

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

<b>Introduction</b> . . . . .	<b>1</b>
<b>Chapter 1: Context</b> . . . . .	<b>3</b>
<b>Chapter 2: Overview</b> . . . . .	<b>11</b>
<b>Chapter 3: Data</b> . . . . .	<b>13</b>
3.1 The Forest Inventory & Analysis Program . . . . .	13
3.2 The Interior West . . . . .	14
3.3 Our Data: Specifics . . . . .	15
3.4 Data Structure & Hierarchy . . . . .	20
<b>Chapter 4: Methods</b> . . . . .	<b>23</b>
4.1 Direct Estimation . . . . .	24
4.1.1 The Horvitz-Thompson Estimator . . . . .	24
4.1.2 The Post-Stratified Estimator . . . . .	24
4.2 Implicit Model Based Indirect Estimation . . . . .	25
4.2.1 Using Area-level Auxiliary Data . . . . .	26
4.2.2 Using Unit-level Auxiliary Data . . . . .	26
4.3 Explicit Model Based Indirect Estimation . . . . .	26
4.3.1 The Unit-level EBLUP . . . . .	27
4.3.2 The Area-level EBLUP . . . . .	27
4.4 A Hierarchical Bayesian Approach . . . . .	27
<b>Chapter 5: Results</b> . . . . .	<b>29</b>
5.1 Modeling Overview . . . . .	29
5.2 Unit-level Models . . . . .	30
<b>Chapter 6: Discussion and Conclusion</b> . . . . .	<b>33</b>
<b>Appendix A: The First Appendix</b> . . . . .	<b>35</b>
<b>Appendix B: The Second Appendix, for Fun</b> . . . . .	<b>37</b>
<b>References</b> . . . . .	<b>39</b>



# List of Tables

3.1	Relevant Glimpse of Data . . . . .	15
3.2	Summary Statistics of Relevant Variables . . . . .	19
3.3	Analysis of Variance Model (Biomass Response) . . . . .	21



# List of Figures

1.1	An ecological province . . . . .	4
1.2	The mean as a direct estimator . . . . .	5
1.3	The post-stratified direct estimator . . . . .	6
1.4	The unit-level hierarchical Bayesian model . . . . .	7
1.5	The area-level hierarchical Bayesian model . . . . .	8
3.1	The Interior West region of the United States . . . . .	14
3.2	The Northern Rocky Forest colored by eco-section . . . . .	16
3.3	Mean basal area in Interior West eco-subsections . . . . .	17
3.4	Mean biomass in Interior West eco-subsections . . . . .	17
3.5	Mean tree count per acre in Interior West eco-subsections . . . . .	18
3.6	Mean net volume in Interior West eco-subsections . . . . .	18
3.7	Total canopy cover in the M333 eco-province and Interior West . . . . .	19
3.8	The nested data structure of the Interior West . . . . .	20
5.1	Unit-level correlation . . . . .	29
5.2	Area-level correlation . . . . .	30
5.3	Direct and model-based estimates for the unit-level model . . . . .	31
5.4	Direct and model-based coefficients of variation for the unit-level model	32



# **Abstract**

The preface pretty much says it all.

Second paragraph of abstract starts here.



# Dedication

To my family.



# **Introduction**

This is the introduction to my thesis.



# 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.

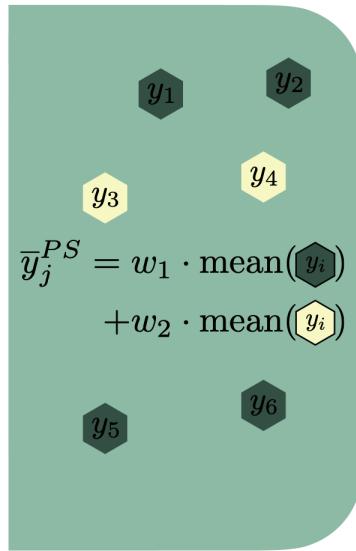


Figure 1.3: The post-stratified direct estimator in the seafoam green eco-subsection. This estimator only relies on sampled areas (hexagons) within the eco-subsection. This estimator takes into account whether the sampled areas are forested (green) or non-forested (beige) areas. It then weights our estimate 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.

