Hierarchical Bayesian Modeling 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.

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. 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) or 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, it introduces high variance and a lack of precision necessary for 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. But what is 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:

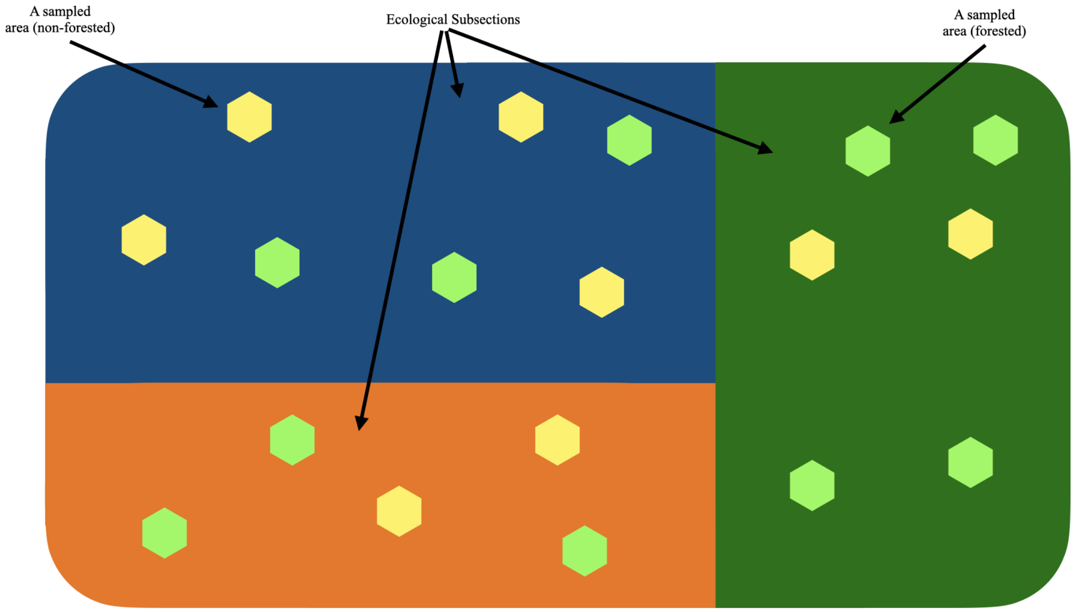

Figure 1.1: An Ecological Province

We are interested in performing inference at the sub-population level, and in Figure 1.1 these sub-populations are represented by the blue, orange, and 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 mean of a variable and the post-stratified estimate of a variable. The post-stratified estimate is similar to the mean, however it accounts for under- and over-sampling of forested areas in a given sub-population. These estimates 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 green sub-population j from Figure 1.1.

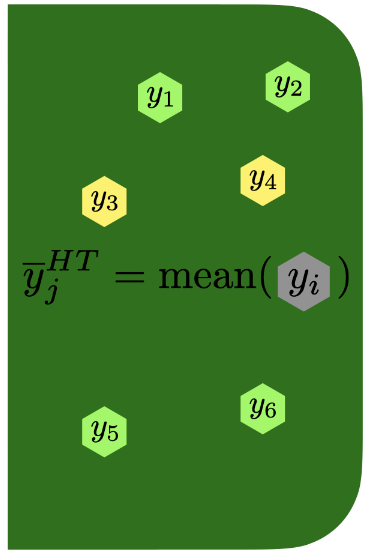

Figure 1.2: The mean as a direct estimator

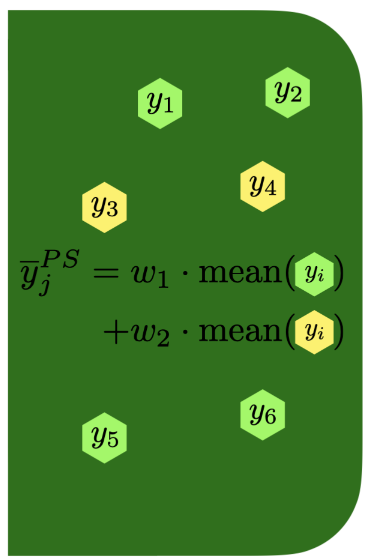

Figure 1.3: The post-stratified direct estimator

The second kind of estimator, indirect estimators with implicit models, rely on data outside of the area of interest to produce their estimate, 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 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, 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. This idea of design bias also occurs in the hierarchical Bayesian models presented in this thesis, however they generally reduce variance further as they are model-based estimators. Thus, we will not explore indirect estimators with implicit models further throughout 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 is primarily investigates to 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.

