A Comparison of Forest Estimations: Post-Stratification & Generalized Regression Estimators

Abstract

The National Forest Inventory and Analysis (FIA) Program of the United States Forest Service collects and analyzes data on many important forest attributes. The current FIA procedure is to utilize post-stratification (PS) estimation to improve estimation precision. The Interior West (a region of the United States) stratifies estimates by forest and non-forest areas. This research compares the use of PS estimation with the generalized regression estimator (GREG) estimation technique in the Interior West. In estimating four forest attributes, we consider and compare 5 GREG estimators to the PS, specifically their relative efficiences. We find the GREG estimator generally improves the precision of forest attribute estimates, but not for counties with a small number of ground plots, nor for counties with a small amount of our variable of interest, e.g. counties that are nearly all non-forest.

Keywords: Forestry, Generalized Regression Estimator, Post-Stratification Estimator, Forest Attributes, Geospatial Data, Interior West

1 Introduction

The National Forest Inventory and Analysis (FIA) Program of the USDA Forest Service has been in continuous operation since 1930, shortly following the passage of the McSweeney-McNary Forest Research Act of 1928. Once established, the FIA became the primary source of conducting and continuously updating a comprehensive inventory and analysis of the present and prospective conditions of the renewable resources of the United States ("Forest Inventory and Analysis National Program - About Us" n.d.).

As the FIA creates and maintains a broad inventory of resources, data is available on numerous forest attributes ranging from merchantable timber and other wood products, risks associated with fire, fuels and potential fire hazards, conditions of wildlife habitats, insects or diseases, biomass, carbon storage, forest health, and other general characteristics of forest ecosystems.

In creating and maintaining a broad-scale resource inventory, the FIA is responsible for monitoring forest ecosystem attributes across the United States. Given the diversity of forests and rangelands composing the United States, groups of states are divided into unique geographic divisions, as noted by the United States Census Bureau. One of these divisions, and the division of interest for our research, is the Interior West. The Interior Western covers the states of Arizona (AZ), Colorado (CO), Idaho (ID), Montana (MT), Nevada (NV), New Mexico (NW), Utah (UT), and Wyoming (WY) ("Interior West Forest Inventory & Analysis - About Us" n.d.).

The FIA program, both in the Interior West and broadly, is important as it is the sole source of consistent annual forest survey data across the entire country. The FIA provides objective and scientifically credible information on how much forest there is, what it looks like, whether the forest area is increasing or decreasing, whether we are gaining or losing species, how quickly trees are growing, dying, and being harvested, and how forest ecosystems change over time. In sum, the FIA collects information to tell us about the current and prospective state of our environment. Thus, it is important that the estimates of different forest attributes are accurate and comprehensive.

FIA data primarily comes from two sources: plot-level and pixel-level. Plot-level data, also known as field data, comes from field plots distributed across each state, at a sample intensity of about one plot every 6,000 acres (Bechtold and Patterson 2015). Most field data is related to tree and understory vegetation components of a forest. By comparison, pixel-level data comes from remote-sensing sources, such as satellite imagery, whereby data is gathered on aspects of the environment such as latitude, longitude, and elevation. Compared to plot-level data, pixel-level data has greater sampling intensity; much more of the Interior West is being sampled in the pixel data.

We combine the two sources of FIA data—pixel and plot-level data—to estimate forest attributes, utilizing two survey estimation methods. However, we are not the first to utilize

both plot-level and pixel-level data in an attempt to improve estimation. Survey estimation is a well-developed field, and has led to the creation of numerous survey estimation packages to streamline estimation techniques—of note is the *'mase'* package for R (McConville et al. 2018). Many survey estimation statistical packages implement advanced survey estimation techniques such LASSO, elastic net, and ridge regression.

Though progress has been made in the literature on the application and utilization of advanced estimation techniques such as penalized regression, FIA still relies primarily on post-stratification for producing its estimates. As FIA still relies on post-stratification, we wanted to explore the use of another, well-documented technique—generalized regression estimation (GREG)—and compare it to post-stratification estimates. Post-stratification is a special case of the GREG that incorporates one categorical predictor.

Overall, the objective of our research is to compare the precision of the GREG estimator to the post-stratification estimator in estimating forest attributes in the Interior Western US. The forest attributes we are interested in estimating are forest biomass, basal area, volume of live trees, and count of live trees. Within the Interior West, we analyze whether the GREG estimator is more efficient than the post-stratification estimator. We are interested in quantifying how much smaller (if at all) the standard error of the GREG estimator is under different predictor schemas compared to the current post-stratified estimator.