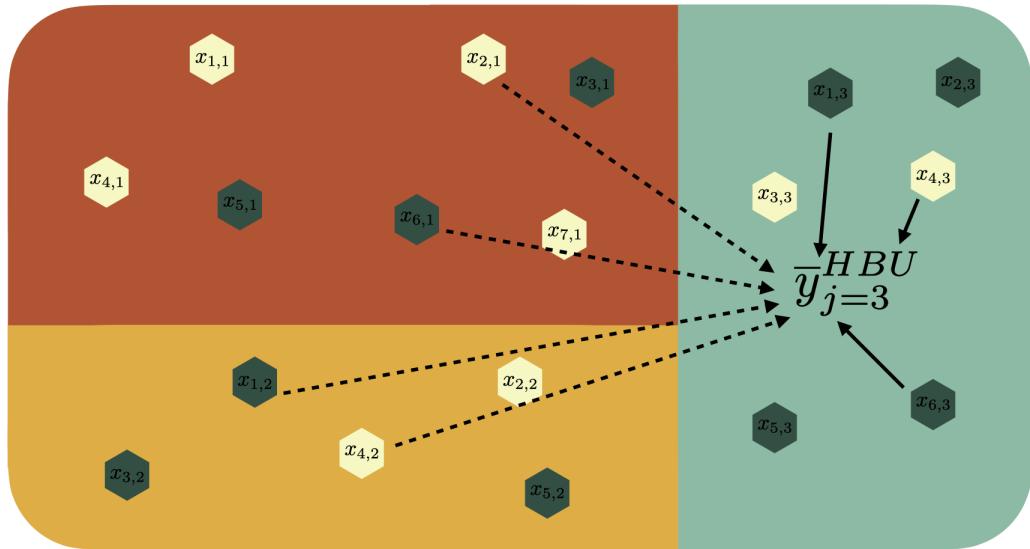


Figure 1.4: The unit-level hierarchical Bayesian model producing an estimate in our seafoam green eco-subsection. This estimator relies on auxiliary information from all eco-subsections within a given eco-province to produce estimates as shown by the arrows. Notably, more information is used in the eco-subsection 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.

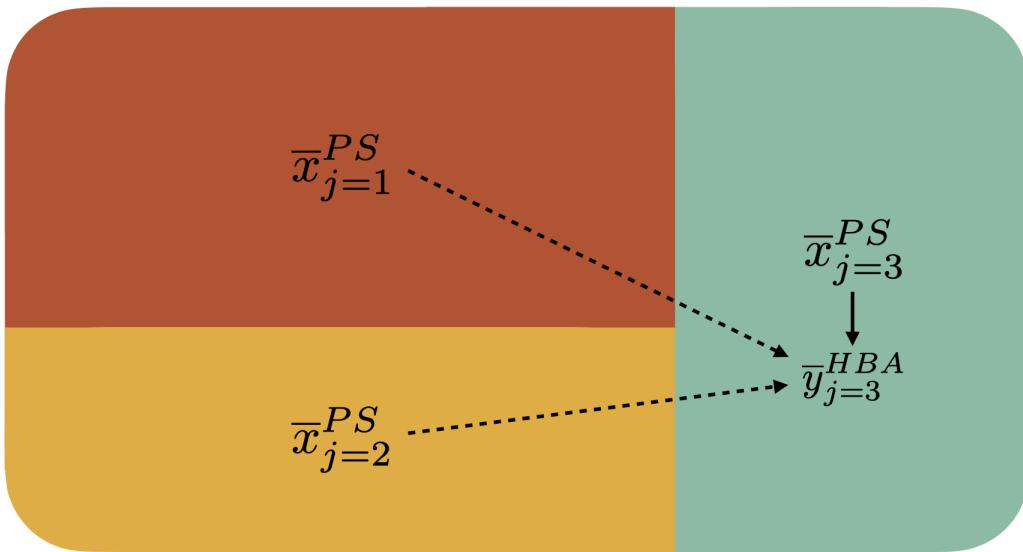


Figure 1.5: The area-level hierarchical Bayesian model producing an estimate in our seafoam green eco-subsection. This estimator relies on auxiliary information from all eco-subsections within a given eco-province to produce estimates as shown by the arrows. Notably, more information is used in the eco-subsection 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  $\mu_{y_j}$ :

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

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:

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

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:

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

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**

## **Overview**



# Chapter 3

## Data

### 3.1 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 (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.

## 3.2 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 3.1: 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.

### 3.3 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 3.1: 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:

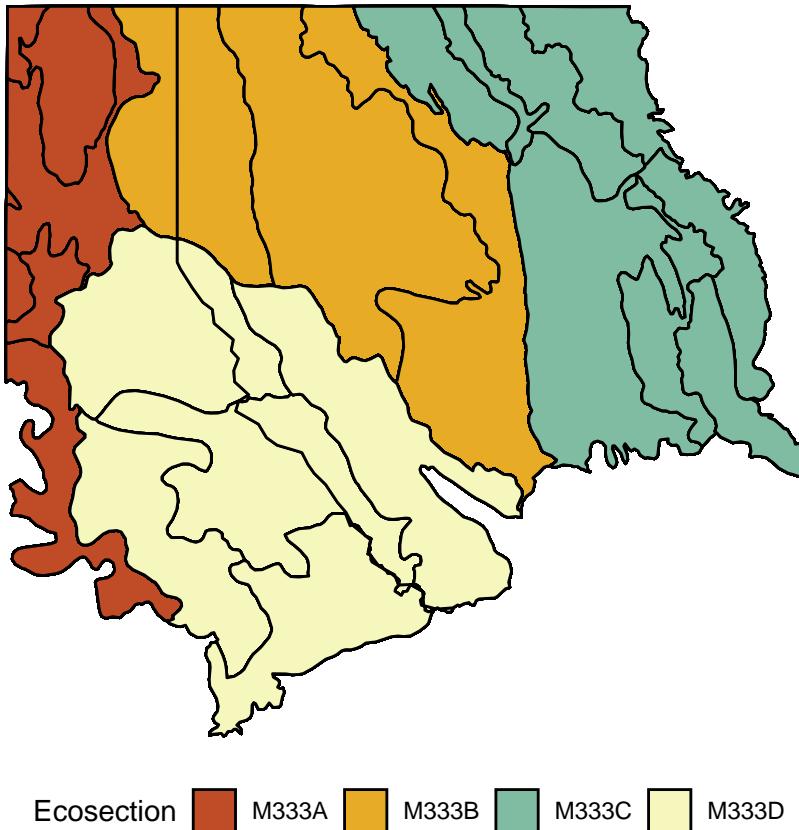


Figure 3.2: The Northern Rocky Forest colored by eco-section

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 ( $\text{ft}^3$ ). These variables are coded as `BALIVE_TPA`, `CNTLIVE_TPA`, `BIO LIVE_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 3.3: Mean basal area in Interior West eco-subsections

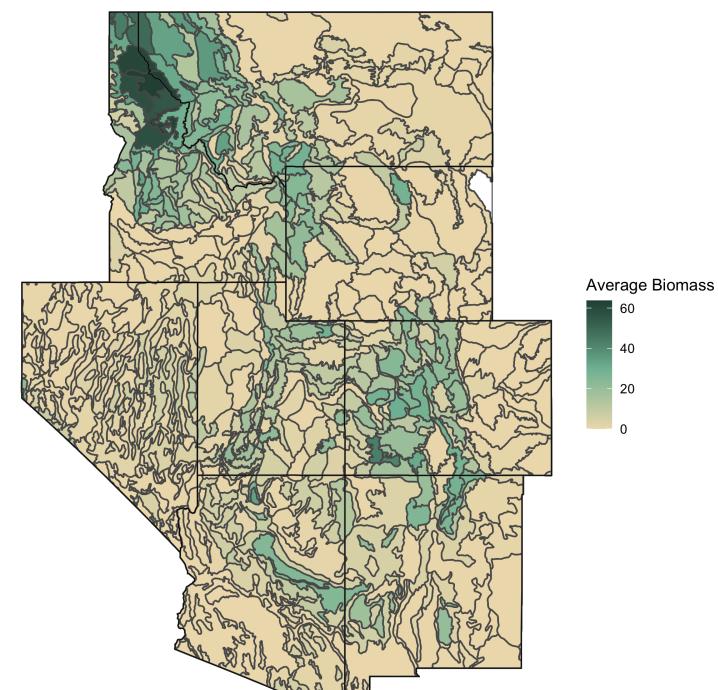


Figure 3.4: Mean biomass in Interior West eco-subsections



Figure 3.5: Mean tree count per acre in Interior West eco-subsections



Figure 3.6: 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.

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



Figure 3.7: 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 3.2: 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.

