j^{PS} + (1 - \hat{\gamma}_j)\hat{\mu}_j^{RSA} \tag{3.12}$$

where

$$\hat{\gamma}_j = \frac{\hat{\sigma}_\nu^2}{\sigma_j^2 + \hat{\sigma}_\nu^2} \tag{3.13}$$

Again, $\hat{\gamma}$ is the ratio of between-area variation and total variation. We note that when most of the variation is between-area variation, the area-level EBLUP will rely on the post-stratified estimator and otherwise it will rely more heavily on the area-level regression-synthetic estimator. We compute the variance of this estimator by

using the MSE of this estimator. The MSE is expressed by the following equation (Hidiroglou & You, 2016):

$$\hat{V}(\hat{\mu}_j^{FRA}) = MSE(\hat{\mu}_j^{FRA}) = g_{1j} + g_{2j} + 2g_{3j} \tag{3.14}$$

where

$$\begin{aligned} g_{1j} &= \hat{\gamma}_j \sigma_j^2, \\ g_{2j} &= \hat{\sigma}_\nu^2 (1 - \hat{\gamma}_j)^2 \mathbf{z_j'} \Big(\sum_j \hat{\gamma}_j \mathbf{z_j} \mathbf{z_j'} \Big)^{-1} \mathbf{z_j}, \\ g_{3j} &= (\sigma_j^2)^2 \cdot (\hat{\sigma}_\nu^2 + \sigma_j^2)^{-3} \cdot 2(\sum_j (\hat{\sigma}_\nu^2 + \sigma_j^2)^{-2})^{-1} \end{aligned}$$

where

$$\mathbf{z_j} = \begin{bmatrix} 1 \\ \overline{X}_j \end{bmatrix}$$

We can intuitively think about each $g_{\#j}$ as follows: g_{1j} accounts of for within-area variation, g_{2j} accounts for variation in estimating the regression parameter β, and g_{3j} accounts for model variance estimation (Hidiroglou & You, 2016).

3.4 A Hierarchical Bayesian Approach

So far in this chapter, we have explored common frequentist approaches to small area estimation. However, the research goal of this thesis is to study the performance of the hierarchical Bayesian model for small area estimation. With a hierarchical Bayesian model, we derive the posterior distribution of our variable of interest with either Markov Chain Monte Carlo (MCMC) methods or through numerical integration. We do this by considering both the likelihood (data given unknown parameters) and prior distributions. This allows us to use Bayes' Theorem in order to attain our posterior.

Before introducing a slew of notation regarding Bayesian statistics, let's take a step back to recall the type of questions we have been asking so far. Each estimator in this chapter so far has attempted to quantify the mean of our response variable y in each eco-subsection j, that is estimating μ_{y_j}. The way in which we attain our estimate $\hat{\mu}_{y_j}$ and its variance is by asking the question: "What sort of $\hat{\mu}_{y_j}$ would we expect to get under hypothetical resampling?" This question is the basis of the name "frequentist" as our estimates in this frequentist paradigm rely on the hypothetical repeated (frequent) sampling. When we choose to use Bayesian methods to estimate μ_{y_j}, we ask a different question: "What is our knowledge of μ_{y_j} based on the data and prior information?" (Bray, 2020).

These questions bring up a philosophical difference between Bayesian and frequentist statistics about the addition of prior information. One may object to this, however in this thesis we add very little prior information and rely quite heavily on the data to much of the work for us. Secondly, the addition of prior information allows us to attain a posterior distribution for our estimate of μ_{y_j}, rather than a point estimate.

Now with this philosophical difference in mind, recall Bayes' Theorem below:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} \tag{3.15}$$

$$\propto P(B \mid A) \cdot P(A) \tag{3.16}$$

Bayes' Theorem allows us to derive one conditional probability, $P(A \mid B)$, given that we know the inverse conditional probability and each marginal distribution (Bayes, 1763). This theorem is the basis of our hierarchical Bayesian models. Replacing A with our parameter of interest, μ_{y_j}, and B with our data we get the following equation:

$$f(\mu_{y_j} \mid \text{data}) \propto P(\text{data} \mid \mu_{y_j}) \cdot P(\mu_{y_j}) \tag{3.17}$$

In our application, we are interested in making inferences about μ_{y_j} given the data we have. This is represented on the left hand side of Equation (3.17). We call this quantity the posterior distribution of μ_{y_j} given the data. As we continue to read left to right, we have our likelihood (or, data) multiplied with the distribution of μ_{y_j}. Clearly we do not know the true value of μ_{y_j} or else we would not attempt to estimate μ_{y_j}. In order to estimate our posterior distribution we must multiply something with the likelihood function, so we choose a "prior" distribution which reflects our prior belief of the distribution of μ_{y_j}. We can visualize obtaining a posterior distribution below:

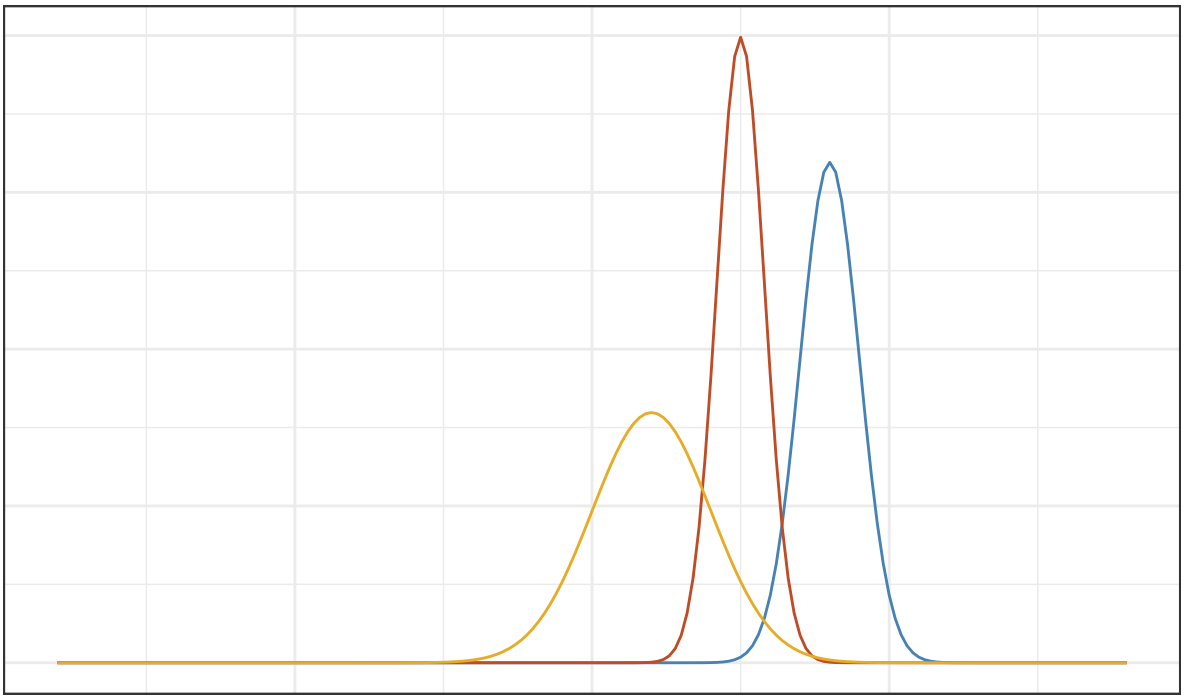

Figure 3.1: Hypothetical prior, likelihood, and posterior distributions. The posterior is inbetween the prior and likelihood as it is the normalized product of these distributions.

Here we can see that our estimate is informed by both the data and the prior distribution, rather than in the frequentist paradigm where only the data is used. We also see that if the prior is "flat," our posterior would just be our likelihood function, leading us to the same result as we achieve through a frequentist analysis, just though a different method.

These are the concepts and basics of Bayesian statistics applied to the problem at hand: estimating μ_{y_j}. We introduce two methods which estimate $\hat{\mu}_{y_j}$: the unit-level hierarchical Bayesian estimator and the area-level hierarchical Bayesian estimator. These estimators are analogous to the unit-level EBLUP and the area-level EBLUP, respectively. Notably, when our variance parameters are known and we use flat priors to attain our posterior the hierarchical Bayesian approach and EBLUPs give the same estimates and measures of variability (Rao, 2014). However, we of course do not know the underlying parameter values of our variances. We now introduce each estimator.

3.4.1 The Unit-level Hierarchical Bayesian Estimator

For the unit-level hierarchical Bayesian estimator, we start with Equation (3.7) as we did for the unit-level frequentist EBLUP. We now just apply a hierarchical Bayesian approach to this equation. We must assume a prior distribution on β, σ_ν^2, and σ_e^2.