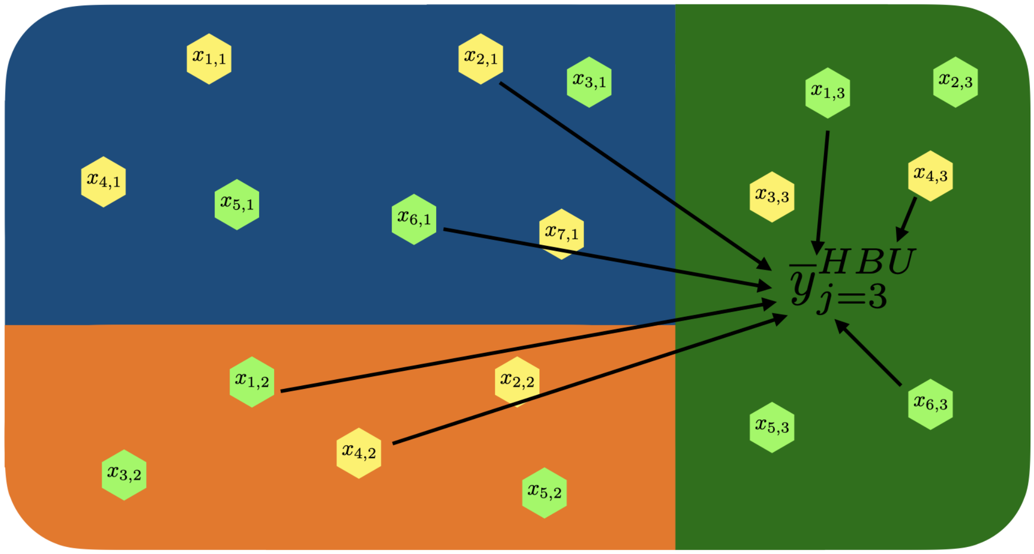

Figure 1.4: The unit-level hierarchical Bayesian model

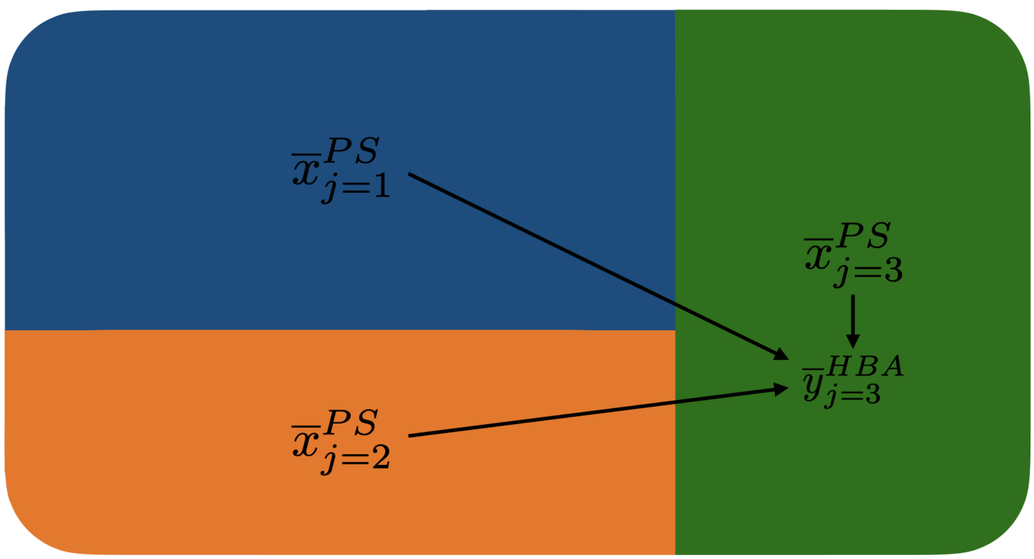

Figure 1.5: The area-level hierarchical Bayesian model

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. Notably, Figures 1.4 and 1.5 would be the same for the frequentist EBLUP unit- and area- level models as these diagrams do not depict the details of how strength is borrowed across eco-subsections. These EBLUP models borrow strength in a way that minimizes bias given modeling assumptions are correct whereas the hierarchical Bayesian models sacrifice adding some bias while further reducing variance.

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. This means that, given the modeling assumptions are met, our estimate ($\bar{y}_j$) of our parameter (μ_{y_j}) for each sub-population will have the following property:

$$E[\bar{y}_j] - \mu_{y_j} = 0. \tag{1.1}$$

That is, the expected value of the statistic, $\bar{y}_j$, 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:

$$\mathrm{MSE}(\bar{y}_j) = \mathrm{Var}(\bar{y}_j) + \mathrm{Bias}(\bar{y}_j,\ \mu_{y_j})^2 \tag{1.2}$$

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 Forest Inventory & Analysis Program (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 FIA because if our models can predict ground-collected variables well, 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 refered 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:

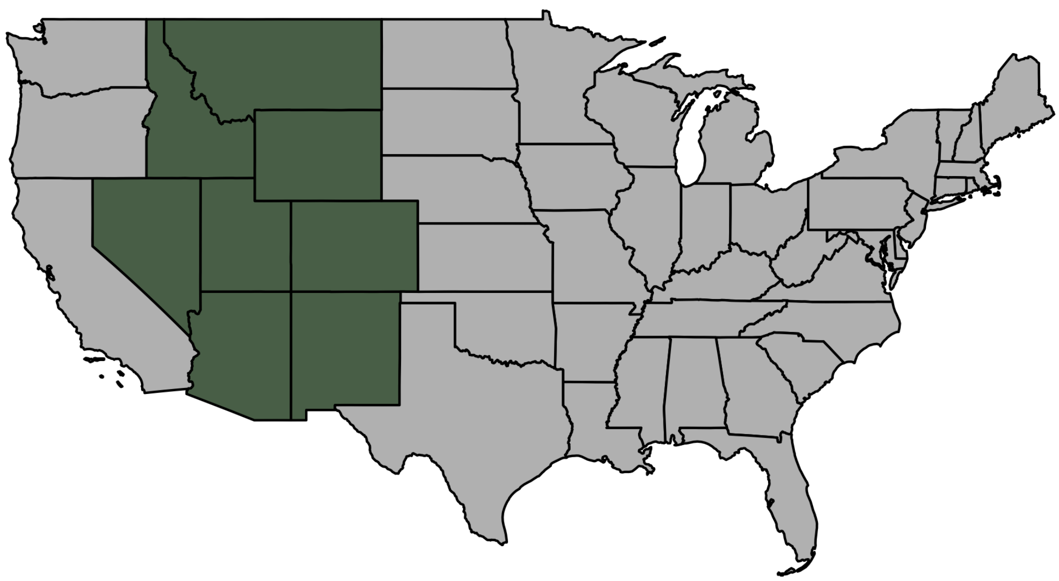

Figure 3.1: The Interior West Region of the United States

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 Forest Inventory and Analysis Program (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 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 ecosubsection. 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 ecosubsection. 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 & lattitude, elevation, predictor variables, response variables, ecosubsection, ecosection, and 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 ecosections, and these ecosections contain ecosubsections. In our data, on average, each ecosection contains approximately 7.06 ecosubsections, and each province contains an average of 4.86 ecosections. So, an average province then contains just over 34 ecosubsections.

We can take a look at the Northern Rocky Forest province, colored by ecosection, with lines dividing each ecosubsection to see this structure in action:

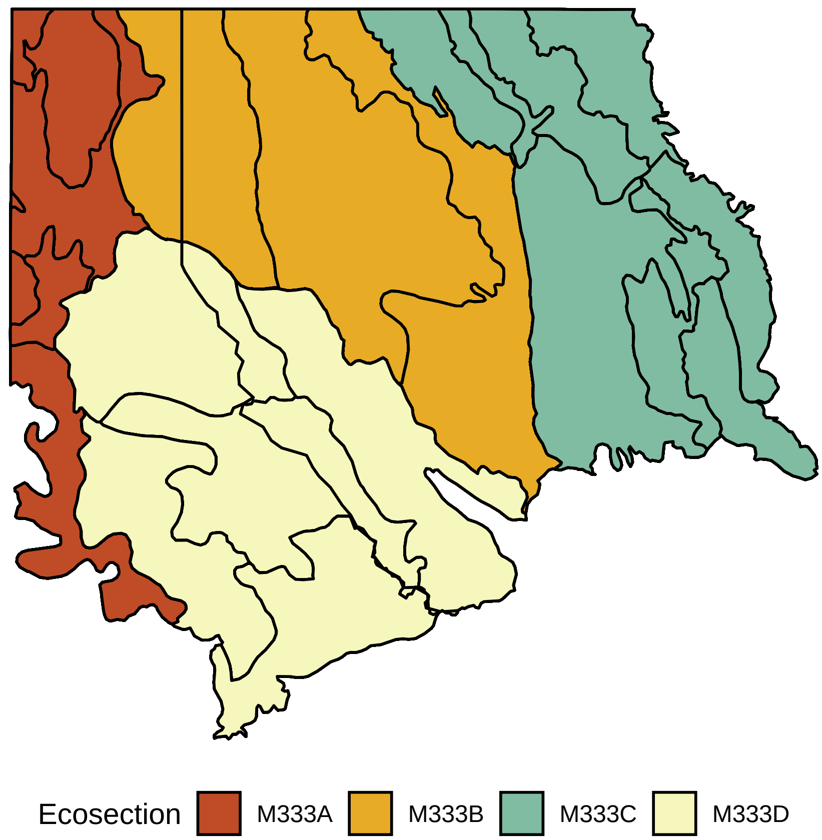

Figure 3.2: The Northern Rocky Forest, Colored By Ecosection

The data we have covers a total of 14 provinces, 68 ecosections, and 480 ecosubsections. The hierarchical struture of the data and nestedness of the ecosubsections within ecosections within 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 ecosubsection in the four following maps of the interior west.