## 3.4 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:

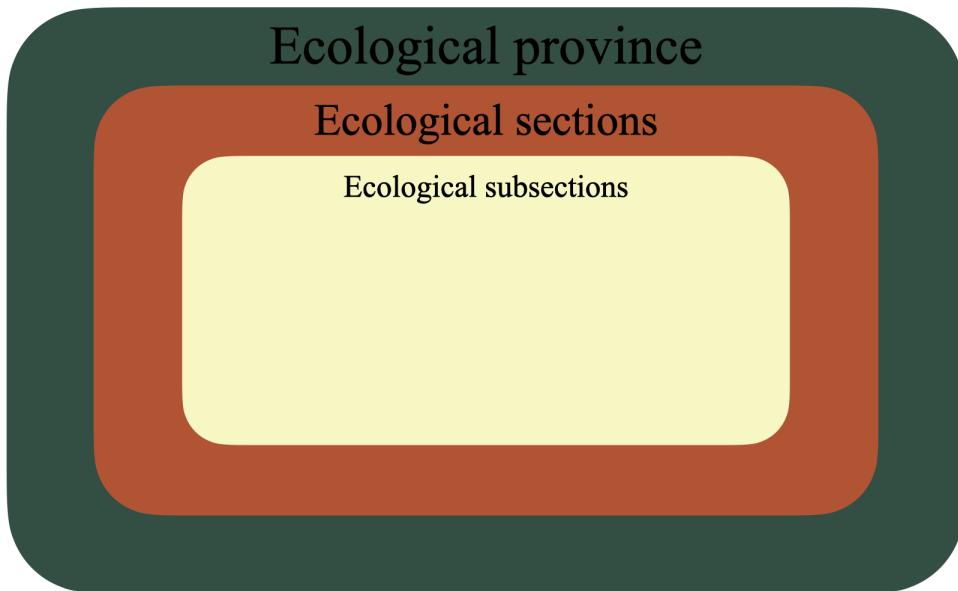


Figure 3.8: The nested data structure of the Interior West. The green area is the eco-province which is divided into eco-sections (red area) which is in turn divided into eco-subsections (beige area).

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 3.3: Analysis of Variance Model (Biomass Response)

term	df	sumsq	meansq	statistic	p.value
province	13	6512457	500958	2921	0
section	54	967169	17911	104.4	0
subsection	412	2247965	5456	31.82	0
Residuals	85605	14679154	171.5	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 4

## 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 across small areas and produce estimates with implicit use of a model. 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 use mixed or hierarchical models.

This thesis explores the application of two less commonly applied model-based estimators, the hierarchical Bayesian unit-level model and the hierarchical Bayesian area-level model. We compare these novel Bayesian models to the frequentist EBLUP unit- and area-level models and two common direct estimators, the 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 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 study variable  $y$ , such as trees per acre or biomass, in a small area. So, let  $\mu_{y_j}$  be the population mean of the study variable in eco-subsection  $j$  in the Interior West. To denote the estimate produced of  $\mu_{y_j}$  we will use  $\hat{\mu}_{y_j}$  with a superscript denoting which estimator is being used. In summary, each of our estimators aims to estimate  $\mu_{y_j}$ , the population mean of some study variable in the  $j$ th eco-subsection and our estimated value is denoted by  $\hat{\mu}_{y_j}$ .

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  $s_j$ .

## 4.1 Direct Estimation

There are two direct estimators that we will explore 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 will now explore both of these estimators in depth.

### 4.1.1 The Horvitz-Thompson Estimator

What might be the most intuitive approach to estimating the population mean  $\mu_{y_j}$  is taking the sample mean, i.e. using the “Horvitz-Thompson estimator” as foresters like to say. This estimator can be expressed as follows:

$$\hat{\mu}_{y_j}^{HT} = \frac{1}{n_j} \sum_{i \in s_j} y_i \quad (4.1)$$

The variance estimate,  $S$ , for Horvitz-Thompson estimator is calculated by:

$$S_j^2 = \frac{1}{n_j - 1} \sum_{i \in s_j} \left( y_i - \hat{\mu}_{y_j}^{HT} \right)^2 \quad (4.2)$$

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  $i$ th 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 has high variance relative to other estimators we will discuss and it is biased unless we have sampled the correct proportion of forested areas. The post-stratified estimator begins to address both the bias and variance of the Horvitz-Thompson estimator.

### 4.1.2 The Post-Stratified Estimator

The post-stratified estimator is very similar to the Horvitz-Thompson estimator however, as stated above, it addresses bias and variance that occurred from using the Horvitz-Thompson estimator. While decreasing bias and 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:

$$\hat{\mu}_{y_j}^{PS} = \sum_{k=1}^2 w_k \cdot \hat{\mu}_{y_{j,k}}^{HT} \quad (4.3)$$

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 design-unbiased estimate of  $\mu_{y_j}$ .

The post-stratified estimator is a great alternative to the Horvitz-Thompson estimator when auxiliary information 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.

## 4.2 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. First we discuss a common estimator used when area-level auxiliary data is available and then a common estimator for unit-level auxiliary data.

### 4.2.1 Using Area-level Auxiliary Data

When we choose to use area-level auxiliary data and implicit model based indirect estimation, Rao (2014) suggests the regression-synthetic estimator. Here, in order to estimate our parameter of interest,  $\mu_{y_j}$ , we fit an ordinary least squares linear regression on our  $p$  auxiliary variables over each small area in our large domain:

$$\hat{\mu}_{y_j}^{REG} = \hat{\beta}_0 + \hat{\beta}_1 x_{j,1} + \cdots + \hat{\beta}_p x_{j,p} \quad (4.4)$$

We denote this estimator with *REG* rather than a seemingly more sensible notation of *RSE* in order to avoid confusion with the residual standard errors of a model. We note that the  $\hat{\beta}$ 's are calculated through standard methods of ordinary least squares and in the case of our analysis  $p = 1$ . Further, it is important to recognize that in the case of the regression-synthetic estimator each small area receives the same  $\hat{\beta}$ 's. This is the realization of the requirement that the auxiliary information must have the same effect on each small area.