We now specify Equation (3.7) as a hierarchical Bayesian model:

$$\begin{aligned} &y_{ij} \mid \beta,\ \nu_j, \sigma_e^2 \sim \mathrm{N}(x_{ij}\beta + \nu_j,\ \sigma_e^2), \\ &\nu_j \mid \sigma_\nu^2 \sim \mathrm{N}(0,\ \sigma_\nu^2), \\ &f(\beta,\ \sigma_\nu^2,\ \sigma_e^2) = f(\beta) \cdot f(\sigma_\nu^2) \cdot f(\sigma_e^2). \end{aligned} \tag{3.18}$$

where $f(\beta)$, $f(\sigma_\nu^2)$, and $f(\sigma_e^2)$ are priors specified below:

$$\begin{aligned} &f(\beta) \propto 1, \\ &\frac{f(\sigma_e^2)}{f(\sigma_\nu^2)} = \mathrm{IG}(\mathrm{df} = 0,\ \mathrm{scale} = 0). \end{aligned}$$

Now that we have specified the estimator, it is easy to attain the small area estimate and variance. For the estimate we have the posterior distribution of μ_{y_j} given :

$$\hat{\mu}_{y_j}^{HBU} = E[\mu_{y_j} \mid \mathbf{y_j}] \tag{3.19}$$

where $\mathbf{y_j}$ is the vector of response values for sampled units in the jth small area. Finally, for the variance we have:

$$\hat{V}(\hat{\mu}_{y_j}^{HBU}) = V(\mu_{y_j} \mid \mathbf{y_j}). \tag{3.20}$$

3.4.2 The Area-level Hierarichical Bayesian Estimator

For the area-level hierarchical Bayesian estimator, we start with Equation (3.11) as we did for the area-level frequentist EBLUP. We now just apply a hierarchical Bayesian approach to this equation. We must assume a prior distribution on β and σ_ν^2. We now specify Equation (3.11) as a hierarchical Bayesian model:

$$\begin{aligned} &\hat{\mu}_{y_j}^{PS} \mid \mu_{y_j},\ \beta,\ \sigma_\nu^2 \sim \mathrm{N}(\mu_{y_j},\ \psi_j), \\ &\mu_{y_j} \mid \beta,\ \sigma_\nu^2 \sim \mathrm{N}(x_j\beta,\ b_j^2\sigma_\nu^2), \\ &f(\beta,\ \sigma_\nu^2) = f(\beta) \cdot f(\sigma_\nu^2). \end{aligned} \tag{3.21}$$

where $f(\beta)$ and $f(\sigma_\nu^2)$ and $f(\sigma_e^2)$ are priors specified below:

$$\begin{aligned} &f(\beta) \propto 1, \\ &f(\sigma_\nu^2) = \mathrm{IG}(\mathrm{df} = 10000 \cdot m,\ \mathrm{scale} = 1). \end{aligned}$$

where m is the number of small areas in the eco-province we fit the estimator in. Now that we have specified the estimator, it is easy to attain the small area estimate and variance. We attain these estimates through numerical integration. For the estimate in eco-subsection j we use the expected value of the posterior distribution of μ_{y_j}:

$$\hat{\mu}_{y_j}^{HBA} = E[\mu_{y_j} \mid \hat{\mu}_{y_j}^{PS}] \tag{3.22}$$

And for the variance in eco-subsection j we have the variance of the same posterior distribution:

$$\hat{V}(\hat{\mu}_{y_j}^{HBA}) = V(\mu_{y_j} \mid \hat{\mu}_{y_j}^{PS}) \tag{3.23}$$

Chapter 4

Results

This chapter addresses the performance of the six estimators in this thesis: the sample mean, the post-stratified estimator, the unit-level EBLUP, the area-level EBLUP, the unit-level hierarchical Bayesian estimator, and the area-level hierarchical Bayesian estimator.

We used the R statistical software to compute our results (R Core Team, 2020). The sample mean was computed using the *sae* (Molina & Marhuenda, 2015). The post-stratified estimates were computed using *mase* (K. McConville, Tang, Zhu, Cheung, & Li, 2018). The frequentist EBLUP estimates were computed using *sae* (Molina & Marhuenda, 2015). The hierarchical Bayesian estimates were computed using *hbsae* (Boonstra, 2012). The results were tidied, processed, and visualized using many *tidyverse* packages (Wickham et al., 2019). Our data spans the entire Interior West, however we were forced to exclude a few eco-provinces. Namely two eco-provinces with a very small number of eco-subsections which caused errors in producing the unit-level hierarchical Bayesian estimates. We also had to not include estimates in a few eco-subsections with either none or very close to no sampled areas with non-zero values for our variables of interest. That is, areas which are in extremely non-forested areas. While this is disappointing, we still were able to fit these estimators to the great majority of our data.

This thesis uses all six of these estimators to produce estimates for basal area (square-foot), tree count per acre, above-ground biomass (lbs), and net volume (ft^3). The EBLUP and hierarchical Bayesian estimators use one explanatory variable, total canopy cover, to produce estimates. Our estimation occurs at the eco-subsection level, and thus we have 10,176 estimates produced (six estimators, four response variables, and 424 eco-subsections).

We are generally concerned with producing estimates that have both low variance and low bias. This chapter primarily aims to answer questions regarding these two quantities. In order to do so, we will summarize our findings both globally and by briefly examining a subset of our results in the Northern Rocky Forest. This subset examination allows us to dig in to results without much aggregation or incomprehensible plots due to the large amount of results we have. We explore both metrics and visualizations of variance and bias across the interior west in order to deeply understand the performance of our estimators. This thesis is, in many ways, a study

navigating the bias-variance trade-off in depth in a real world setting. While quantifying variance of estimates is a generally straightforward task, producing an accurate depiction of bias is extremely difficult without knowing population parameters.

4.1 The Big Picture

We will now investigate the performance of the six estimators used in this thesis. In order to explore variance, the primary metric we use is the coefficient of variation (CV). This metric, for each estimator, is defined as the standard deviation of our estimate divided by the sample mean in that small area ($\hat{\mu}_{y_j}^{HT}$). This allows us to normalize our variation across different response variables and areas that are more or less forested. We can express the coefficient of variation as follows:

$$CV_{y_j} = \frac{\hat{\sigma}_{y_j}}{\hat{\mu}_{y_j}^{HT}} \tag{4.1}$$

It is notable that when the mean of the variable of interest is small, we will sometimes get strangely large coefficients of variation for estimators that borrow strength. This phenomenon can generally be thought of as more to do with intrinsic (and often very good) qualities of the estimator rather than correlation with poor performance of the estimator when the sample mean is small. Notably, estimators which borrow strength to a large degree in situations where the sample mean is low and the areas which we borrow strength from have significantly higher values will have higher variance due to this heterogeneity. We will explore this idea in depth further on in this chapter.

In order to visualize the coefficients of variation in an appealing way, we have filtered out all observations greater than one in the following plot:

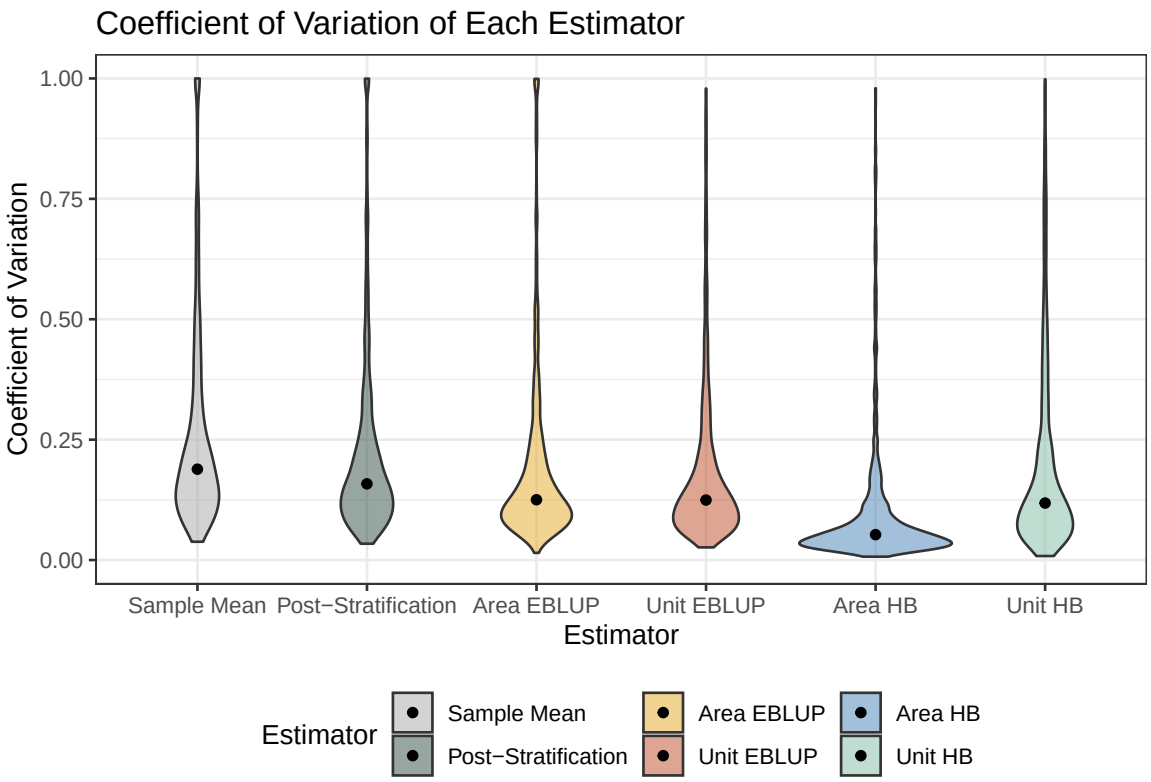

Figure 4.1: Distribution of the coefficient of variation of each eastimator. The black dot in each violin segment represents the median coefficient of variation value. The width of the violin segments corresponds to the density of points in the given value range. Values greater than 1 truncated in plot, however still considered in median calculation.