2 Methods

2.1 Objective

Our goal is to assess the GREG estimator (McConville, Moisen, and Frescino 2020) and its variance when compared to the PS estimator, which is what is currently being utilized by the IW Forest Inventory Analysis (FIA). For both the GREG and PS estimators, we denote the estimation of the variable of interest as $\hat{\mu}_y$, and denote whether we're referring to PS or GREG by subscripting the $\hat{\mu}_y$ being estimated.

2.2 The GREG estimator

The GREG estimator requires a linear model utilizing auxiliary data, and combines both the predicted values for our pixel data with the residuals of the known plot values to find a predicted mean value for forest attribute, $\hat{\mu}_{y,GREG}$. A simple version of a GREG estimator, done at a county level, is:

$$\hat{\mu}_{y,GREG} = \bar{y} - \bar{x}_n^T \hat{\beta} + \bar{x}_N^T \hat{\beta}$$

Where $\bar{y}$ is the mean variable of interest in a county, $\bar{x}_n$ is the mean auxiliary data for the plots where the variable of interest was measured, and $\bar{x}_N$ is the auxiliary pixel data available for the whole county. For example, $\bar{y}$ could be the count of trees, and $\bar{x}$ could be canopy cover. The $\hat{\beta}$'s are obtained using a multiple linear regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i = x_i^T \beta + \epsilon_i$$

Running the regressions at the county level is just one possible resolution to run these regressions, and an arbitrary one at that. To potentially strengthen our model, we could instead, for example, run the regression on the 14 ecological provinces found in the Interior West (detailed in Data Preparation). An alternative multiple linear regression model could be run to obtain $\hat{\beta}$'s at resolution unit j.

But this creates an issue: the estimates should be at the county level, not another. So we need to introduce w_j, which is the proportion of the county of interest which lies in region j, and recreate our GREG estimate by summing the different parts of counties which lie in different resolution units.

$$\hat{\mu}_{y,GREG} = \bar{y} - \sum_{j=1}^{q} w_j \bar{x}_n^T \hat{\beta}_j + \sum_{j=1}^{q} w_j \bar{x}_N^T \hat{\beta}_j$$

This is the final GREG model we use in this paper, with our chosen resolution being ecological province.

2.3 The Post-Stratification estimator

The post stratification estimator uses the product of one categorical auxiliary data to estimate means of the variables of interest, denoted $\hat{\mu}_{y,PS}$. For the Interior West (IW), pixels are classified as forest stratum or nonforest stratum. Generally, the post stratification procedure incorporates a categorical variable with D categories; the variable is expressed in the linear model, as in section 3.1, using indicator variables, where x_{ij} denotes category j for $j = 1, \ldots, D$. We can reduce this expression to the group mean model: $y_i = \sum_{j=i}^{D} B_j x_{ij} + \epsilon_i$.

In the estimated coefficient vector, the jth entry of the estimated coefficient vector reduces to the stratum mean estimator of y for category j.

$$\hat{B}_{sj} = \frac{1}{n_j} \sum_{i \in s_j} y_i = \tilde{u}$$

where s_j are the sample units in category j, and n_j is the sample size, i.e. number of plots, in category j. Under these specifications, the GREG may be denoted as a weighted average of the post-stratified means (McConville, Moisen, and Frescino 2020).

$$\hat{\mu}_{y,PS} = \frac{1}{N} \sum_{j=1}^{D} \frac{N_j}{n_j} \sum_{i \in s_j} y_i = \frac{1}{N} \sum_{j=1}^{D} N_j \tilde{\mu}_{y_j}$$

2.4 Assessing the Predictors

To find our best linear model for our GREG estimator, we made two decisions: what predictors to use, and what resolution to build the linear models at. Our resolutions of interest were: Total IW, State-level, County-level, Ecological Province, and Ecological Section. We construct the estimators at the county level, but we still examine whether it is optimal to use one model built on the entire dataset, use a different model for each county, or use different models on a resolution in between these two extremes.

To evaluate an appropriate resolution level, we created models composed of every combination of the selected predictors at every resolution, and examined the output of these models, specifically the adjusted R^2 and the standard deviation of the adjusted R^2s. We chose the resolution that had the least variance across the different units of the resolution. This would ensure that the model we end up with performed at a high level covering the entire Interior West. The adjusted R^2 statistic is calculated as:

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

where RSS represents the Residual Sum of Squares, TSS represents the total sum of squares, n the number of observations, and d the number of predictors.