### 4.2.2 Using Unit-level Auxiliary Data

An approach to implicit model based indirect estimation when unit-level auxiliary information is available is to use the

## 4.3 Explicit Model Based Indirect Estimation

Often times, implicit model based indirect estimators still do not reduce variance enough to give us estimates with reasonable standard errors. 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

### 4.3.1 The Unit-level EBLUP

### 4.3.2 The Area-level EBLUP

$$Y_{i,j} = \vec{X}_{i,j}^T \vec{\beta} + \nu_j + \epsilon_{i,j} \quad (4.5)$$

where  $Y_{i,j}$  is the response variable,  $\vec{\beta}$  is the vector of coefficients of fixed-effect predictors,  $\vec{X}_{i,j}^T$  is the vector of fixed-effect predictors,  $\nu_j$  is the random-effects term. Note that this is a varying intercepts model where the intercept can change based on the group that the observation is in, however the coefficient estimates do not vary between groups. Another important aspect of the model is the assumption of normally distributed errors and random effects:

$$\nu \sim N(0, \sigma_\nu^2), \quad (4.6)$$

$$\epsilon \sim N(0, \sigma_\epsilon^2) \quad (4.7)$$

## 4.4 A Hierarchical Bayesian Approach

(still needs revised and expanded)

While we have explored the frequentist version of a hierarchical model, this thesis primarily studies *Bayesian* hierarchical models and compares their performance to their frequentist counterparts. With a Bayesian hierarchical 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 data (likelihood) and prior distributions. This allows us to use Bayes' Theorem in order to get our posterior. Similarly to the frequentist counterpart, we can specify the varying-intercepts hierarchical Bayesian model as follows:

$$Y_{i,j} \sim N(\nu_j + \vec{\beta} \vec{X}_i, \sigma^2) \quad (4.8)$$

$$\nu_j \sim N(\mu_\nu, \sigma_\nu^2) \quad (4.9)$$

$$\mu_\nu \sim N(a, b) \quad (4.10)$$

$$\sigma \sim \text{Inv-Gamma}(c, d) \quad (4.11)$$

In this model, we have the response variable,  $Y$ , which is modeled to have a Gaussian posterior distribution with mean  $\nu_j + \vec{\beta} \vec{X}_i$  which can change intercept based on the level that a given observation is in. Note that  $\mu_\nu$  is given a hyperprior distribution where  $a$  and  $b$  are numbers that often specify a weakly informative or uninformative prior. Often we will set  $a = 0$  and  $b$  equal to some large number to specify a very small amount of prior information. Also, we give the within-area variance a regularizing prior with the Inverse Gamma.



# Chapter 5

## Results

### 5.1 Modeling Overview

We explore both unit- and area-level models in this thesis, where unit-level models fit the model to the plot (unit) level data, and the area-level models fit to data which has been aggregated to the ecosubsection (area) level. These models types each have their own costs and benefits, and while we lose some data structure with the area-level estimates we gain a large amount of precision. We can see this when looking at the correlation between the predictor `nlcd11` and one of our response variables, `BIOLIVE_TPA`, at both the unit- and area-levels with ordinary least squares regression lines fit to the data:

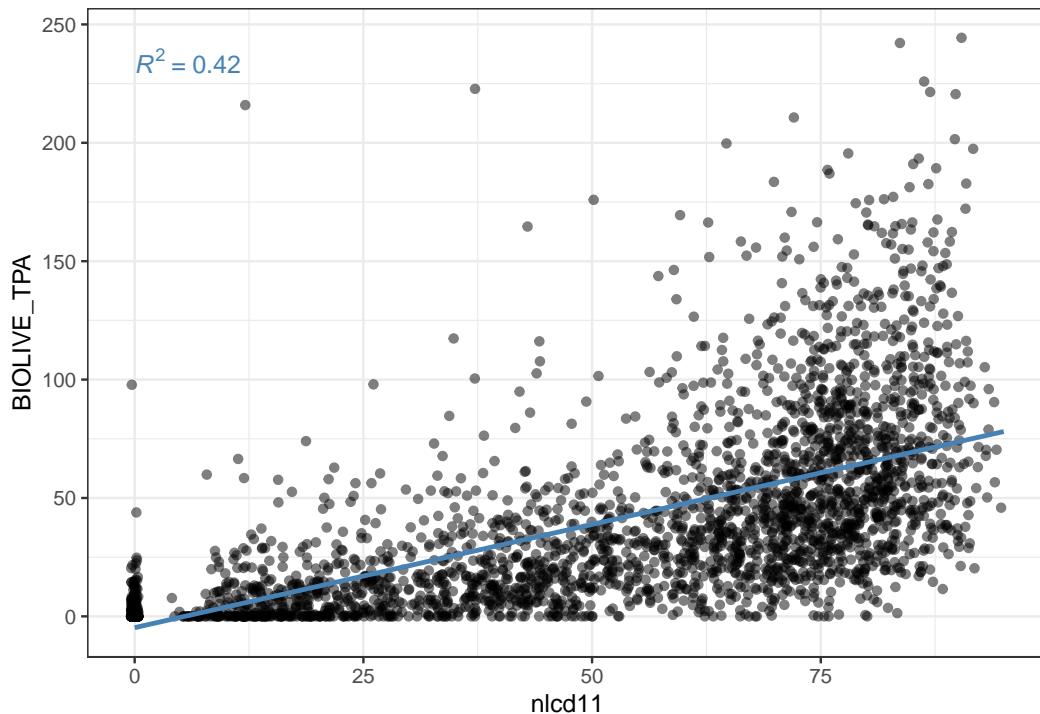


Figure 5.1: Unit-level correlation

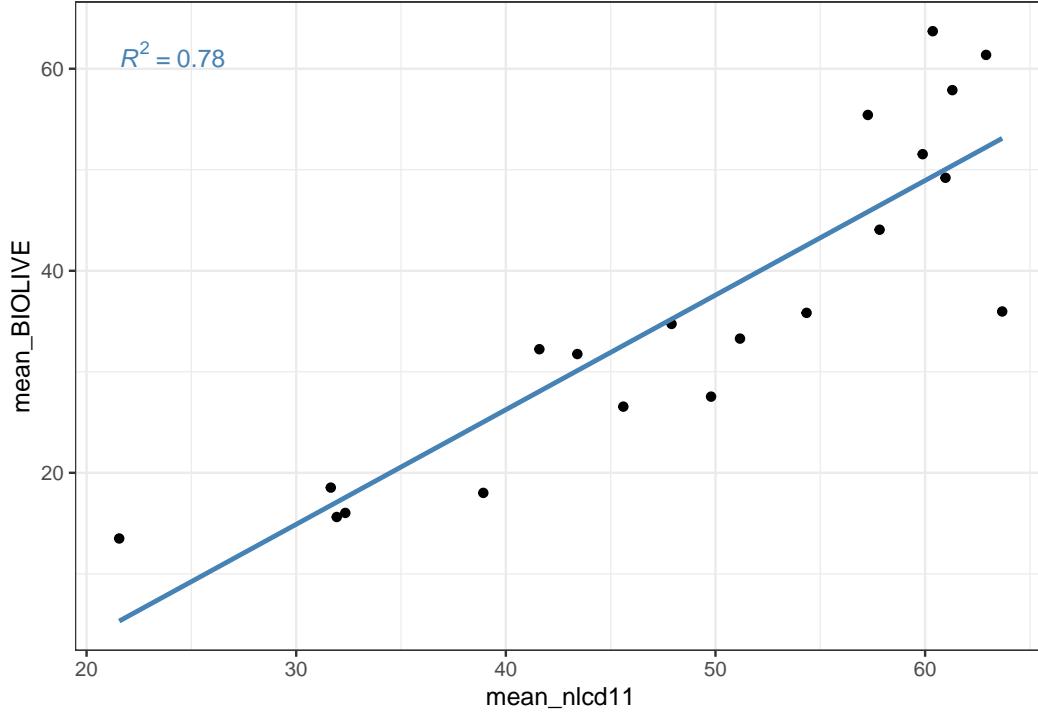


Figure 5.2: Area-level correlation

Notably, the  $R^2$  value for the area-level simple linear regression is much higher than the  $R^2$  value at the unit-level. This is of course compromised by the number of data points ( $n_{area} = 20$ ,  $n_{unit} = 3003$ ). Also, fitting a polynomial regression curve to the unit level data hardly improves the fit ( $R^2 = 0.44$ ).

We, however, are not fitting simple linear regressions. In this chapter, we explore the benefits of Bayesian hierarchical models which use varying-slopes to lower the variance in our estimates at the cost of a small amount of bias.