Figure 4.1 shows the coefficient of variation for each estimator. We see a slight reduction in the coefficient of variation when moving from the sample mean to post-stratification, and a decently large reduction in variation when adding auxiliary data with either EBLUP estimator. We also see that the unit-level hierarchical Bayesian estimator performs quite similarly to the unit-level EBLUP. This makes sense as at the unit-level we have so much data that it outweighs the contribution of any priors significantly. Finally, we see a huge reduction in the coefficient of variation compared to all other estimators when we fit the area-level hierarchical Bayesian estimator. The median coefficient of variation for this area-level hierarchical Bayesian estimator (0.055) is less than half of the next best performing median coefficient of variation from the area-level EBLUP (0.125). Examining the quantiles of these estimator's coefficient of variation continues to show the superiority of the area-level hierarchical Bayesian estimator while also demonstrating the metric's tendency to have some extremely large values:

Table 4.1: Quantiles of Each Estimator's Coefficient of Variation

Estimator	0%	25%	50%	75%	100%
Sample Mean	0.038	0.123	0.189	0.318	1.000
Post-Stratification	0.034	0.105	0.158	0.255	1.000
Area EBLUP	0.015	0.087	0.125	0.213	0.999
Unit EBLUP	0.026	0.085	0.131	0.244	52.408
Area HB	0.007	0.034	0.055	0.112	56.401
Unit HB	0.008	0.078	0.150	0.486	299.673

While Figure 4.1 and Table 4.1 gives us a good sense of the distribution, we must acknowledge that we truncated some values in Figure 4.1 and further investigate why the coefficient of variation tends to get very large in some cases. By examining the distribution and performance of the coefficients of variation by truncating the tails in Figure 4.1 we are able to gain insight that would be much harder to see while including the tails. Now, we can look more in depth at the tails of these distributions. Three estimators had coefficients of variation that exceeded 1. The below table shows the count and proportion of each:

Table 4.2: Coefficient of variation estimates greater than one

Estimator	Count	Proportion
Post-Stratification	0	0.000
Area EBLUP	0	0.000
Sample Mean	0	0.000
Area HB	81	0.048
Unit EBLUP	103	0.061
Unit HB	289	0.170

We see that there are three estimators with coefficient of variations greater than one. Both the frequentist and hierarchical Bayesian unit-level estimators have many values greater than one, and the area-level hierarchical Bayesian has some as well. To understand why these particular estimators have many estimates greater than one, we first need to explore the small areas which have these coefficient of variation values. To do so, we can examine the mean response value for estimates where the coefficient of variation is greater than one and when it is less than one.

Table 4.3: Mean Estimates Where Coefficent of Variation is Greater Than and Less Than One

Response	Mean Estimate (CV < 1)	Mean Estimate (CV > 1)
BALIVE_TPA	39.655	17.473
BIOLIVE_TPA	11.522	3.760
CNTLIVE_TPA	181.724	96.469
VOLNLIVE_TPA	615.601	130.547

From this table we are able to confirm that we see high coefficients of variation more often in areas with lower values for our response variable. It now becomes apparent why some estimators have many coefficients of variation greater than one and others have none or very few. The three estimators which have coefficients of variation greater than one all will borrow a great deal of strength from outside eco-subsections when those outside eco-subections have a much higher response variable value than the eco-subsection of interest. In other words, with these three estimators, eco-subsections with lower response variable sample means will produce larger estimates. These larger estimates require larger a larger variance as we are unsure whether we had a non-representative sample (too low) in the given eco-subsection or if this eco-subsection truly has a low value for its response variable(s). However, when we are selecting methods of estimation which borrow strength, this is a property we like to see. In fact, estimators which do not borrow strength in these situations and have small coefficients of variation could be producing under-estimates with an artificial amount of precision. We can first visualize this phenomenon by examining the coefficient of variation of each estimator across the Interior West:

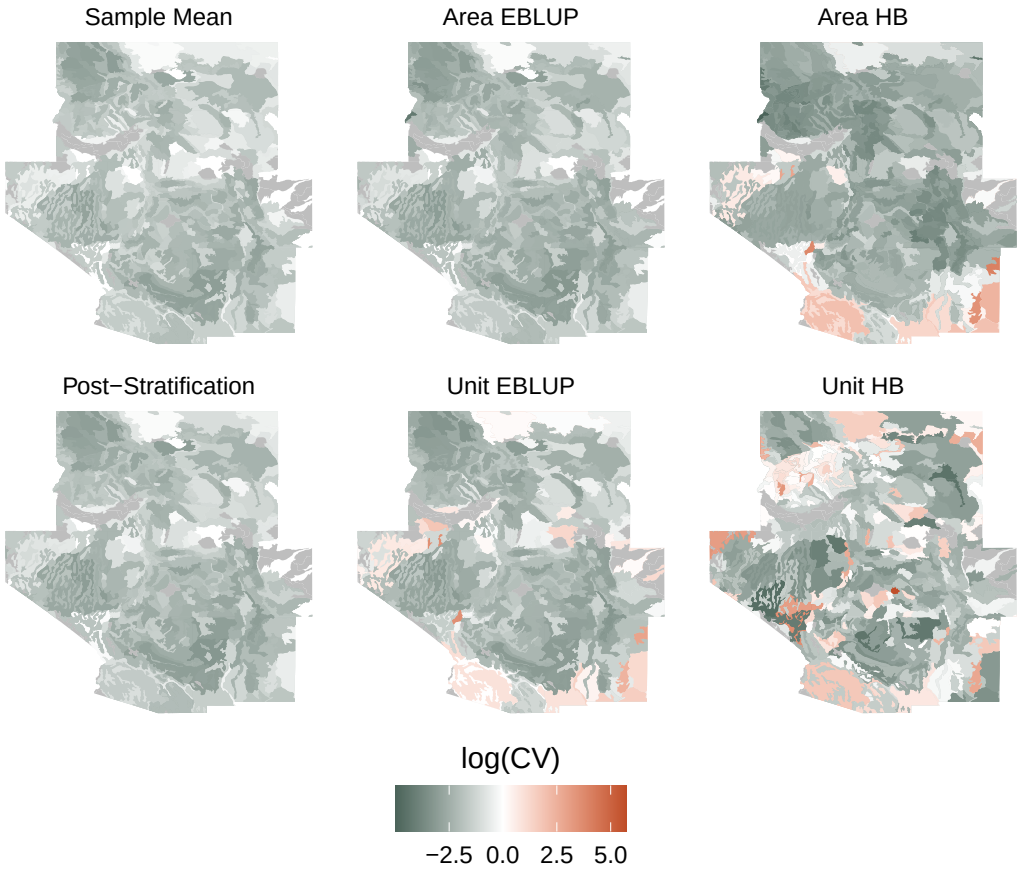

Figure 4.2: Average log coefficient of variation of each estimator in each eco-subsection plotted onto a map of the Interior West. The average is taken over the four response variables. The natural log is used to perserve a reasonable color scale. Grey areas represent areas where models were not fit.

We see that the three estimators which have coefficient of variation values higher than one generally produce these values in heavily deserted areas in the southern parts of the Interior West and some deserts in the north parts of Nevada. It is sensible that these areas may truly have low values for our response variables given that the post-stratified estimator is doing its job correctly and the population totals for forested and non-forested areas are correct. On the other hand, it may be reasonable to use a higher variance estimate in these cases and borrow strength from surrounding areas. While it is hard to say which approach to handling these situations gives a better estimate, it highlights distinct ways in which the hierarchical Bayesian estimators (and the unit-level EBLUP) perform differently than the direct estimators and area-level EBLUP.

4.2 Zooming In: The Northern Rocky Forest

In order to give ourselves a closer look at *how* borrowing of strength differs between the area-level hierarchical Bayesian estimator and the area-level EBLUP estimator, we zoom in to basal area in the Northern Rocky Forest. Immediately, when we examine the Northern Rocky Forest we notice the behavior of the hierarchical Bayesian area-level estimator "pulling up" lower valued eco-subsections compared to their post-stratified estimate:

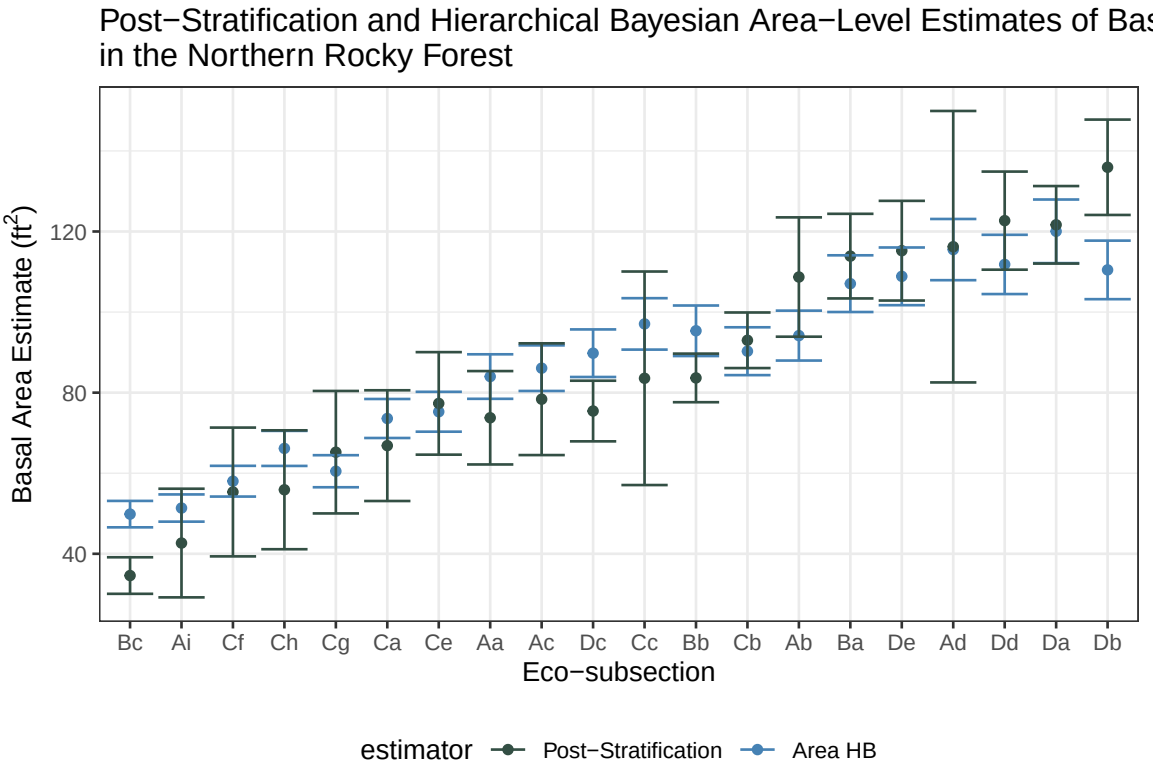

Figure 4.3: The hierarchical Bayesian area-level estimator and post-stratified estimator predicting basal area in the Northern Rocky Forest. The error bars depict two standard errors above an below the estimate of our response variable.

