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

I want to thank a few people.

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

Introduction **1**

Chapter 1: Context **3**

Chapter 2: Overview **5**

Chapter 3: Data **7**
3.1 The Forest Inventory & Analysis Program 7
3.2 The Interior West 7
3.2.1 The Northern Rocky Mountain Forest 8
3.3 Our Data: Specifics 8
3.4 Data Structure & Hierarchy 14

Chapter 4: Methods **15**

Chapter 5: Results **17**
5.1 Modeling Overview 17
5.2 Unit-level Models 18

Chapter 6: Discussion and Conclusion **21**

Appendix A: The First Appendix **23**

Appendix B: The Second Appendix, for Fun **25**

References **27**

List of Tables

List of Figures

3.1 The Interior West Region of the United States 8
3.2 Mean Basal Area in Interior West Ecosubsections 10
3.3 Mean Biomass in Interior West Ecosubsections 11
3.4 Mean Tree Count per acre in Interior West Ecosubsections 11
3.5 Mean Net Volume in Interior West Ecosubsections 12

5.1 Unit-level correlation 17
5.2 Area-level correlation 18
5.3 Direct and model-based estimates for the unit-level model 19
5.4 Direct and model-based coefficients of variation for the unit-level model 20

Abstract

The preface pretty much says it all.

Second paragraph of abstract starts here.

Dedication

You can have a dedication here if you wish.

Introduction

This is the introduction to my thesis.

Chapter 1

Context

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 m by 30 m 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 m by 30 m hexagonal units by using a geographically-based systematic sampling design (Source: McConville et al, 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. This plot level data is what we are working with throughout the duration of the thesis.

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:

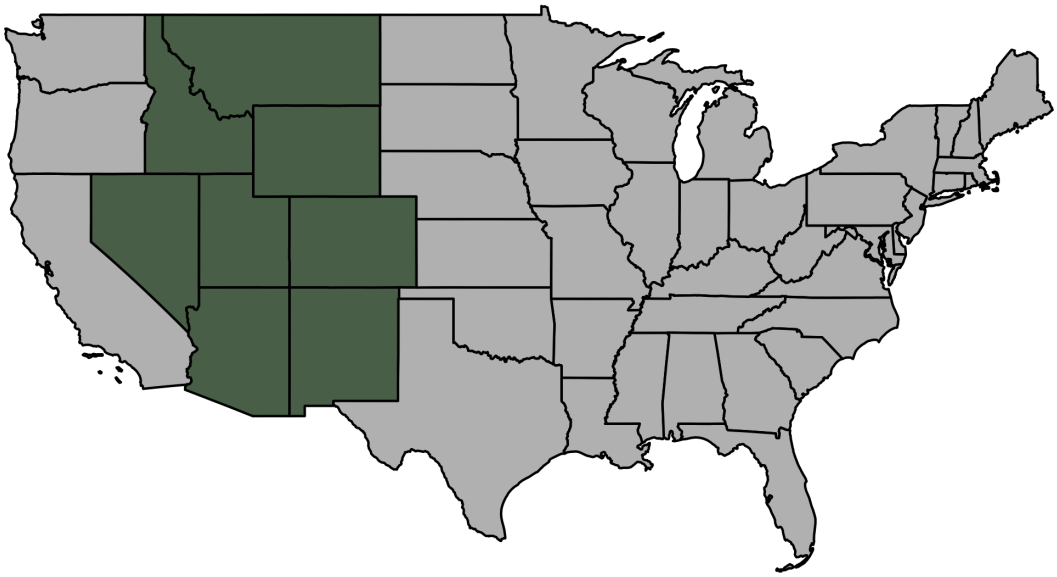


Figure 3.1: The Interior West Region of the United States

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.