## 5.2 Unit-level Models

At the unit-level, the small area estimates for each ecosubsection are made by post-aggregation of the plot level output of our model. We fit these models using varying slopes model, which can be written as:

$$\begin{aligned} Y_i &\sim N(\alpha_j + \vec{\beta} \vec{X}_i, \sigma^2) \\ \alpha_j &\sim N(\mu_\alpha, \sigma_\alpha^2) \\ \mu_\alpha &\sim N(a, b) \end{aligned}$$

Here, we have  $Y_i$ , our response variable (BIOLIVE\_TPA), which is modeled to have a Gaussian posterior distribution with mean  $\alpha_j + \vec{\beta} \vec{X}_i$  which can change intercept based on the level that a given observation is in. Note that we are predicting  $Y$  at the unit-level, so we compute  $Y_i$  for every plot in the Northern Rocky Forest, and we allow  $\alpha_j$ , the intercept, to vary over each of the 20 ecosubsections within the Northern

Rocky Forest. Then, we must aggregate our result by taking the mean of our  $Y_i$ 's in each small area. After fitting this model and performing the aggregation, we can look at the estimates of the mean biomass predicted by the model compared to the direct estimator:

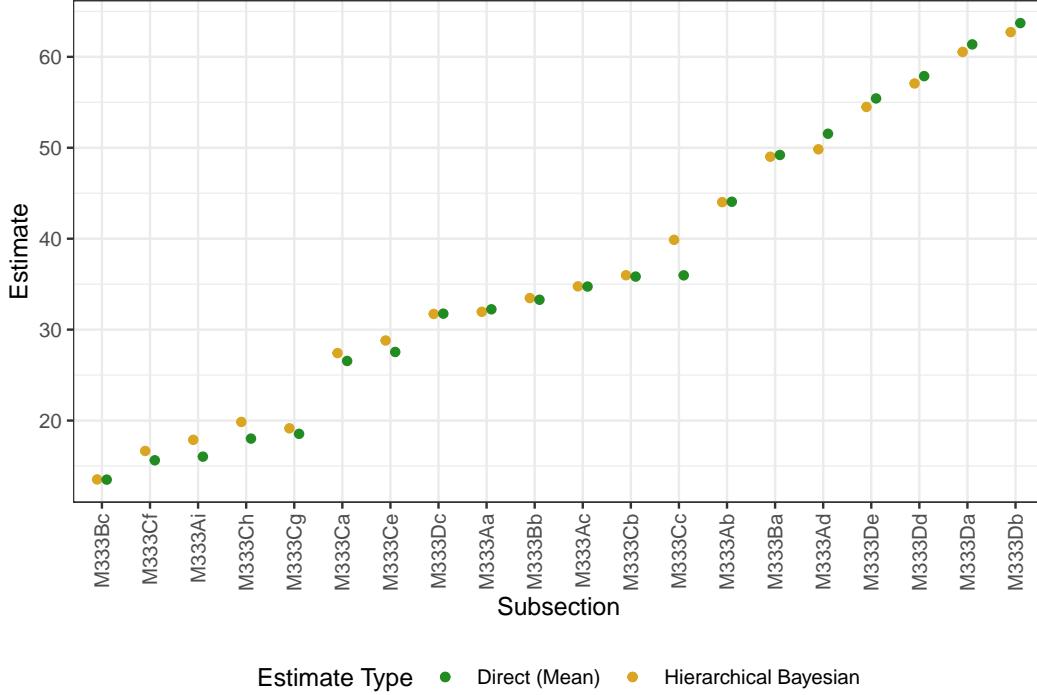


Figure 5.3: Direct and model-based estimates for the unit-level model

These estimates make sense in the context of hierarchical Bayesian modeling because we can see the shrinkage of the estimates towards the overall mean. We also see more shrinkage in ecosubsections which have less plots, particularly M333Cc ( $n_j = 28$ ), M333Ai ( $n_j = 38$ ), and M333Ad ( $n_j = 26$ ). This is again consistent with our intuition as small areas with less data should rely more heavily on the overall mean biomass level of the Northern Rocky Forest.

We can also begin to look at the increase in precision which is gained from this unit-level hierarchical Bayesian model by examining the coefficient of variation for the model and the direct estimator in each ecosubsection. For the direct estimator, the coefficient of variation of a certain ecosubsection  $j$  is defined as

$$CV_{\text{direct}} = \frac{\sqrt{\text{var}(Y_{i,j})}}{\text{mean}(Y_{i,j})} \quad (5.1)$$

where  $Y_{i,j}$  considers all  $i = 1, \dots, n_j$  units in the  $j$ th ecosubsection. Similarly, for the model-based estimator we define the coefficient of variation as

$$CV_{\text{model}} = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^{n_j} (Y_{i,j} - \hat{Y}_{i,j})^2}}{\text{mean}(Y_{i,j})} \quad (5.2)$$