We also observe that the eco-subsections with high response values will get "pulled down," however the coefficient of variation does not increase significantly here as the sample mean is still relatively high. The "pulling up" and "pushing down" of estimates is a trait expected of these hierarchical Bayesian models, and it is the realization of "borrowing strength" or "pooling information" (McElreath, 2020).

The frequentist area-level estimator takes a different approach to making these estimates. Notably, in the case of small within-area variation we see that the EBLUP will put a heavy weight on the post-stratified estimate from Equation (3.12). We can see this occur by comparing the area-level frequentist estimator to the post-stratified estimate with the same eco-province and response variable as above:

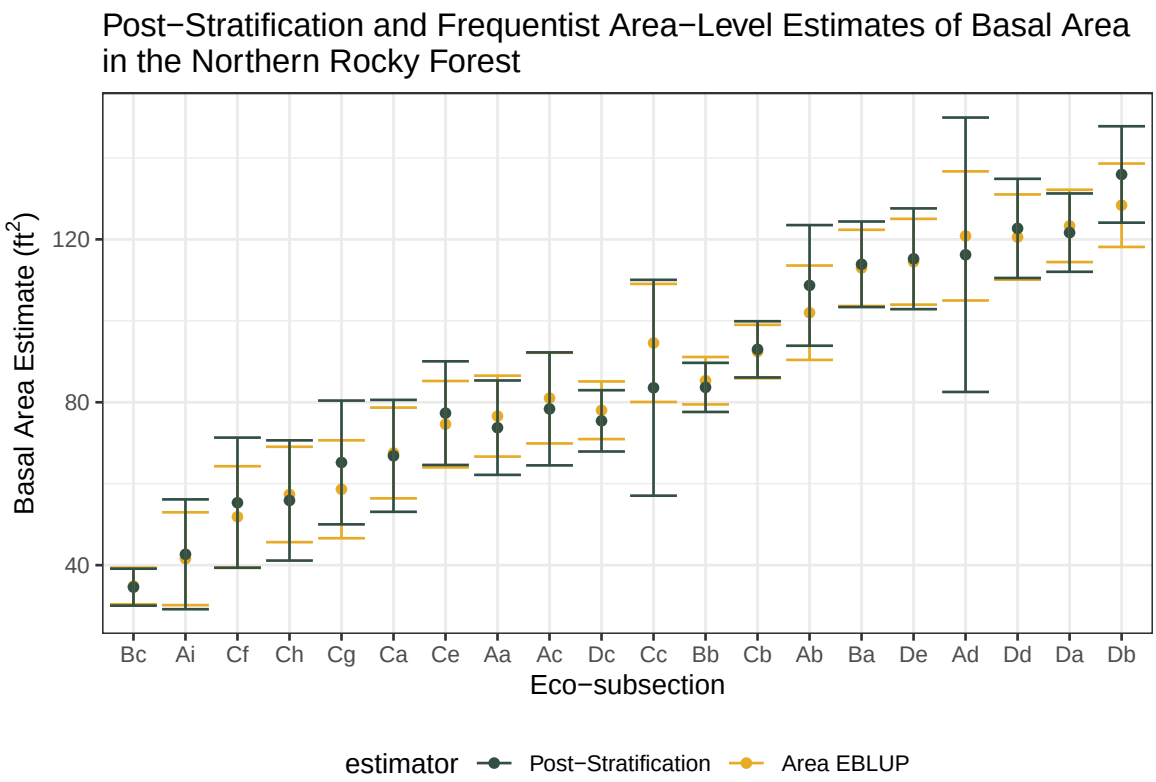

Figure 4.4: The frequentist area-level estimator and post-stratified estimator predicting basal area in the Northern Rocky Forest. The error bars depict two standard errors above an below the estimate of our response variable.

Overall, we can see that the frequentist area-level estimator performs much more similarly to the post-stratified estimator, a trait that may or may not be appealing based on the confidence we have in representative sampling. However, it is important to note that the frequentist area-level estimator is still borrowing strength, it just seems that it will often borrow less strength than the hierarchical Bayesian area-level estimator. This is expected, if we recall Equation (3.12) we note that when a large proportion of the total variation is between-area variation, the area-level EBLUP will rely heavily on the post-stratified estimate. Thus, in eco-subsections with low within-area variation the area-level EBLUP will produce an estimate very similar to the post-stratified estimate. It is important to recall that much of our data contains many observations which equal zero (Figure 2.7 and Table 2.2). So, in eco-subsections with a low number of sampled plots and a high proportion of sampled plots with response values equal to zero we may artificially lower our variance. The hierarchical Bayesian estimators have the unique property of considering the uncertainty in the variance estimate when producing estimates, rather than relying on just a point value.

4.3 Stepping Back

When we narrow our vision to the Northern Rocky Forest it does seem like the area-level hierarchical Bayesian estimator clearly outperforms the analogous frequentist estimator. However, we must return to the Interior West in order to understand trends of estimator performance over space. To do this, we simply examine the eco-subsections in which the area-level hierarchical Bayesian estimator has lower variance than the post-stratified estimator and the area-level EBLUP:

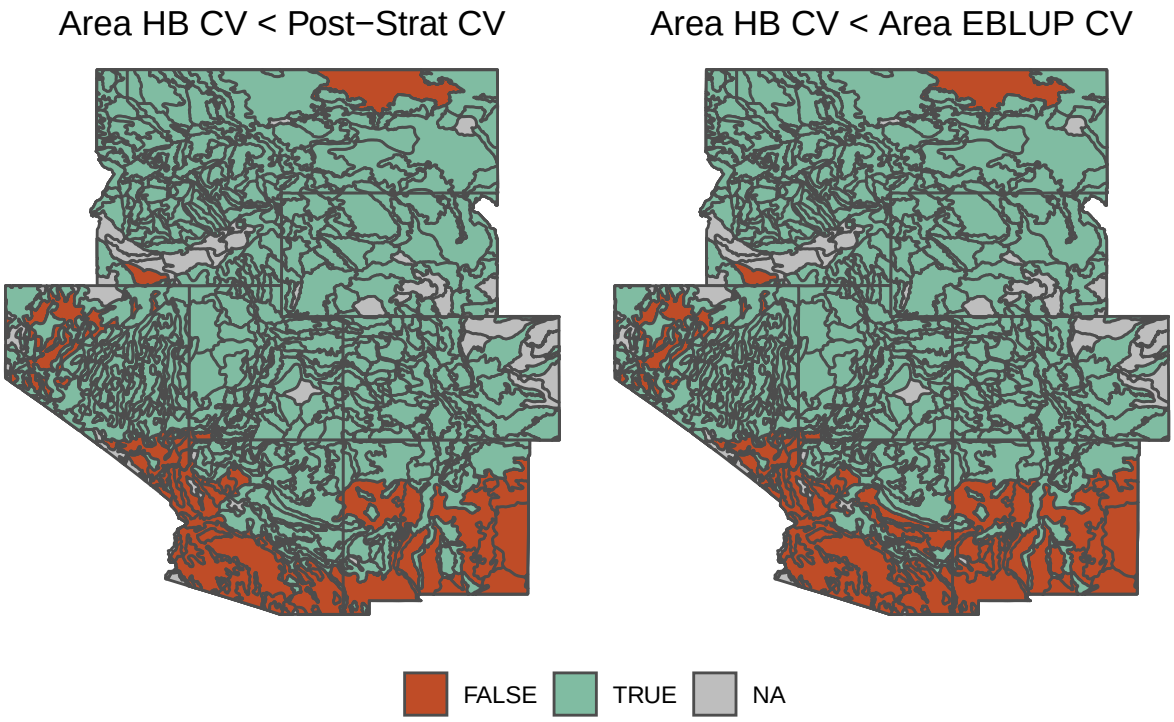

Figure 4.5: The area-level hierarcical Bayesian estimator's coefficient of variation compared with the coefficient of variation for the post-stratified estimator (left) and area-level EBLUP (right). Eco-subsections in seafoam green indicate areas where the coefficient of variation is lower for the area-level hierarchical Bayesian estimator, red areas indicate that the coefficient of variation of the hierarchical Bayesian estimator is greater than the estimator we are comparing it to. Grey areas indicate areas where we did not fit these estimators.

Generally, we continue to see the trend in Figure 4.2 where the area-level hierarchical Bayesian estimator producing estimates with higher variance in heavily deserted areas in the southern parts of the Interior West and deserts in the north parts of Nevada. The areas in which the area-level hierarchical Bayesian estimator has more variation than the post-stratified estimator and the area-level EBLUP are clearly non-randomly distributed across the Interior West. This gives us great insight as to where we may want to use this hierarchical Bayesian estimator and further supports our idea that the area-level hierarchical Bayesian estimator has a high coefficient of

variation in non-forested areas.