3.2.1 The Northern Rocky Mountain Forest

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 three remotely sensed pre-

dictor variables (`forprob`, `forbio`, and `nlcd11`), location information, and ecosubsection. The `forprob` and `forbio` predictors were colleted by FIA (Blackard et al., 2008). The `nlcd11` variable was collected by the Environmental Protection Agency ("Completion of the 2011 national land cover database for the conterminous united states – representing a decade of land cover change information," 2015).

The `response` dataframe contains 86085 observations and 67 variables, most notably four predictor variables collected by FIA crew members (`BALIVE_TPA`, `CNTLIVE_TPA`, `BIOLIVE_TPA`, and `VOLNLIVE_TPA`), location information, and ecosubsection. 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:

PLOT	LON	LAT	ELEV	forgrp	forprob	demLF	forbio	BALIVE_TPA
83657	-111.3261	35.02106	6680	180	1	2080	12.67684	236.1169
87963	-109.9398	36.59399	5550	0	0	1700	0.00000	0.0000
84186	-109.9925	36.27860	7510	180	1	2305	10.57715	105.3212
87499	-109.9058	35.32838	5630	0	0	1717	0.00000	0.0000
88091	-109.9024	34.83752	5510	0	0	1672	0.00000	0.0000
80842	-109.9774	33.52990	5920	180	1	1828	14.46534	149.7189

CNTLIVE_TPA	BIOLIVE_TPA	VOLNLIVE_TPA	subsection	section	province
555.3678	43.08024	2066.3352	M313Ak	M313A	M313
0.0000	0.00000	0.0000	313Ac	313A	313
108.3248	18.02096	988.9084	313Ap	313A	313
0.0000	0.00000	0.0000	313Db	313D	313
0.0000	0.00000	0.0000	313Db	313D	313
273.5608	21.76656	1381.8511	313Cd	313C	313

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. 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 two explanatory variables. The response

variables used are basal area (square-foot), trees per acre, above-ground biomass (lbs), and net volume (ftˆ3). 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.