Note that the numerator is now the root mean squared error of the  $j$ th ecosubsection. This is equivalent to taking the square root of the variance as we did in the direct estimator's coefficient of variation, given that our model perfectly meets our modeling assumptions. Knowing that this will never perfectly be the case, we take the root mean squared error to get a more realistic estimate. Now, we can visualize this statistic for each ecosubsection:

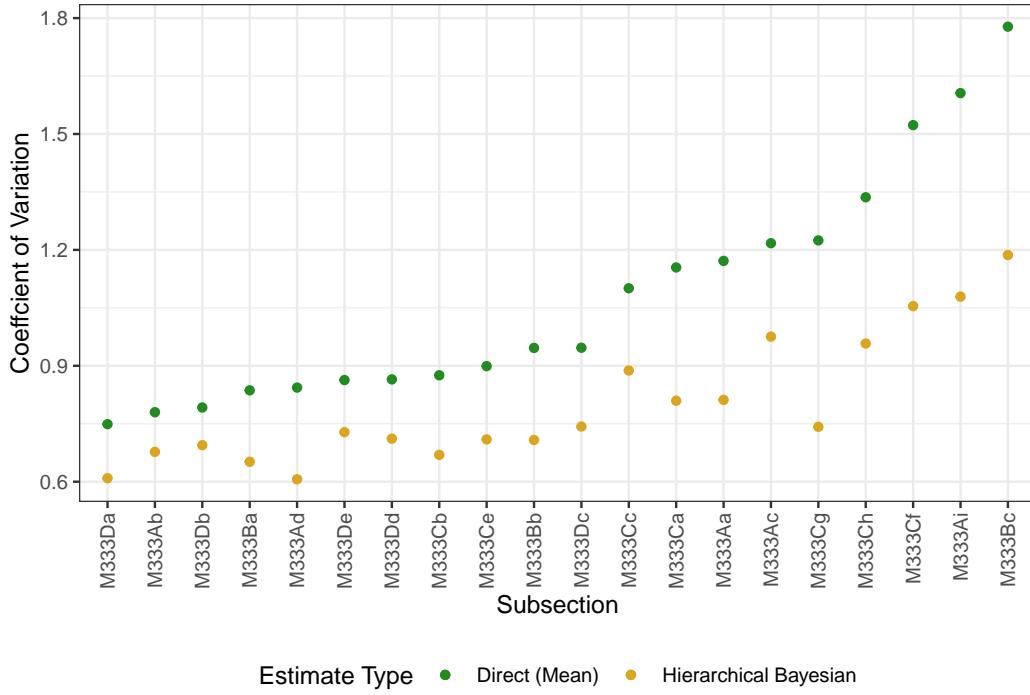


Figure 5.4: Direct and model-based coefficients of variation for the unit-level model

We see reductions in every coefficient of variation from the direct estimator to our model-based approach, with an average reduction in of 24.17%. However, the variation we see is still much larger than wanted, with the ecosubsection with the lowest coefficient of variation just over 0.6 and the overall coefficient of variation of the model at a value of 0.76. These large coefficients of variation indicate that even though we were able to reduce the variance in the estimate by an average of 24.17%, the will not perform well enough to be used as a reliable predictor of average biomass.

# Chapter 6

## Discussion and Conclusion

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

### More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.



# Appendix A

## The First Appendix

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

### In the main Rmd file

```
# This chunk ensures that the thesisdown package is
# installed and loaded. This thesisdown package includes
# the template files for the thesis.
if (!require(remotes)) {
  if (params$`Install needed packages for {thesisdown}`) {
    install.packages("remotes", repos = "https://cran.rstudio.com")
  } else {
    stop(
      paste('You need to run install.packages("remotes")',
            "first in the Console."))
  }
}
if (!require(thesisdown)) {
  if (params$`Install needed packages for {thesisdown}`) {
    remotes::install_github("ismayc/thesisdown")
  } else {
    stop(
      paste(
        "You need to run",
        'remotes::install_github("ismayc/thesisdown")',
        "first in the Console."
      )
    )
  }
}
library(thesisdown)
```

```
# Set how wide the R output will go
options(width = 70)
```

In Chapter ??:

## **Appendix B**

### **The Second Appendix, for Fun**



# References

- FIA. (2020). Forest inventory and analysis national program. *What is FIA?* Retrieved from [https://www.fia.fs.fed.us/about/about\\_us/](https://www.fia.fs.fed.us/about/about_us/)
- 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](https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NERL&dirEntryId=309950)
- McConville, K. S., Moisen, G. G., & Frescino, T. S. (2020). A tutorial on model-assisted estimation with application to forest inventory. *Forests*, 11(2), 244.
- Rao, J. N. (2014). Small-area estimation. *Wiley StatsRef: Statistics Reference Online*, 1–8.
- 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.