We have largely discussed the variance of these estimators thus far, as we cannot know the true bias of an estimator without knowing the population parameters μ_{y_j} in each eco-subsection. However, in an attempt to quantify the magnitude of difference between two estimators we introduce percent relative difference, which we define below.

$$PRD(\hat{\mu}_1,\ \hat{\mu}_2) = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\mu}_2} \cdot 100$$

While we could explore percent relative difference between all combinations of estimators, the Forest Inventory and Analysis Program's go-to estimator is post-stratification, so we will compare the remaining five estimators to post-stratification. Also, post-stratification, as discussed in Methods, is an unbiased estimator given assumptions about our auxiliary data. Thus, we will rely on it for some proxy of truth. However we must realize that this proxy is quite weak. Below, we can see the quantiles of percent relative difference to post-stratification.

Table 4.4: Quantiles of Percent Relative Difference to the Post-Stratified Estimator

Estimator	0%	25%	50%	75%	100%
Sample Mean	-60.997	-3.826	0.056	5.610	1102.114
Unit HB	-99.928	1.326	21.816	94.770	24224.087
Area HB	-74.491	-13.192	13.095	66.254	105994.261
Unit EBLUP	-80.319	3.521	25.471	99.885	27724.471
Area EBLUP	-79.641	-3.747	-0.165	2.802	160.843

These quantiles continue to enforce the idea that the area-level EBLUP generally performs very similarly to the post-stratified estimator (albeit with lower variance). Notably though, the percent relative difference between the area-level hierarchical Bayesian estimator and the post-stratified estimator is small compared to both unit-level models. The median values in Table 4.4 can be thought of some sort of proxy for each estimator's bias, but this is a difficult equivalency to make when we do not have the parameter values. While this proxy for bias may be weak, it is reassuring that for the area-level hierarchical Bayesian estimator our median (50% quantile) percent relative difference is somewhat low compared to other common estimators.

Another aspect of our results that percent relative difference allows us to investigate is the performance of an estimator which we have hardly addressed so far: the hierarchical Bayesian unit-level model. This thesis claims to be a hierarchical Bayesian approach to small area estimation, yet the performance of one of the hierarchical Bayesian estimators has yet to be discussed in depth in this chapter. This is because the area-level hierarchical Bayesian estimator outperforms it by a huge margin in terms of variance. However, the case of the unit-level hierarchical Bayesian estimator is interesting in itself, even if it does not increase precision of estimates.

To understand its performance, let's examine its percent relative difference to the unit-level EBLUP estimator:

Table 4.5: Quantiles of Percent Relative Difference to the Unit-level EBLUP

Estimator	0%	25%	50%	75%	100%
Unit HB	-99.845	-2.736	-0.658	-0.119	4.844

We immediately see that the unit-level hierarchical Bayesian estimator is extremely similar to the unit-level EBLUP. That is, many of their estimates are within just a few percent difference compared to the unit-level EBLUP. This is because the data (or, likelihood) outweighs any prior information we gave the this model to a large degree. At the unit level we have so many observations that any reasonable prior has a negligible effect on the estimator's results. This illustrates the idea that under "flat" or "uninformative" priors that the hierarchical Bayesian estimates will converge to the analogous EBLUP estimator's estimates. Also, recall that not only the estimates should converge to the same point, but the variance should as well. We can see this illustrated in Figure 4.1 and below:

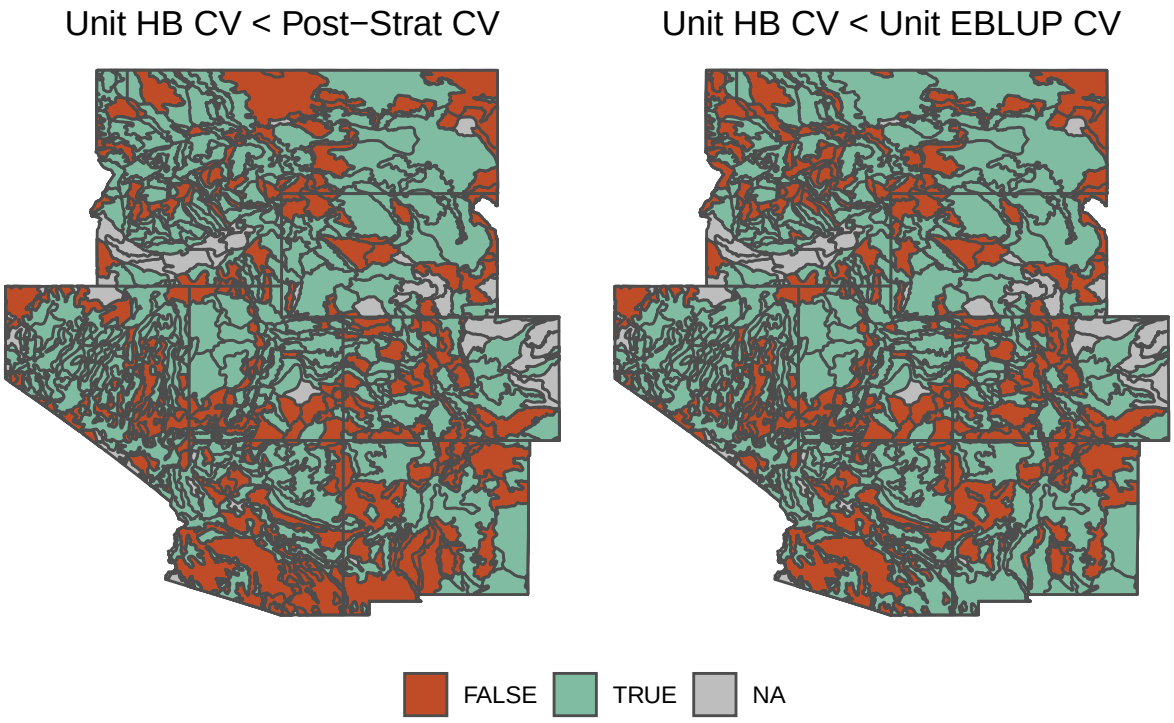

Figure 4.6: The unit-level hierarcical Bayesian estimator's coefficient of variation compared with the coefficient of variation for the post-stratified estimator (left) and unit-level EBLUP (right). Eco-subsections in seafoam green indicate areas where the coefficient of variation is lower for the unit-level hierarchical Bayesian estimator, red areas indicate that the coefficient of variation of the hierarchical Bayesian estimator is greater than the estimator we are comparing it to. Grey areas indicate areas where we did not fit these estimators.

The distribution of areas where the hierarchical Bayesian unit-level estimator has lower variance than the unit-level EBLUP appears to be very random. Just under 46% of areas have the hierarchical Bayesian unit-level model with lower variance than the unit-level EBLUP. We are not making any significant variance gains by implementing the unit-level hierarchical Bayesian estimator, however there are benefits of the implementation of this estimator. First, we now have a case study of a hierarchical

Bayesian estimator converging in variance and estimates to the analogous frequentist model. This allows us to verify a statement often made in Bayesian statistics courses and textbooks with a real world example rather than a mathematical proof. Second, we now have access to a much more flexible estimator than the unit-level EBLUP that does a very similar job to the unit-level EBLUP. We can "tweak" the behavior of this estimator through the use of priors, or even a different linking model if needed. These changes could allow for a more successful estimator at the unit-level without changes many processes computationally. If we were to attempt to approach improving the unit-level model starting with the EBLUP we would be met with restrictions due to the rigidity of the EBLUP's mathematics, and we would have to construct a new estimator entirely.

Chapter 5

Discussion

We have introduced, fit, and presented the results of six small area estimators across the Interior West with four forestry-related response variables. Particularly, we introduce two estimators which are uncommon in forestry research: the hierarchical Bayesian unit- and area-level estimators. The area-level hierarchical Bayesian estimator reduces variance in estimates significantly compared to common direct estimators, frequentist model-based approaches (EBLUPs), and the unit-level hierarchical Bayesian estimator. We see reduction in variance compared to the area-level frequentist estimator and post-stratification in the great majority of small areas when implementing the hierarchical Bayesian area-level estimator. In extremely non-forested small areas, the hierarchical Bayesian area-level estimator consistently has higher variance than post-stratification or the area-level EBLUP. This is likely due to the way in which the area-level hierarchical Bayesian estimator borrows strength.

5.1 Possible Extensions

Now we turn to possible extensions to the work which has been done in this thesis. We explore ideas for further research on adjacent topics and possibilities for applying these methods in other locations.

First, attempting to understand the performance of these hierarchical Bayesian estimators outside of the Interior West region of the United States through implementing these methods across the country would be insightful to the generalization of these estimators. Outside of the Interior West, the Forest Inventory and Analysis Programs has research stations in the Pacific Northwest region, Northern region, and Southern region. The hierarchical Bayesian area-level estimator is also a great candidate for estimation in Alaska. The reasons for its great candidacy are twofold: first, it is expensive and difficult to send foresters to collect data in Alaska due to the extreme rurality of the area; second, Alaska is very forested and we thus would likely see a large reduction in variance of estimates when using the hierarchical Bayesian area-level estimator.

- some sort of hybrid estimator: rely on eblup area when it does well (desert), and bayes otherwise (talk about how bayes performance was very non-random)

- explicit spatial analysis
- simulation study to understand the bias of these estimators
- different regions of the united states: list all FIA regions
- improved software for bayesian sae, more flexible software, non-normal response? (gamma or tweedie)
- different aux data, desert/non-desert could be helpful

Appendix A

Code Appendix

This appendix includes the code used to create the estimates used in this thesis. We include both the helper functions created to run the analyses and the implementation of these helper functions.