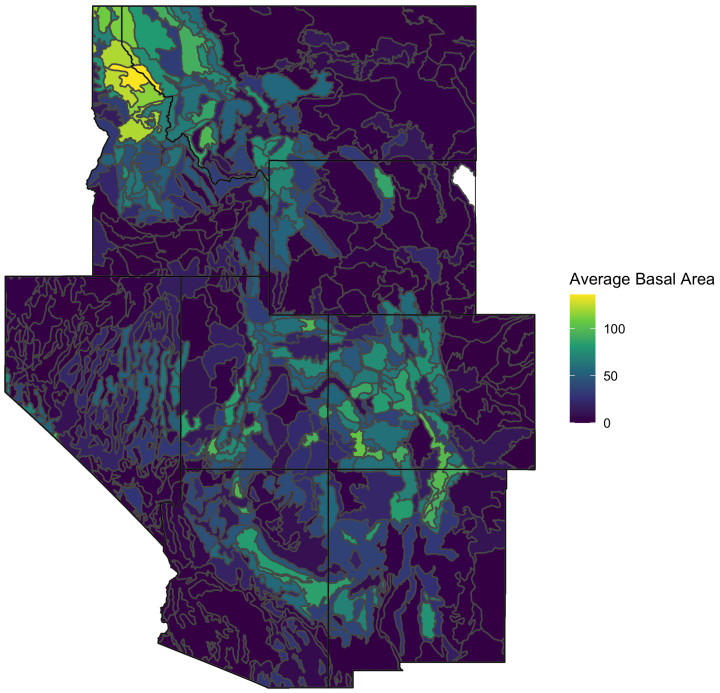


Figure 3.2: Mean Basal Area in Interior West Ecosubsections

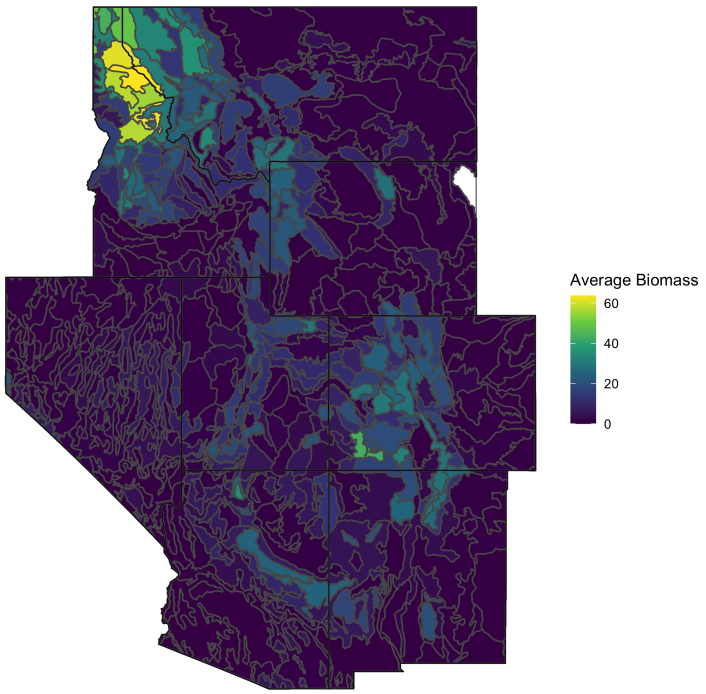


Figure 3.3: Mean Biomass in Interior West Ecosubsections

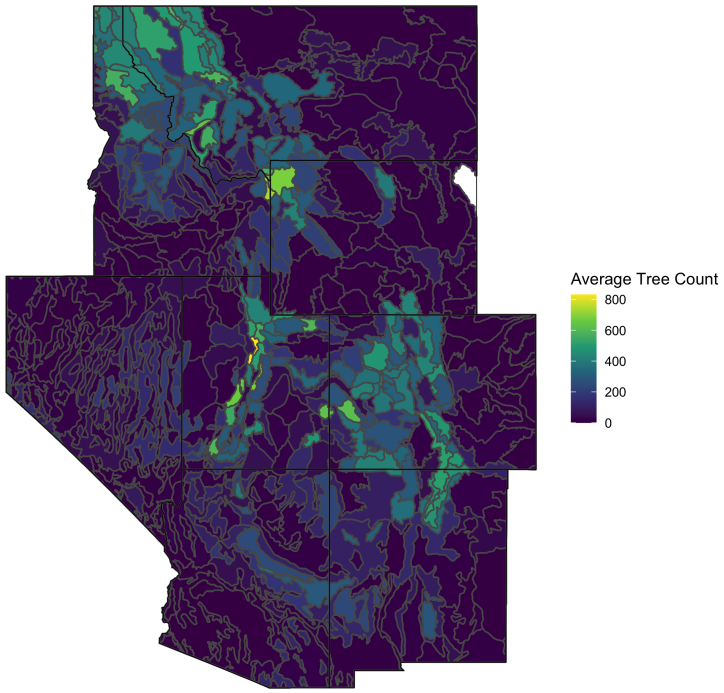


Figure 3.4: Mean Tree Count per acre in Interior West Ecosubsections

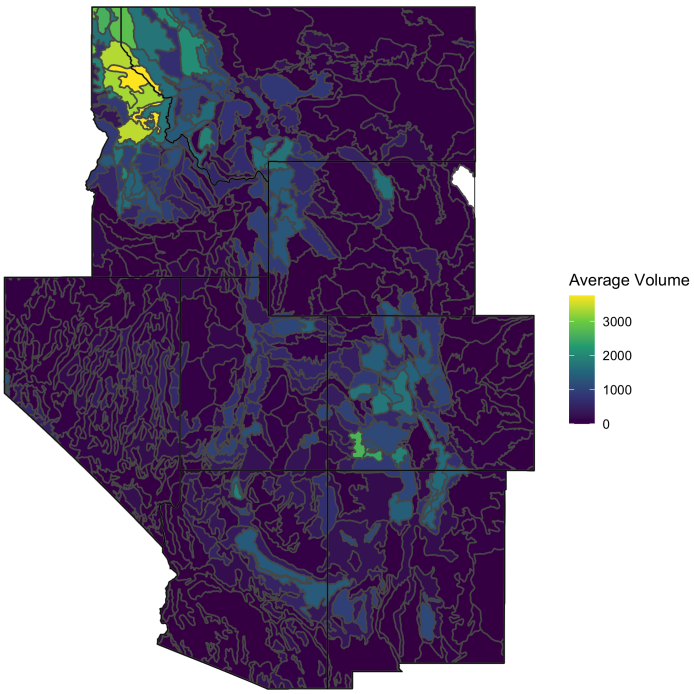
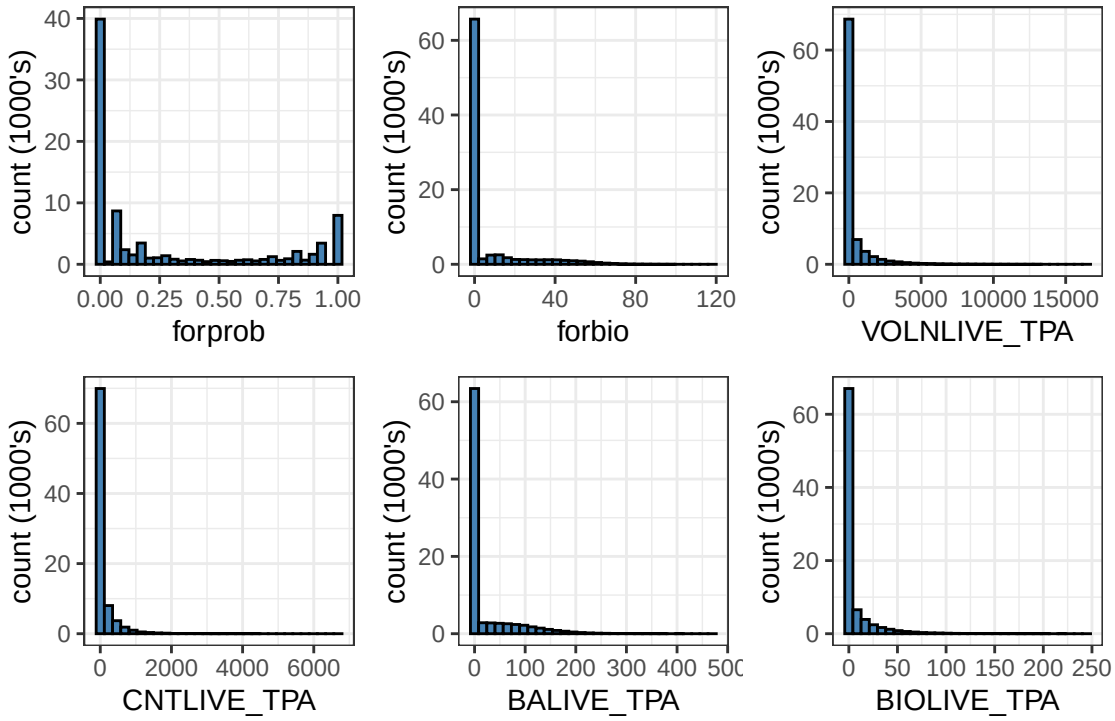


Figure 3.5: Mean Net Volume in Interior West Ecosubsections

While we have four variables which we will model as response variables throughout the analyses, we also have two predictor variables which will be of much use to us. In particular, forest probability and forest biomass (coded as `forprob` and `forbio`.) These variables which we will treat as predictors are remotely sensed variables, 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. While it may seem unnatural to attempt to predict forest biomass (`BIOLIVE_TPA`) with forest biomass (`forbio`), the differences in the data collection process between the ground level data and remotely sensed data are quite different.