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y})^2 TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

TSS measures the toal variance in the response Y, and can be thought of the total variance for Y before any regression is performed. RSS measures the amount of variability explained after performing our regression. Maximizing adjusted R^2 is equivalent to minimizing $\frac{RSS}{n-d-1}$ (James et al. 2013). Adding variables that do not lead to a significant decrease in RSS will then increase $\frac{RSS}{n-d-1}$ as a whole, because we subtract from our denominator the number of variables we include in our model.

Once we decided on the appropriate resolution, we then evaluated what predictors to include in our models. We evaluated combinations of predictors by comparing the adjusted R^2 and the Bayes Information Criterion (BIC) across ecological provinces. We use adjusted R^2 and BIC as these measures penalize models with unnecessary predictors, and we wanted to ensure the inclusion of additional predictors was appropriate. We then grouped our models by their size, where size denotes the number of predictors in a model. For a model with d predictors, n observations, and sample variance of $\hat{\sigma}^2$ the BIC is given by:

$$BIC = \frac{1}{n\hat{\sigma}^2}(RSS + \log(n)d\hat{\sigma}^2)$$

Our optimal models minimize BIC, which tends to place a heavier penalty on models with many variables (James et al. 2013).

The resolution analysis was conducted on the training dataset. After comparing the top mod-

els at each size, we then explored the inclusion of several combinations of the location-based predictors, specifically latitude, longitude, and elevation. In addition to two-predictor combinations, we also included a three-predictor combination—latitude, longitude, and elevation—in models that initially included all three predictors. We again compared the adjusted R^2 and BIC of these new models compared to our original top models with no interactions included. After narrowing down our list of top models, we compared the adjusted R^2s and the BICs from our training data set to those in our test and total data sets.

2.5 Comparing PS and GREG

After choosing our best models, we compared the values and variance of both the post-stratification and GREG estimators. We assessed the variance using a bootstrap variance estimator, which mimics the sampling variability by taking a bootstrap sample of the data. The steps of our bootstrap procedure are:

1. Take a simple random sample with replacement of size n from the original sample, called the bootstrap sample.
2. Compute the estimator, $\hat{u}_y$, on the bootstrap sample.
3. Repeat steps 1 and 2 many times. In this instance, we performed steps 1 and 2 1000 times.

The bootstrap variance estimator is given by:

$$\hat{SE}_B(\hat{u}_y) = \sqrt{\big(\frac{n}{n-1}\big)\big(\frac{N-n}{N-1}\big)\frac{1}{B-1}\sum_{b=1}^{B}(\hat{u}_y^{(b)} - \bar{\hat{u}})^2}$$

,

where $\hat{u}_y^{(b)}$ is the bth bootstrap estimate. The estimator has two bias adjustment terms: $n(n-1)^{-1}$ adjusts for bias induced by taking a bootstrap sample instead of a random sample from the population, and $(N-n)(N-1)^{-1}$ adjusts for not replacement sampling in the original collection of data (McConville, Moisen, and Frescino 2020).

Once we had our best models, we used the total data to build province-level models, construct GREG and PS estimates at the county level along with their standard errors (SEs). We then compare the ratio of the SE of PS to GREG for each county, called the relative efficiency.

$$RE = \frac{SE_{PS}}{SE_{GREG}}$$

If the RE of a county is equal to 1, then the bootstrap standard errors of the two estimators is equal, and the GREG shows no improvement in precision over the PS estimator. If the RE of a county is greater than 1, then we recognize that the GREG has more precise estimates than PS.

3 Data Description

3.1 Data Sources/Data Collection

Our dataset is comprised of two different kinds of data: plot and pixel. Plot level data was collected by the FIA in a quasi-systematic sample of ground plots over a 10 year period, with a base sampling intensity of one plot per every 6,000 acres (Bechtold and Patterson 2015). The plots are based on a systematically sampled hexagonal projection over the US, specifically Phase 2 plot data. These plots were measured in person for a selection of variables. We also have access to auxiliary data from satellite imagery based on 30x30 meter "pixels" which we use to appropriately weigh our estimators.