A.1 Helper Functions

A.1.1 The Sample Mean

```
direct_estimate <- function(data, response, small_area) {
  # Load packages
  library(sae)

  # Create dataframe
  dat <- data.frame(
    y = data[[response]],
    small_area = data[[small_area]]
  )

  # Compute estimate
  sae::direct(y = dat$y,
              dom = dat$small_area,
              replace = TRUE)
}
```

A.1.2 Post-Stratification

```
postStrat2_bio <- function(data, strata) {
  est <- mase::postStrat(y = data[["BIOLIVE_TPA"]],
                  x_sample = data[["FIAstrat"]],
```

```
                  x_pop = strata,
                  data_type = "totals",
                  var_est = T)
  return(data.frame(ps_est = est$pop_mean,
                    sd = sqrt(est$pop_mean_var)))
}
postStrat2_ba <- function(data, strata) {
  est <- mase::postStrat(y = data[["BALIVE_TPA"]],
                  x_sample = data[["FIAstrat"]],
                  x_pop = strata,
                  data_type = "totals",
                  var_est = T)
  return(data.frame(ps_est = est$pop_mean,
                    sd = sqrt(est$pop_mean_var)))
}

postStrat2_voln <- function(data, strata) {
  est <- mase::postStrat(y = data[["VOLNLIVE_TPA"]],
                  x_sample = data[["FIAstrat"]],
                  x_pop = strata,
                  data_type = "totals",
                  var_est = T)
  return(data.frame(ps_est = est$pop_mean,
                    sd = sqrt(est$pop_mean_var)))
}

postStrat2_cnt <- function(data, strata) {
  est <- mase::postStrat(y = data[["CNTLIVE_TPA"]],
                  x_sample = data[["FIAstrat"]],
                  x_pop = strata,
                  data_type = "totals",
                  var_est = T)
  return(data.frame(ps_est = est$pop_mean,
                    sd = sqrt(est$pop_mean_var)))
}
```

A.1.3 Hierarchical Bayesian Unit-Level

```
hb_unit <- function(data, formula, small_area, pop_data) {
  # Load packages
  library(tidyverse)
  library(hbsae)
```

```r
# Create model frame
model_frame <- model.frame(formula, data) %>%
  dplyr::mutate(small_area = data[[small_area]])
colnames(model_frame) <- c("y", "x", "small_area")

# Area population sizes
pop_size <- pop_data %>%
  dplyr::filter(zoneid %in% model_frame$small_area) %>%
  dplyr::select(zoneid, sum) %>%
  dplyr::rename(pop_size = sum) %>%
  dplyr::select(pop_size)

# Create population means matrix
pop_means <- pop_data %>%
  dplyr::filter(zoneid %in% model_frame$small_area) %>%
  dplyr::select(zoneid, mean) %>%
  dplyr::rename(x = mean) %>%
  column_to_rownames("zoneid")

# Create lambda
anova <- aov(y ~ small_area, data = model_frame)
l <- summary(anova)[[1]]["small_area", "F value"]

# Fit the model
mod <- fSAE.Unit(
  y = model.frame(formula, data = data)[, 1],
  X = data.frame(X = model.frame(formula, data = data)[,-1]),
  area = data[[small_area]],
  Narea = pop_size$pop_size,
  Xpop = pop_means,
  fpc = TRUE,
  lambda0 = l,
  silent = T
)

# Calculate CoV
mean_y <- model_frame %>%
  dplyr::group_by(small_area) %>%
  dplyr::summarise(mean_y = mean(y))
CoV <- hbsae::SE(mod) / mean_y$mean_y

## Add to model object
mod$CoV <- CoV
```

```
  # Print model
  mod
}
```

A.1.4 Hierarchical Bayesian Area-Level

```
hb_area <- function(data, formula, small_area,
                    pop_data, post_strat_data) {
  # Load packages
  library(tidyverse)
  library(hbsae)

  # Create unnamed model frame (to call correct y var in a filter)
  mf <- model.frame(formula, data)

  # Create model frame
  model_frame <- model.frame(formula, data) %>%
    dplyr::mutate(small_area = data[[small_area]])
  colnames(model_frame) <- c("y", "x", "small_area")

  # Direct X
  X <- pop_data %>%
    dplyr::filter(zoneid %in% model_frame$small_area) %>%
    dplyr::select(zoneid, mean) %>%
    dplyr::rename(mean_x = mean,
                  small_area = zoneid) %>%
    dplyr::arrange(small_area)

  # Compute direct estimate
  mean <- direct_estimate(model_frame, "y", "small_area") %>%
    dplyr::mutate(var = SD^2)

  dir <- post_strat_data %>%
    filter(response %in% colnames(mf)[1],
           province %in% unique(data$province)) %>%
    arrange(subsection)

  # Create lambda
  anova <- aov(y ~ small_area, data = model_frame)
  l <- summary(anova)[[1]]["small_area", "F value"]

  # Fit the model
  mod <- fSAE.Area(
```

```
    est.init = dir$est,
    var.init = dir$var,
    X = X %>% dplyr::select(mean_x),
    lambda0 = 1
  )

  # Calculate CoV
   CoV <- hbsae::SE(mod) / mean$Direct
   mod$CoV <- CoV

  # Print model
  mod
}
```

A.1.5 Frequentist Unit-Level

```
freq_unit <- function(data, formula, small_area, pop_data) {
  # Load packages
  library(tidyverse)
  library(sae)

  # Create model frame
  model_frame <- model.frame(formula, data) %>%
    dplyr::mutate(small_area = data[[small_area]])
  colnames(model_frame) <- c("y", "x", "small_area")

  # Area population sizes
  pop_size <- pop_data %>%
    dplyr::filter(zoneid %in% model_frame$small_area) %>%
    dplyr::select(zoneid, sum) %>%
    dplyr::rename(pop_size = sum,
                  small_area = zoneid)

  # Create population means matrix
  meanxpop <- pop_data %>%
    dplyr::filter(zoneid %in% model_frame$small_area) %>%
    dplyr::select(zoneid, mean) %>%
    dplyr::rename(x = mean,
                  small_area = zoneid)

  # Fit the model
  mod <- eblupBHF(
    formula = model_frame$y ~ model_frame$x,
```

```
    dom = model_frame$small_area,
    meanxpop = meanxpop,
    popnsize = pop_size
  )
  mod
}
```

A.1.6 Frequentist Area-Level

```
freq_area <- function(data, formula, small_area,
                      pop_data, post_strat_data) {
  # Load packages
  library(tidyverse)
  library(sae)

  # Create model frame
  model_frame <- model.frame(formula, data) %>%
    dplyr::mutate(small_area = data[[small_area]])
  colnames(model_frame) <- c("y", "x", "small_area")
  model_frame

  mf <- model.frame(formula, data)

  dir <- post_strat_data %>%
    filter(response %in% colnames(mf)[1],
           province %in% unique(data$province)) %>%
    arrange(subsection)

  # Direct X
  X <- pop_data %>%
    dplyr::filter(zoneid %in% model_frame$small_area) %>%
    dplyr::select(zoneid, mean) %>%
    dplyr::rename(mean_x = mean,
                  small_area = zoneid) %>%
    dplyr::arrange(small_area)

  # Join pop and dir
  dat <- dir %>%
    left_join(X, by = c("subsection" = "small_area"))

  # Fit the model
  mod <- sae::mseFH(formula = dat$est ~ dat$mean_x,
                    vardir = dat$var)
```

```
  mod

}
```

A.1.7 Coeffiicent of Variation Functions

```
hb_CoV <- function(data) {
  # Load packages
  library(tidyverse)
  library(hbsae)

  # Grab CoV
  data$CoV
}

freq_unit_CoV <- function(data, formula, small_area,
                          pop_data, B = 100) {
  # Load packages
  library(tidyverse)
  library(sae)

  # Create empty items for looping
  boots <- list()
  fit <- list()
  mean_df <- list()
  final <- data.frame()

  # Create model frame
  model_frame <- model.frame(formula, data) %>%
    dplyr::mutate(small_area = data[[small_area]])
  colnames(model_frame) <- c("y", "x", "small_area")

  # Nest by small area
  data_nested <- model_frame %>%
    mutate(id = small_area) %>%
    group_by(small_area) %>%
    nest()

  # Bootstrap
  for(i in 1:B){
    for(j in 1:length(unique(model_frame$small_area))) {
      boots[[j]] <- sample_n(
        data_nested[[2]][[j]],
```

```
        size = length(data_nested[[2]][[j]]$y),
        replace = TRUE
      )
      boots_df <- bind_rows(boots)
    }

    fit[[i]] <- freq_unit(boots_df, y ~ x, "id", pop_data)

    mean_df[[i]] <- data.frame(fitted = fit[[i]]$eblup$eblup,
                               subsection = fit[[i]]$eblup$domain)
    if (i %% 50 == 1) {
      print(i)
    }
  }

  # Create final output
  final <- bind_rows(mean_df) %>%
    group_by(subsection) %>%
    summarize(sd = sd(fitted, na.rm = TRUE))
  mean_y <- model_frame %>%
    dplyr::group_by(small_area) %>%
    dplyr::summarise(mean_y = mean(y, na.rm = TRUE))

  COV <- final$sd / mean_y$mean_y
  names(COV) <- final$subsection

  COV
}
```