These variables are almost all right-skewed, and all take value zero quite often. To get a good sense of the distributions of the six key variables, we can look at histograms of each variable:



It is notable that the `forprob` variable is bimodal and modes zero and one, while all other variables are extremely right-skewed. This is likely because when `forprob` is decided from the remotely sensed data, there are likely areas that are either very clearly forest or very clearly not forest (residential areas, for example).

Apart from making histograms of our data, we can also summarize the data to see some summary statistics of our six key variables:

variable	mean	quantile_25	median	quantile_75	min	max	na_count
forbio	6.66	0	0.00	0.00	0	118.00	0
forprob	0.27	0	0.07	0.56	0	1.00	1
BIOLIVE_TPA	6.23	0	0.00	1.98	0	244.35	0
BALIVE_TPA	22.75	0	0.00	14.75	0	469.39	0
CNTLIVE_TPA	98.60	0	0.00	30.09	0	6677.93	0
VOLNLIVE_TPA	342.32	0	0.00	74.69	0	16435.55	0

From this table, we can see how heavily skewed these key variables are, with over 25% of the observations for each variable being 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.

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`. By just looking at the MSE of the ANOVA results, we can see that we should expect more homogeneity within ecosubsections:

variable	province_MSE	ecosection_MSE	ecosubsection_MSE
forbio	614761.0	21713.4	6163.8
forprob	362.2	15.1	4.2
BIOLIVE_TPA	500958.2	17910.5	5456.2
BALIVE_TPA	3568986.3	118683.2	47835.7
CNTLIVE_TPA	85365657.9	2512475.9	1287922.7
VOLNLIVE_TPA	1559809302.3	60049667.7	18445636.8

(* Need a better way to output this table and show province level model is significant)

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

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:

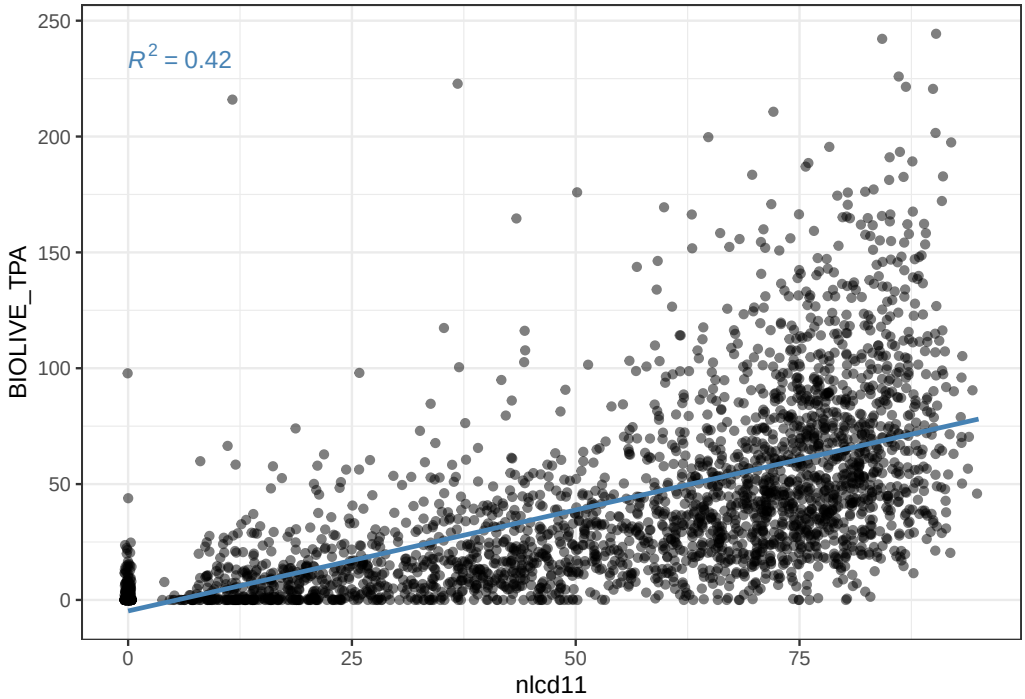


Figure 5.1: Unit-level correlation

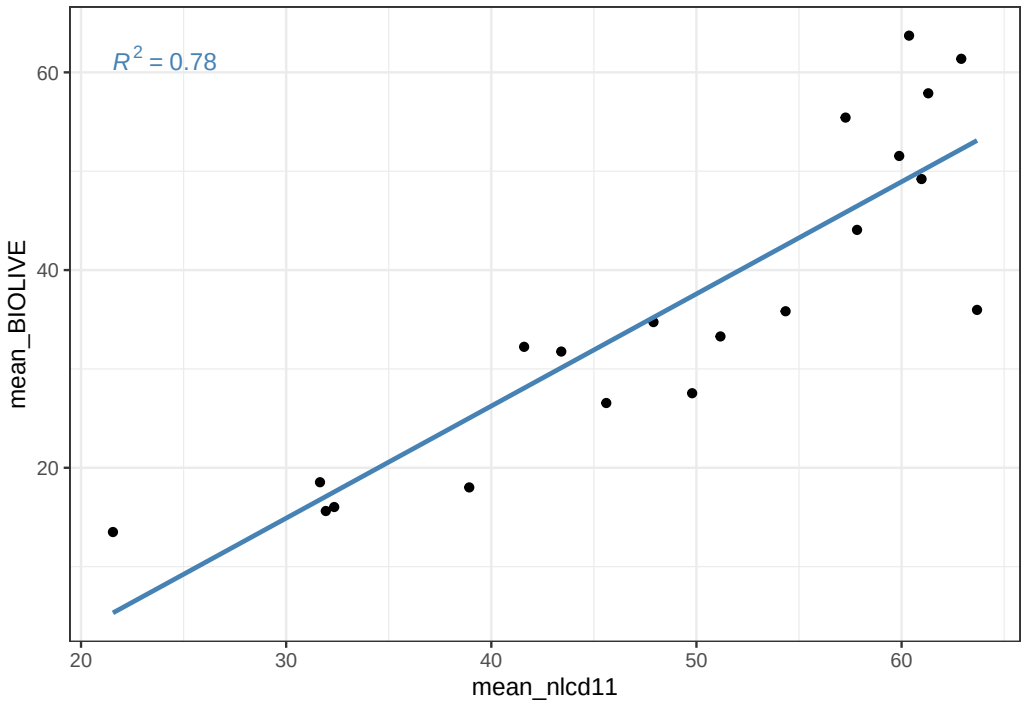


Figure 5.2: Area-level correlation

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:

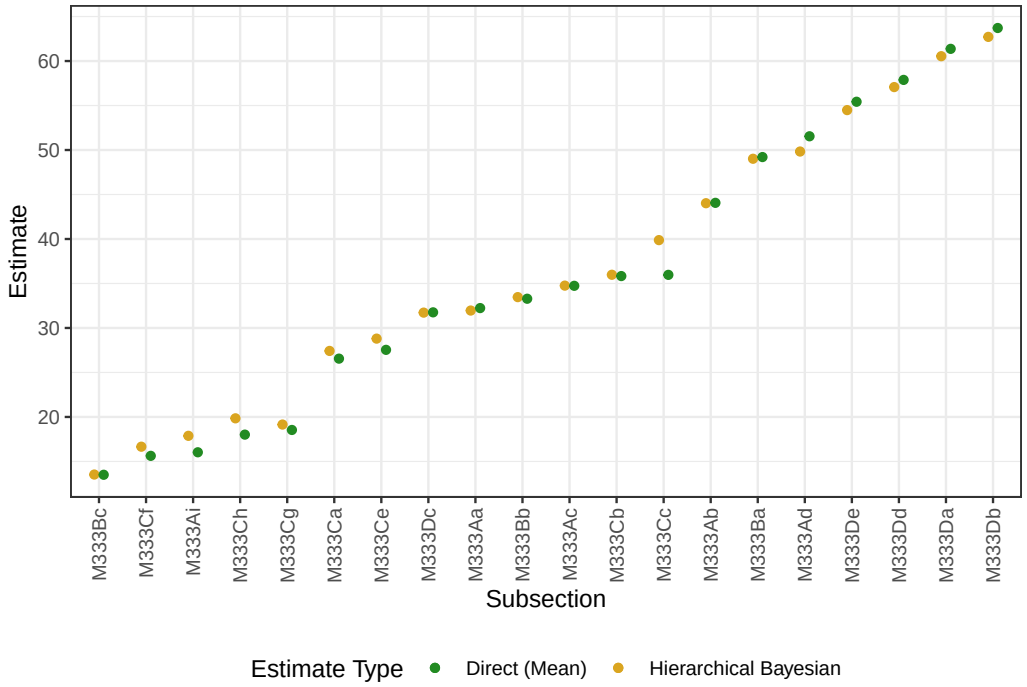


Figure 5.3: Direct and model-based estimates for the unit-level model

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 equvalent 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:

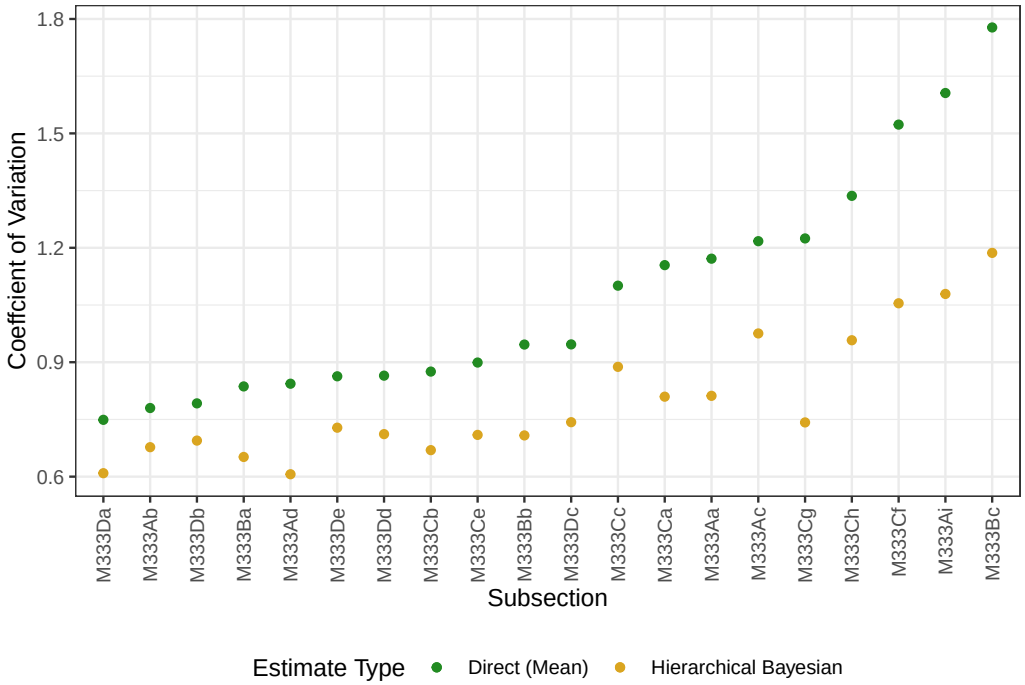


Figure 5.4: Direct and model-based coefficients of variation for the unit-level model

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)
```

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

In Chapter ??:

Appendix B

The Second Appendix, for Fun

References

Blackard, J., Finco, M., Helmer, E., Holden, G., Hoppus, M., Jacobs, D., ... others. (2008). Mapping us forest biomass using nationwide forest inventory data and moderate resolution information. *Remote Sensing of Environment*, *112*(4), 1658–1677.

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

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