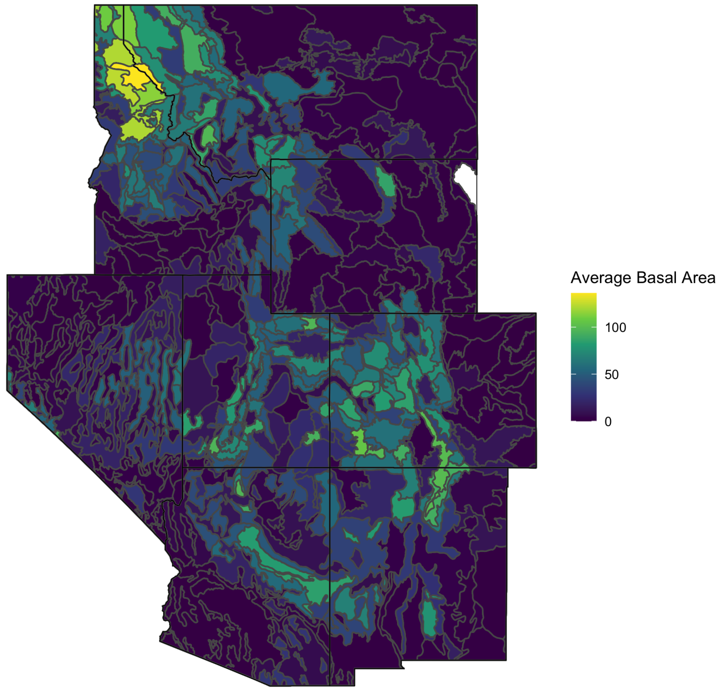

Figure 3.3: Mean Basal Area in Interior West Ecosubsections

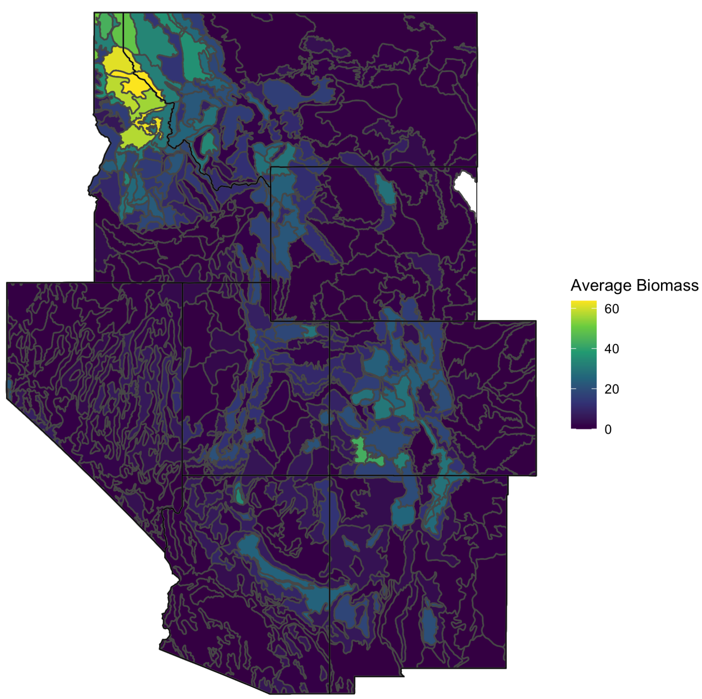

Figure 3.4: Mean Biomass in Interior West Ecosubsections

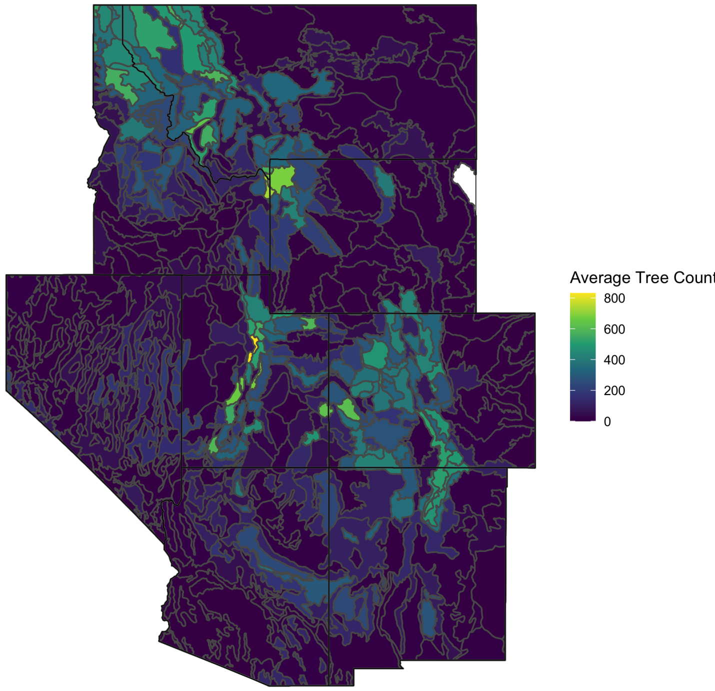

Figure 3.5: Mean Tree Count per acre in Interior West Ecosubsections

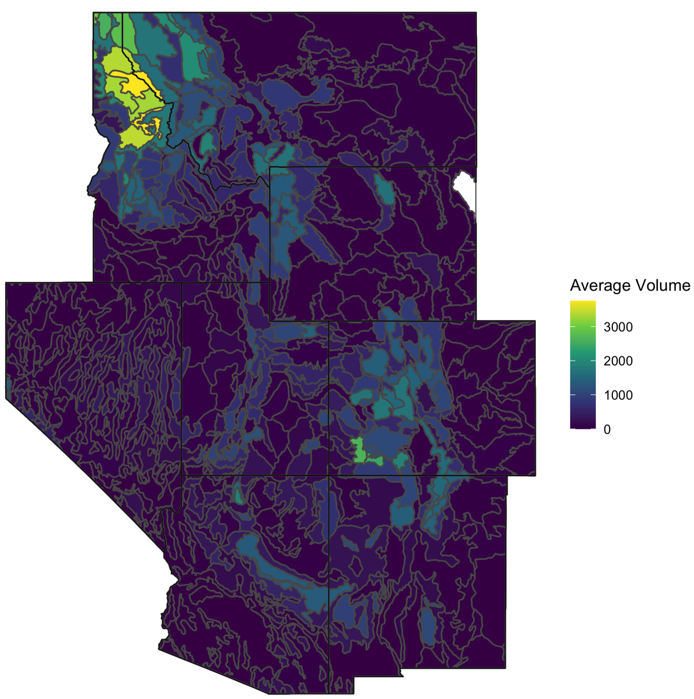

Figure 3.6: Mean Net Volume in Interior West Ecosubsections

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:

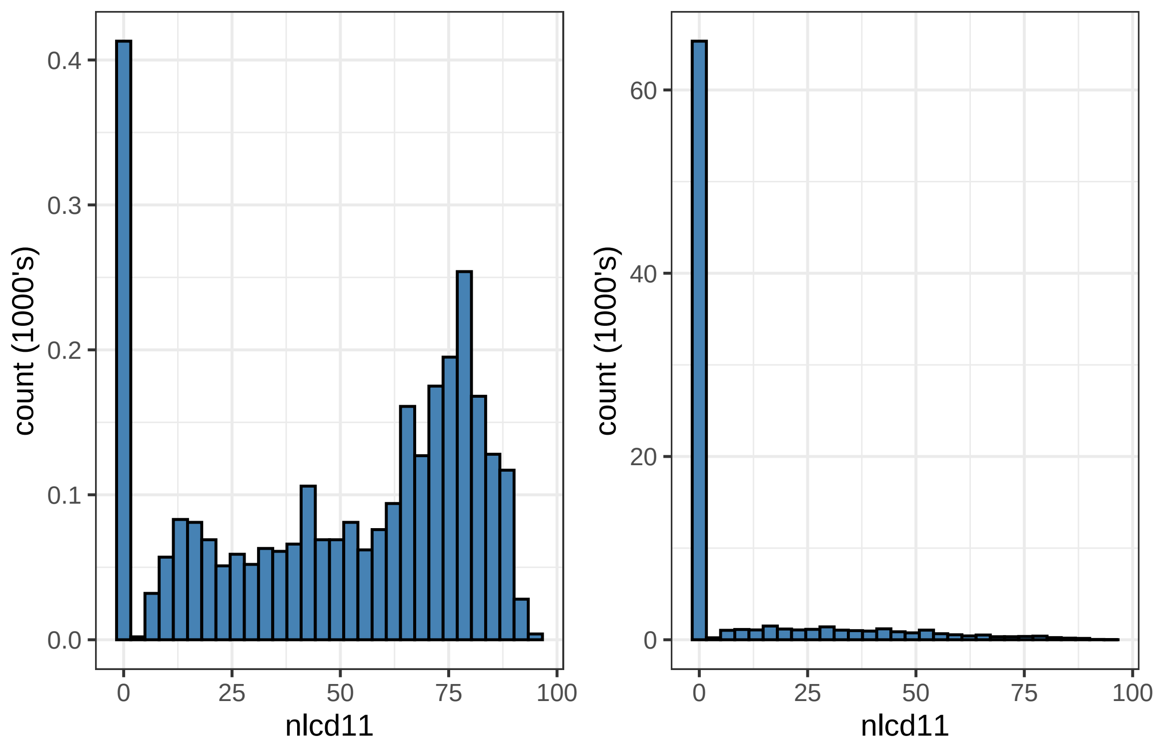

Figure 3.7: Distribution of Total Canopy Cover in the M333 Province (Top) and the Entire Interior West (Bottom)

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 ecosubsections are nested within ecosections which are in turn nested within 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 an example up close of Idaho's ecosubsections colored by their province, and then their ecosection:

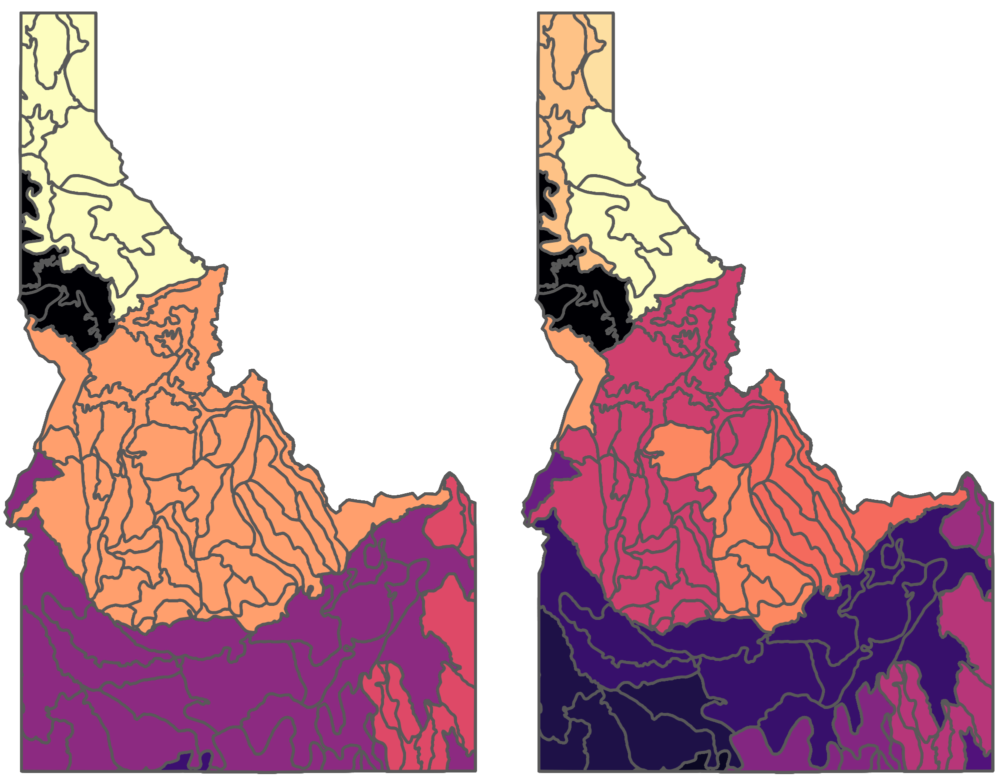

Figure 3.8: Idaho Colored by Province (Left) and Ecosection (Right)

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 succintness, we can look at the ANOVA results for one of the repsonse 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 ecosubsections:

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 province are more homogeneous than observations throughout the Interior West. Thus, if we want ecosubsection level estimates of variables, it makes sense to borrow information from other ecosubsections 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

4.1 Current Approaches

Currently, there are three main types of estimators used to estimate the value of forest attributes: direct, indirect, and model-based estimators. Direct estimators are commonly thought of as the simplest estimators as they do not borrow strength nor do they use auxilary variables for estimation. Direct estimators are hence easy to use and interpret, but we often do not get precise enough estimates with these estimators. Indirect and model-based estimators, such as the EBLUP estimastor, aim to reduce variance in estimates by using data from surrounding areas and (in the case of the model-based estimator) using auxilary data.

4.1.1 Direct and Indirect Estimation

Direct and indirect estimation when auxilary data is not available or information from auxilary datasets is not wanted. One of the most common direct estimators is the mean:

$$\overline{Y_j} = \frac{\sum_{i=1}^{n_j} Y_{i,j}}{n_j} \tag{4.1}$$

where $\overline{Y_j}$ is the mean of the variable of interest, i indexes over each observation in the small area j, $Y_{i,j}$ is the ith value of θ in the jth small area, and n_j is the total number of observations in that small area.

Another commonly used estimator which belongs to the family of direct estimators is the post-stratified estimator:

(PS EQN)

The post-stratified estimator is similar to the mean, however it includes a term which corrects for over or under sampling of forests in a given small area. If we have over or under sampled forests within our small area, the mean will actually be a biased estimator, while the post-stratified estimator is provenly unbiased. While it is not always practical to use the post-stratified estimator due to its need for population information on the forested vs. not-forested variable. However, we do have this population data for the entire interior west region of the United States, so we will be comparing our model based estimators to the baseline post-stratified direct estimator.

(indirect estimator examples, ask kelly abt this)

4.1.2 Model-Based Estimation

Model-based estimators are extremely useful when auxilary data is available and correlated with the response variables of interest. Most commonly, the estimator used is the EBLUP estimator, which, similarly to the post-stratified estimator, provenly unbiased. This is a random effects model which we can see specified below:

$$Y_{i,j} = \vec{X}_{i,j}^T \vec{\beta} + \nu_j + \epsilon_{i,j} \tag{4.2}$$

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, ν_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:

$$\nu \sim \mathrm{N}(0, \sigma_\nu^2), \tag{4.3}$$

$$\epsilon \sim \mathrm{N}(0, \sigma_\epsilon^2) \tag{4.4}$$

4.2 A Hierarchical Bayesian Approach

While we have explored the frequentist version of a hierarchical model, this thesis primarily studies *Bayesian* hierarhical models and compares their performance to their frequentist counterparts. With a Bayesian hierarhical 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:

$$Y_{i,j} \sim \mathrm{N}(\nu_j + \vec{\beta}\vec{X}_i,\ \sigma^2) \tag{4.5}$$

$$\nu_j \sim \mathrm{N}(\mu_\nu,\ \sigma_\nu^2) \tag{4.6}$$

$$\mu_\nu \sim \mathrm{N}(a,\ b) \tag{4.7}$$

$$\sigma \sim \text{Inv-Gamma}(c,\ d) \tag{4.8}$$

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 μ_ν 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:

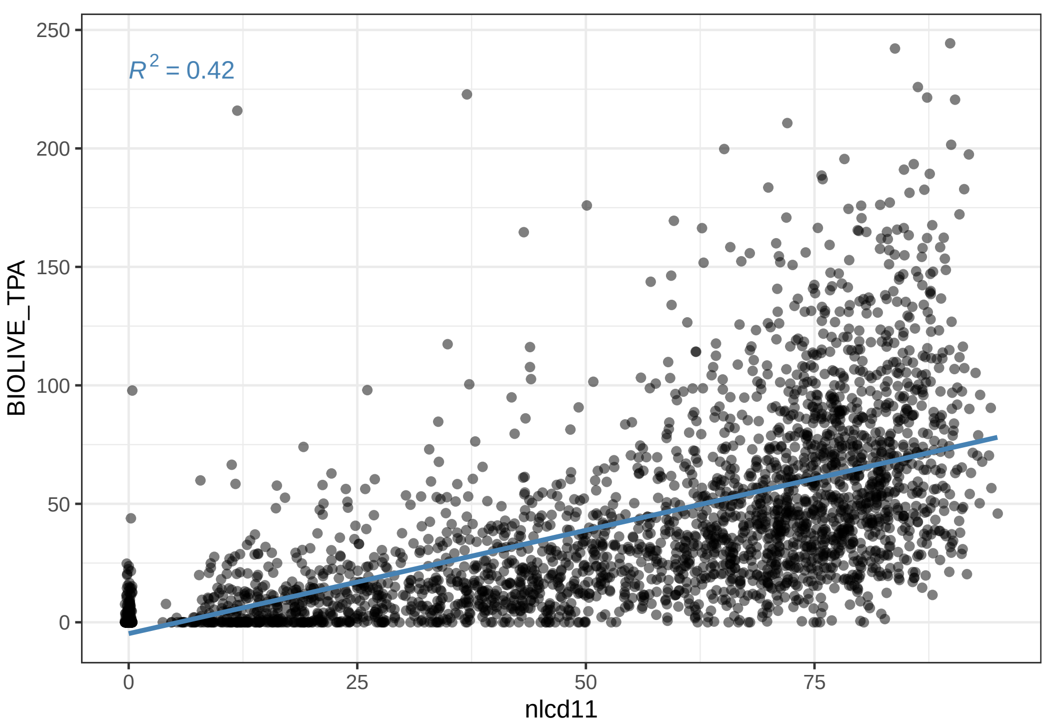

Figure 5.1: Unit-level correlation

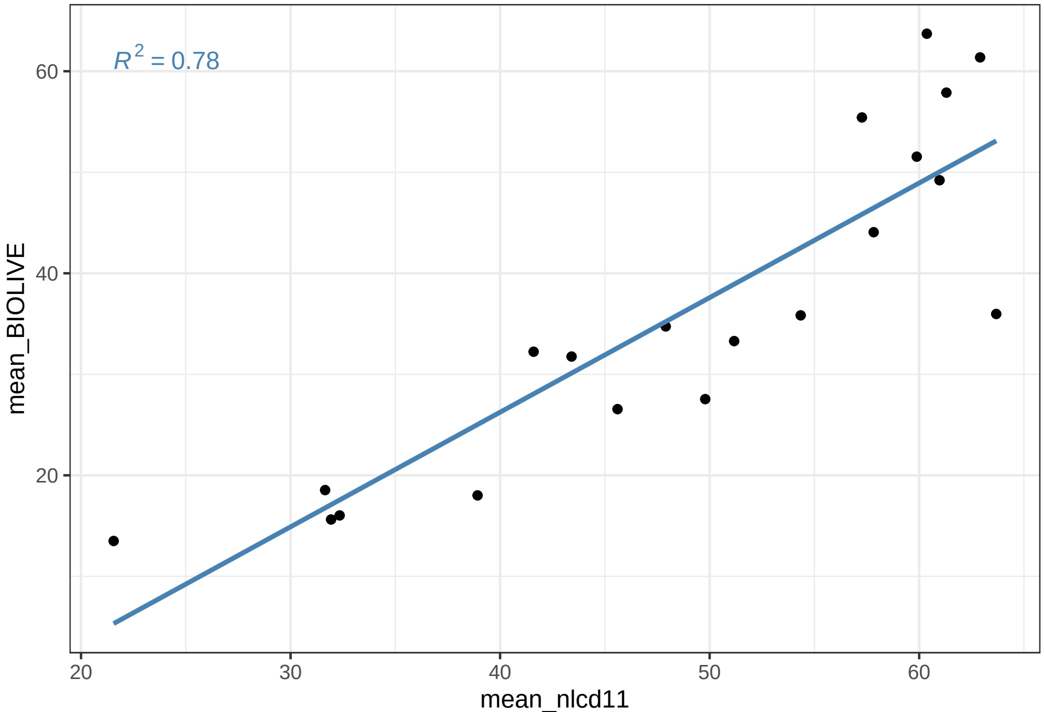

Figure 5.2: Area-level correlation

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 α_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 preforming the aggregation, we can look at the estimates of the mean biomass predicted by the model compared to the direct estimator:

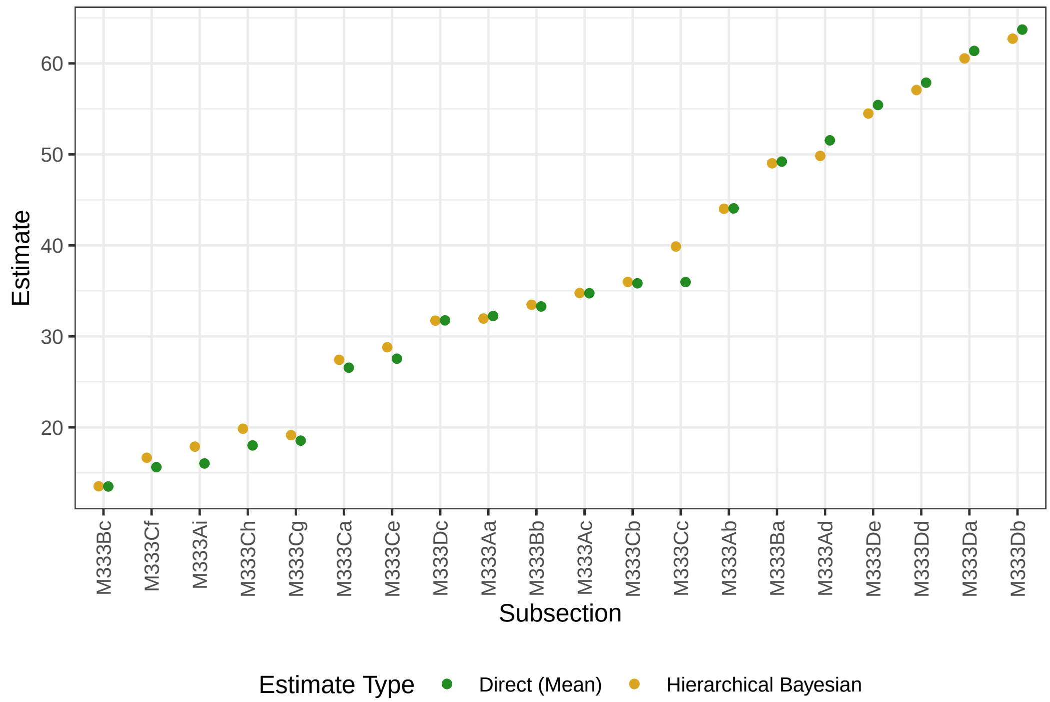

Figure 5.3: Direct and model-based estimates for the unit-level model

These estimates make sense in the context of hierarhical 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 hierarchial 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

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

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

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

Note that the numerator is now the root mean squared error of the jth ecosubsection. This is equivalent to taking the square root of the variance as we did in the direct estimator's coefficent 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 statisitc for each ecosubsection:

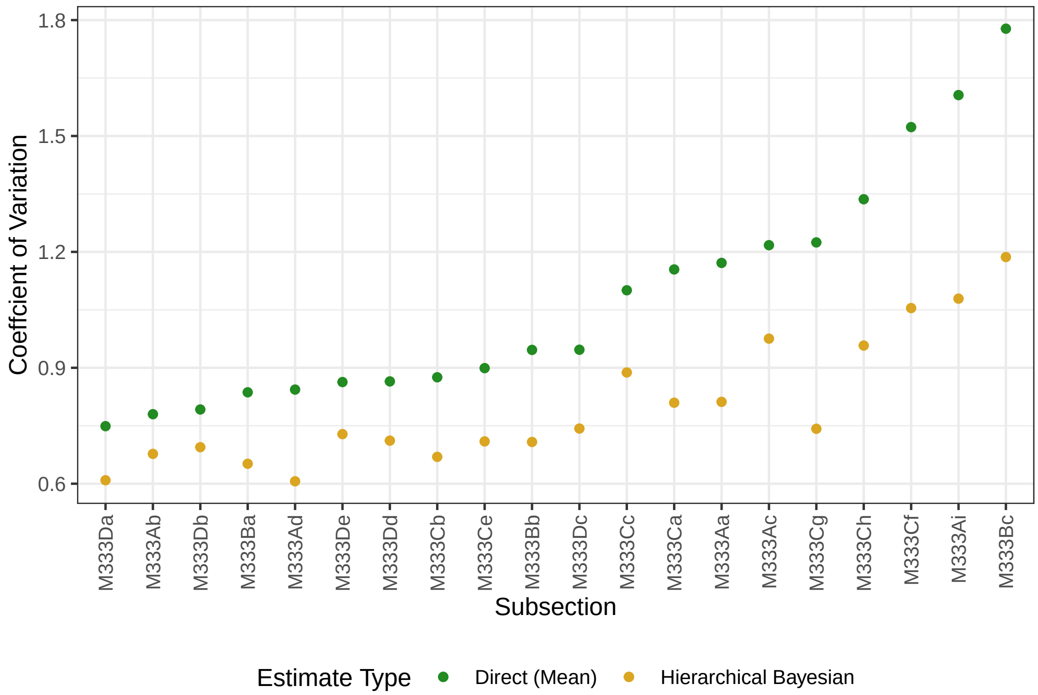

Figure 5.4: Direct and model-based coefficients of variation for the unit-level model

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 coefficent 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 readibility and/or setup.

In the main Rmd file

```
# This chunk ensures that the thesisdown package is
# installed and loaded. This thesisdown package includes
# the template files for the thesis.
if (!require(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)
```

```r
# 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/

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

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.

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.