A.2 Fitting Models

A.2.1 Data Set-up & Preprocessing

```
library(tidyverse)
library(mase)
library(hbsae)
library(sae)
```

```
intwest <- read_csv("data/subsets/df.csv")
tccpop <- read_csv("data/population/tcc_pop.csv")
strata <- read_csv("data/population/strata.csv")
```

```
# Set-up strata:
intwest <- intwest %>%
  mutate(FIAstrat = case_when(
    FIAstrat == "1" ~ "Sampled-Forest",
    FIAstrat == "2" ~ "Sampled-Nonforest",
    FIAstrat == "0" ~ "Sampled-Nonforest",
    FIAstrat == "3" ~ "Sampled-Nonforest",
  ))
```

```
# Filter out subsections that cause errors in computation
no0_subsections <- intwest %>%
  group_by(subsection) %>%
  summarize(mean_y = mean(BIOLIVE_TPA),
            mean_x = mean(nlcd11)) %>%
  filter(mean_y > 0 & mean_x > 0) %>%
  dplyr::select(subsection) %>%
  pull()

intwest_no0 <- intwest %>%
  filter(subsection %in% no0_subsections) %>%
  filter(!(province %in% c("M261", "M334"))) %>%
  filter(!(subsection %in% c("342Fi", "331Kj", "342Dh", "341Dc")))
```

```
# Create list of dataframes
iw <- split(intwest_no0, f = intwest_no0$province)
```

A.2.2 Direct Estimation

```
strata <- strata %>%
  filter(zoneid %in% unique(intwest_no0$subsection)) %>%
  mutate(
    fnf_no_water = case_when(
      fnf_no_water == 1 ~ "Sampled-Forest",
      fnf_no_water == 2 ~ "Sampled-Nonforest"
    )
  )
strata <- strata %>%
  dplyr::select(-zoneprop)

# split strata into list and drop column split by
strata_list <- lapply(
  split(strata, f = strata$zoneid),
```

```
  function(strata) { strata$zoneid <- NULL; strata}
  )

# split subsections into list
subsection_list <- split(intwest_no0,
                         intwest_no0$subsection)


post_strat_bio <- map2(.x = subsection_list,
                       .y = strata_list,
                       .f = postStrat2_bio)
post_strat_ba <- map2(.x = subsection_list,
                      .y = strata_list,
                      .f = postStrat2_ba)
post_strat_voln <- map2(.x = subsection_list,
                        .y = strata_list,
                        .f = postStrat2_voln)
post_strat_cnt <- map2(.x = subsection_list,
                       .y = strata_list,
                       .f = postStrat2_cnt)

results <- data.frame(
  ps_est = c(bind_rows(post_strat_bio)$ps_est,
             bind_rows(post_strat_ba)$ps_est,
             bind_rows(post_strat_cnt)$ps_est,
             bind_rows(post_strat_voln)$ps_est),
  ps_sd = c(bind_rows(post_strat_bio)$sd,
             bind_rows(post_strat_ba)$sd,
             bind_rows(post_strat_cnt)$sd,
             bind_rows(post_strat_voln)$sd),
  subsection = rep(names(post_strat_bio), 4),
  response = c(
    rep("BIOLIVE_TPA", length(post_strat_bio)),
    rep("BALIVE_TPA", length(post_strat_ba)),
    rep("CNTLIVE_TPA", length(post_strat_cnt)),
    rep("VOLNLIVE_TPA", length(post_strat_voln)))
)
```

```
dirmean <- list()
dirmean[1:length(iw)] <-
  lapply(iw,
         direct_estimate,
         response = "BIOLIVE_TPA",
         "subsection")
```

```
dirmean[(length(iw) + 1):(2 * length(iw))] <-
  lapply(iw,
         direct_estimate,
         response = "BALIVE_TPA",
         "subsection")
dirmean[(2*length(iw) + 1):(3*length(iw))] <-
  lapply(iw,
          direct_estimate,
          response = "CNTLIVE_TPA",
          "subsection")
dirmean[(3*length(iw) + 1):(4*length(iw))] <-
  lapply(iw,
          direct_estimate,
          response = "VOLNLIVE_TPA",
          "subsection")

dir <- bind_rows(dirmean) %>%
  dplyr::select(Domain, Direct, CV) %>%
  mutate(cov_dirmean = CV / 100) %>%
  rename(est_dirmean = Direct) %>%
  dplyr::select(-CV) %>%
  mutate(response = c(
    rep("BIOLIVE_TPA", length(post_strat_bio)),
    rep("BALIVE_TPA", length(post_strat_ba)),
    rep("CNTLIVE_TPA", length(post_strat_cnt)),
    rep("VOLNLIVE_TPA", length(post_strat_voln))))

results <- results %>%
  left_join(dir, by = c("subsection" = "Domain",
                        "response" = "response"))
```

```
# Create CoV for ps-estimator
results <- results %>%
  mutate(
    cov_dirps = ps_sd / est_dirmean
  )

# change names and rearrange
results <- results %>%
  rename(est_dirps = ps_est,
         sd_dirps = ps_sd) %>%
  relocate("subsection", "response",
           "est_dirps", "cov_dirps")
```