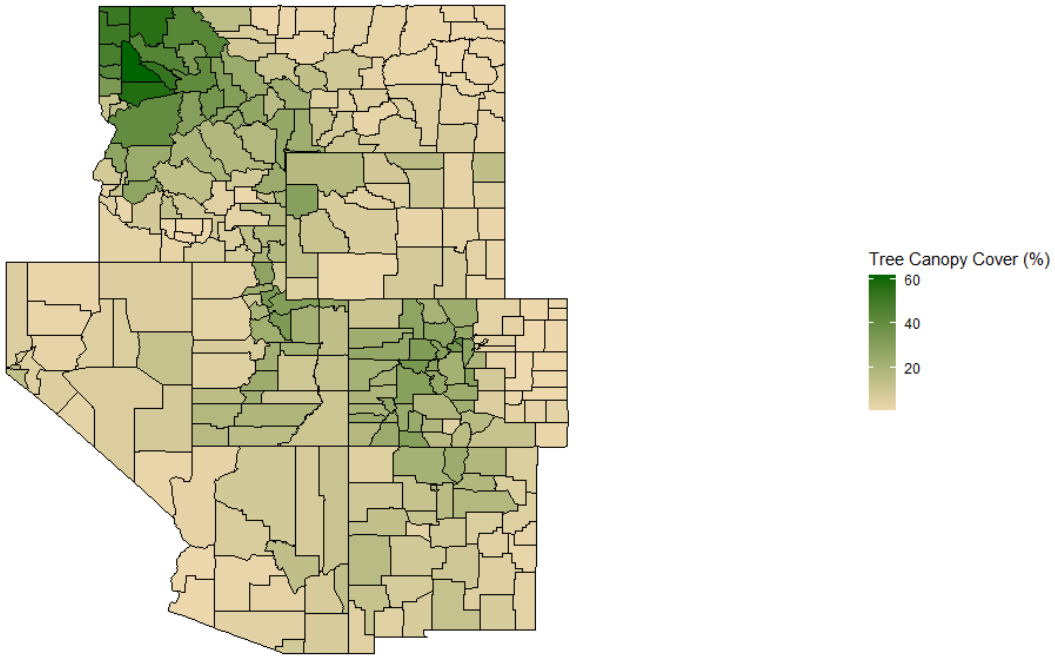

Figure 1: Tree Canopy Cover Using Pixel Data

3.2 Key Assumptions

Though our data used was collected from 2004 to 2013, we do not take into account differences by year. For our analysis, we treat all years as part of a single cycle. Likewise, though plot-level data was collected during different seasons, we do not take into account any seasonal differences.

Furthermore, for ease of analysis, we assume minimal temporal (spatial) autocorrelation. However, this assumption is explored in greater detail by other researchers. We elaborate on this further in our Discussion, but for the time being we recommend research such as

that of Magnussen and Fehrmann, who explore variance estimators under different levels and structure of autocorrelation (Magnussen and Fehrmann 2019).

We also expect human bias in data collection, particularly in tree canopy cover percentage, with clusters in increments of 5, as shown in Figure 2. However, we assume the impact this bias has on our estimates are negligible.