A.2.3 Model-Based Estimation

```
# set up list for area level models
ps_dat <- results %>%
  dplyr::select(est_dirps, sd_dirps,
                subsection, response) %>%
  dplyr::mutate(var = sd_dirps^2) %>%
  dplyr::rename(est = est_dirps) %>%
  dplyr::select(est, var, subsection, response) %>%
  dplyr::mutate(
    section = str_remove_all(subsection, "[:lower:]"),
    province = str_sub(section, end = -2))

ps_l <- split(ps_dat, list(ps_dat$province))
```

```
# HB Unit
set.seed(1)
bayes_unit <- list()
bayes_unit[1:length(iw)] <- lapply(
  iw,
  hb_unit,
  formula = BIOLIVE_TPA ~ nlcd11,
  small_area = "subsection",
  pop_data = tccpop
)
bayes_unit[(length(iw) + 1):(2 * length(iw))] <-
  lapply(
    iw,
    hb_unit,
    formula = BALIVE_TPA ~ nlcd11,
    small_area = "subsection",
    pop_data = tccpop
  )
bayes_unit[(2 * length(iw) + 1):(3 * length(iw))] <-
  lapply(
    iw,
    hb_unit,
    formula = CNTLIVE_TPA ~ nlcd11,
    small_area = "subsection",
    pop_data = tccpop
  )
bayes_unit[(3 * length(iw) + 1):(4 * length(iw))] <-
  lapply(
    iw,
```

```
    hb_unit,
    formula = VOLNLIVE_TPA ~ nlcd11,
    small_area = "subsection",
    pop_data = tccpop
  )

res_hbu <- data.frame(
  subsection = names(unlist(lapply(bayes_unit, EST))),
response = c(
    rep("BIOLIVE_TPA", length(post_strat_bio)),
    rep("BALIVE_TPA", length(post_strat_ba)),
    rep("CNTLIVE_TPA", length(post_strat_cnt)),
    rep("VOLNLIVE_TPA", length(post_strat_voln))),
  est_hb_unit = unlist(lapply(bayes_unit, EST)),
  cov_hb_unit = unlist(lapply(bayes_unit, hb_CoV))
)

results <- results %>%
  left_join(res_hbu, by = c("subsection" = "subsection",
                            "response" = "response"))
```

```
# HB Area
set.seed(1)
bayes_area <- list()
bayes_area[1:length(iw)] <-
  lapply(
    iw,
    hb_area,
    formula = BIOLIVE_TPA ~ nlcd11,
    small_area = "subsection",
    pop_data = tccpop,
    post_strat_data = ps_dat
  )
bayes_area[(length(iw) + 1):(2 * length(iw))] <-
  lapply(
    iw,
    hb_area,
    formula = BALIVE_TPA ~ nlcd11,
    small_area = "subsection",
    pop_data = tccpop,
    post_strat_data = ps_dat
  )
bayes_area[(2 * length(iw) + 1):(3 * length(iw))] <-
  lapply(
```

```
    iw,
    hb_area,
    formula = CNTLIVE_TPA ~ nlcd11,
    small_area = "subsection",
    pop_data = tccpop,
    post_strat_data = ps_dat
  )
bayes_area[(3 * length(iw) + 1):(4 * length(iw))] <-
  lapply(
    iw,
    hb_area,
    formula = VOLNLIVE_TPA ~ nlcd11,
    small_area = "subsection",
    pop_data = tccpop,
    post_strat_data = ps_dat
  )

aranged_df <- intwest_no0 %>%
  filter(province %in% names(iw)) %>%
  arrange(province) %>%
  arrange(subsection)
hb_area_res <- data.frame(
  subsection = rep(unique(aranged_df$subsection), 4),
  est_hb_area = unlist(lapply(bayes_area, EST)),
  cov_hb_area = unlist(lapply(bayes_area, hb_CoV)),
  response = c(
    rep("BIOLIVE_TPA", length(post_strat_bio)),
    rep("BALIVE_TPA", length(post_strat_ba)),
    rep("CNTLIVE_TPA", length(post_strat_cnt)),
    rep("VOLNLIVE_TPA", length(post_strat_voln))
  )
)

results <- results %>%
  left_join(hb_area_res, by = c("subsection" = "subsection",
                                "response" = "response"))
```

```
# Freq Unit
set.seed(1)
frequnit <- list()
frequnit[1:length(iw)] <-
  lapply(iw,
         freq_unit,
         formula = BIOLIVE_TPA ~ nlcd11,
```

```
          "subsection",
          tccpop)
frequnit[(length(iw) + 1):(2 * length(iw))] <-
  lapply(iw,
          freq_unit,
          formula = BALIVE_TPA ~ nlcd11,
          "subsection",
          tccpop)
frequnit[(2 * length(iw) + 1):(3 * length(iw))] <-
  lapply(iw,
          freq_unit,
          formula = CNTLIVE_TPA ~ nlcd11,
          "subsection",
          tccpop)
frequnit[(3 * length(iw) + 1):(4 * length(iw))] <-
  lapply(iw,
          freq_unit,
          formula = VOLNLIVE_TPA ~ nlcd11,
          "subsection",
          tccpop)

frequnit_list <- list()
  for(i in 1:(4 * length(iw))) {
    frequnit_list[[i]] <- frequnit[[i]]$eblup
  }

frequnit_df <- frequnit_list %>%
  bind_rows() %>%
  mutate(response = c(
    rep("BIOLIVE_TPA", length(post_strat_bio)),
    rep("BALIVE_TPA", length(post_strat_ba)),
    rep("CNTLIVE_TPA", length(post_strat_cnt)),
    rep("VOLNLIVE_TPA", length(post_strat_voln)))) %>%
  rename(
    subsection = domain,
    est_freq_unit = eblup
  ) %>%
  dplyr::select(-sampsize)

# Coef of variation
frequnitcov <- c()
frequnitcov <- c(
  Reduce(c,
       lapply(iw,
```

```
        freq_unit_CoV,
        formula = BIOLIVE_TPA ~ nlcd11,
        small_area = "subsection",
        pop_data = tccpop,
        B = 500)),
  Reduce(c,
        lapply(iw,
        freq_unit_CoV,
        formula = BALIVE_TPA ~ nlcd11,
        small_area = "subsection",
        pop_data = tccpop,
        B = 500)),
  Reduce(c,
        lapply(iw,
        freq_unit_CoV,
        formula = CNTLIVE_TPA ~ nlcd11,
        small_area = "subsection",
        pop_data = tccpop,
        B = 500)),
  Reduce(c,
        lapply(iw,
        freq_unit_CoV,
        formula = VOLNLIVE_TPA ~ nlcd11,
        small_area = "subsection",
        pop_data = tccpop,
        B = 500))
)

frequnitcov_df <- data.frame(
  cov_freq_unit = frequnitcov,
  subsection = names(frequnitcov),
  response = c(
    rep("BIOLIVE_TPA", length(post_strat_bio)),
    rep("BALIVE_TPA", length(post_strat_ba)),
    rep("CNTLIVE_TPA", length(post_strat_cnt)),
    rep("VOLNLIVE_TPA", length(post_strat_voln))
  )
)

frequnit_df <- frequnit_df %>%
  left_join(frequnitcov_df,
            by = c("subsection" = "subsection",
                   "response" = "response"))
```

```
results <- results %>%
  left_join(frequnit_df,
            by = c("subsection" = "subsection",
                   "response" = "response"))
```

```
# Freq Area
set.seed(1)
freqarea <- list()
freqarea[1:length(iw)] <-
  lapply(
    iw,
    freq_area,
    formula = BIOLIVE_TPA ~ nlcd11,
    "subsection",
    tccpop,
    post_strat_data = ps_dat
)
freqarea[(length(iw) + 1):(2 * length(iw))] <-
  lapply(
    iw,
    freq_area,
    formula = BALIVE_TPA ~ nlcd11,
    "subsection",
    tccpop,
    post_strat_data = ps_dat
)
freqarea[(2 * length(iw) + 1):(3 * length(iw))] <-
  lapply(
    iw,
    freq_area,
    formula = CNTLIVE_TPA ~ nlcd11,
    "subsection",
    tccpop,
    post_strat_data = ps_dat
)
freqarea[(3 * length(iw) + 1):(4 * length(iw))] <-
  lapply(
    iw,
    freq_area,
    formula = VOLNLIVE_TPA ~ nlcd11,
    "subsection",
    tccpop,
    post_strat_data = ps_dat
)
```

```
freqarea_list <- list()
for (i in 1:(4 * length(iw))) {
  freqarea_list[[i]] <- freqarea[[i]]$est$eblup
}
freqarea_cov_list <- list()
for (i in 1:(4 * length(iw))) {
  freqarea_cov_list[[i]] <- sqrt(freqarea[[i]]$mse)
}

freq_area_res <- data.frame(
  subsection = rep(unique(aranged_df$subsection), 4),
  est_freq_area = unlist(freqarea_list),
  se_freq_area = unlist(freqarea_cov_list),
  response = c(
    rep("BIOLIVE_TPA", length(post_strat_bio)),
    rep("BALIVE_TPA", length(post_strat_ba)),
    rep("CNTLIVE_TPA", length(post_strat_cnt)),
    rep("VOLNLIVE_TPA", length(post_strat_voln))
  )
)
results <- results %>%
  left_join(freq_area_res,
            by = c("subsection" = "subsection",
                   "response" = "response"))
results <- results %>%
  mutate(cov_freq_area = se_freq_area / est_dirmean)
```

A.2.4 Writing Data Files & Pivoting to Tidy Format

```
write.csv(results,
          "data/results/final_results.csv")
saveRDS(results,
        file = "data/results/final_results.rds")
```

```
estimates_long <- results %>%
  pivot_longer(cols = c("est_hb_unit", "est_hb_area",
                        "est_freq_unit", "est_freq_area",
                        "est_dirmean", "est_dirps"),
               names_to = "estimator",
               values_to = "estimate") %>%
  dplyr::select(-cov_hb_unit, -cov_hb_area,
                -cov_freq_unit, -cov_freq_area,
```

```
                -cov_dirmean, -cov_dirps) %>%
  mutate(estimator = stringr::str_sub(estimator, start = 5))

cov_long <- results %>%
  pivot_longer(cols = c("cov_hb_unit", "cov_hb_area",
                        "cov_freq_unit", "cov_freq_area",
                        "cov_dirmean", "cov_dirps"),
               names_to = "estimator",
               values_to = "cov") %>%
  dplyr::select(-est_hb_unit, -est_hb_area,
                -est_freq_unit, -est_freq_area,
                -est_dirmean, -est_dirps) %>%
  mutate(estimator = stringr::str_sub(estimator, start = 5))

final_results_long <- estimates_long %>%
  full_join(cov_long) %>%
  mutate(
    section = str_remove_all(subsection, "[:lower:]"),
    province = str_sub(section, end = -2)
  )

final_results_long <- final_results_long %>%
  dplyr::select(-sd_dirps, -se_freq_area)
```

```
write.csv(final_results_long,
          "data/results/final_results_long.csv")
saveRDS(final_results_long,
        file = "data/results/final_results_long.rds")
```

References

Bayes, T. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late rev. Mr. Bayes, FRS communicated by mr. Price, in a letter to john canton, AMFR s. *Philosophical Transactions of the Royal Society of London*, (53), 370–418.

Boonstra, H. J. (2012). *Hbsae: Hierarchical bayesian small area estimation*. Retrieved from `https://CRAN.R-project.org/package=hbsae`

Bray, A. (2020). Math 392: Mathematical statistics.

Breidenbach, J., & Astrup, R. (2012). Small area estimation of forest attributes in the norwegian national forest inventory. *European Journal of Forest Research*, *131*(4), 1255–1267.

Efron, B. (1992). Bootstrap methods: Another look at the jackknife. In *Breakthroughs in statistics* (pp. 569–593). Springer.

FIA. (2020). Forest inventory and analysis national program. *What is FIA?* Retrieved from `https://www.fia.fs.fed.us/about/about_us/`

Hidiroglou, M., & You, Y. (2016). Comparison of unit level and area level small area estimators, *42*, 41–61.

Homer, C. (2015, November). Completion of the 2011 national land cover database for the conterminous united states – representing a decade of land cover change information. *EPA*. Environmental Protection Agency. Retrieved from `https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NERL&dirEntryId=309950`

Hooten, M. B., & Wikle, C. K. (2008). A hierarchical bayesian non-linear spatio-temporal model for the spread of invasive species with application to the eurasian collared-dove. *Environmental and Ecological Statistics*, *15*(1), 59–70.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*, 663–685.

Iriawan, N., & Yasmirullah, S. (2019). An economic growth model using hierarchical bayesian method. http://doi.org/10.5772/intechopen.88650

McConville, K. S., Moisen, G. G., & Frescino, T. S. (2020). A tutorial on model-assisted estimation with application to forest inventory. *Forests*, *11*(2).

McConville, K., Tang, B., Zhu, G., Cheung, S., & Li, S. (2018). *Mase: Model-assisted survey estimation*. Retrieved from `https://cran.r-project.org/package=mase`

McElreath, R. (2020). *Statistical rethinking: A bayesian course with examples in r and STAN*.

Molina, I., & Marhuenda, Y. (2015). sae: An R package for small area estimation. *The R Journal*, *7*(1), 81–98. Retrieved from `https://journal.r-project.org/archive/2015/RJ-2015-007/RJ-2015-007.pdf`

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from `https://www.R-project.org/`

Rao, J. N. (2014). Small-area estimation. *Wiley StatsRef: Statistics Reference Online*.

Tojtovska, B., Ribarski, P., & Ljubic, A. (2019). Application of hierarchical bayesian model in ophtalmological study. In S. Gievska & G. Madjarov (Eds.), *ICT innovations 2019. Big data processing and mining* (pp. 109–120). Cham: Springer International Publishing.

Ver Planck, N. R., Finley, A. O., & Huff, E. S. (2017). Hierarchical bayesian models for small area estimation of county-level private forest landowner population. *Canadian Journal of Forest Research*, *47*(12), 1577–1589.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. http://doi.org/10.21105/joss.01686