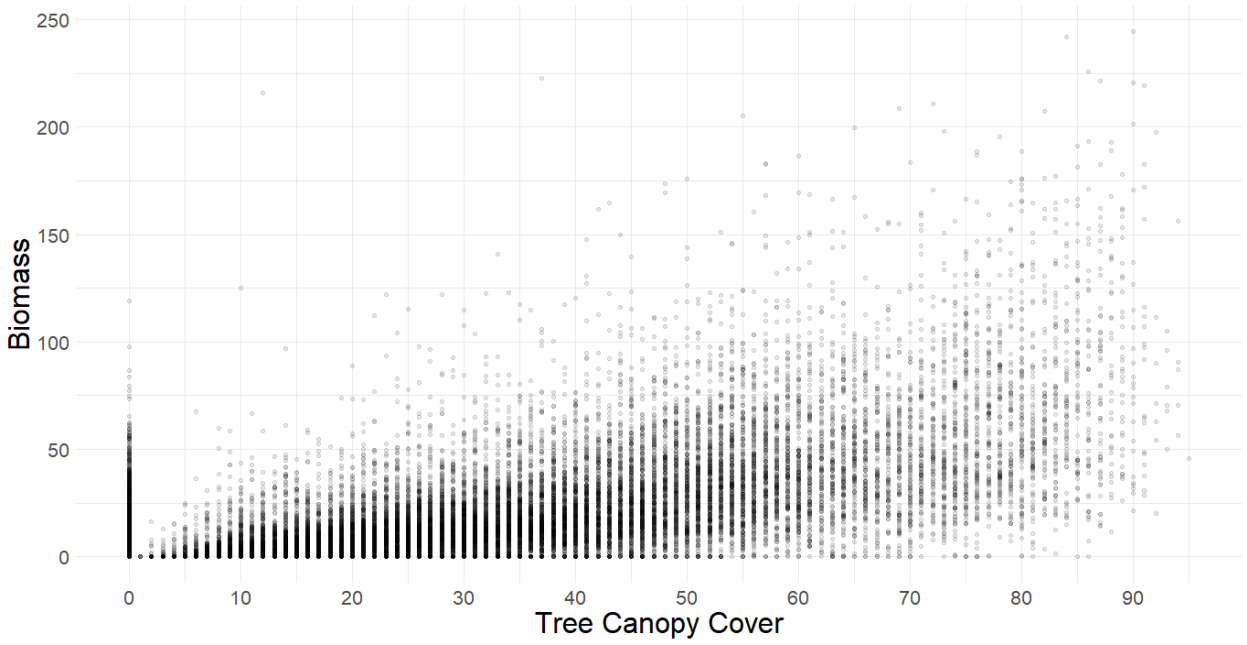

Figure 2: Bivariate Plot of Canopy Cover and Biomass

3.3 Key Variables

Estimation Variables

Our focus is on estimating average value of four forest attributes by county: basal area (square-foot), trees per acre, above-ground biomass (pounds), and net volume (cubic-foot), to evaluate the relative variance of the GREG estimator to the current post-stratification method utilized by FIA. These variables are extrapolated and summed to the plot-level using only live trees. These forest attributes were chosen because they are frequently analyzed, and highly correlated to other forest attributes of interest. Therefore, estimates of one variable are indicative to the estimates of the other variables.

Figures 3, 4, 5, and 6 display the average values of the variable of interest for each plot in a county. Note that the counties with higher values in one variable, for example biomass, will generally have higher values in the other variables, count of live trees, basal area, and volume of live trees.

Predictors

To estimate our four variables of interest, we considered seven different pixel-level variables: Forest probability, Forest biomass (Mg/ha), Forest Type Groups, Elevation (ft), Longitude

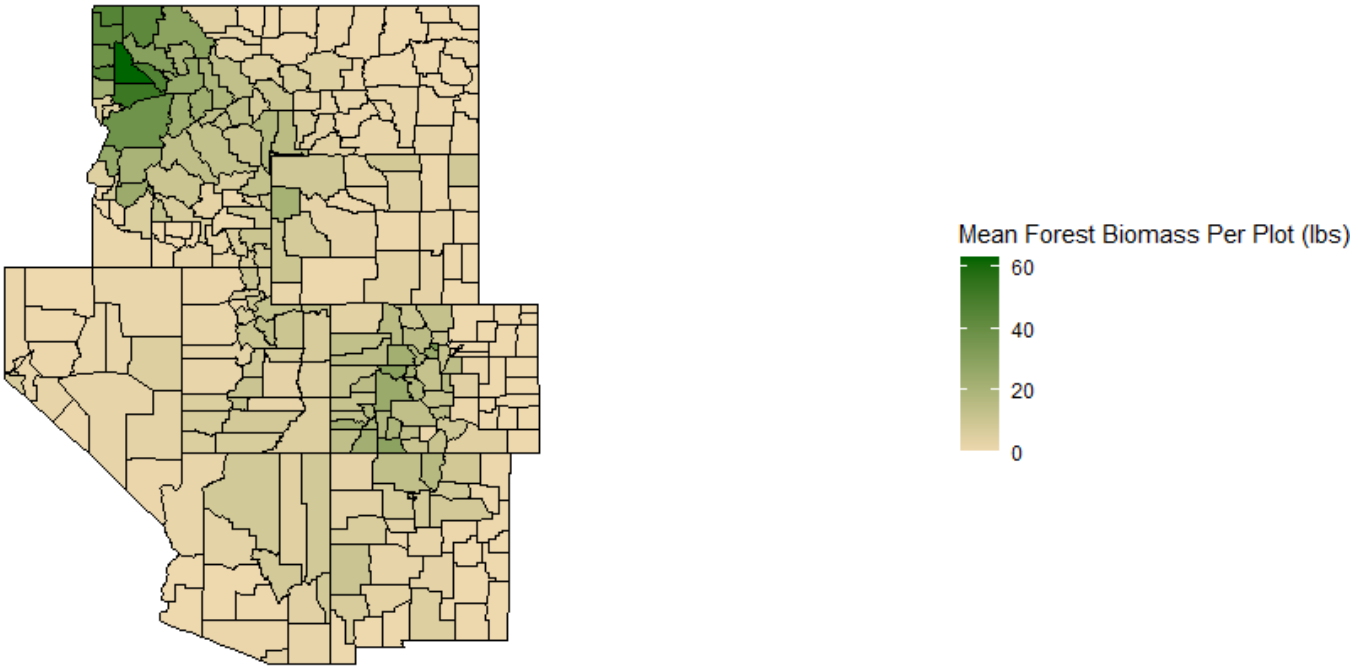

Figure 3: Average Biomass (lbs) per Plot by County

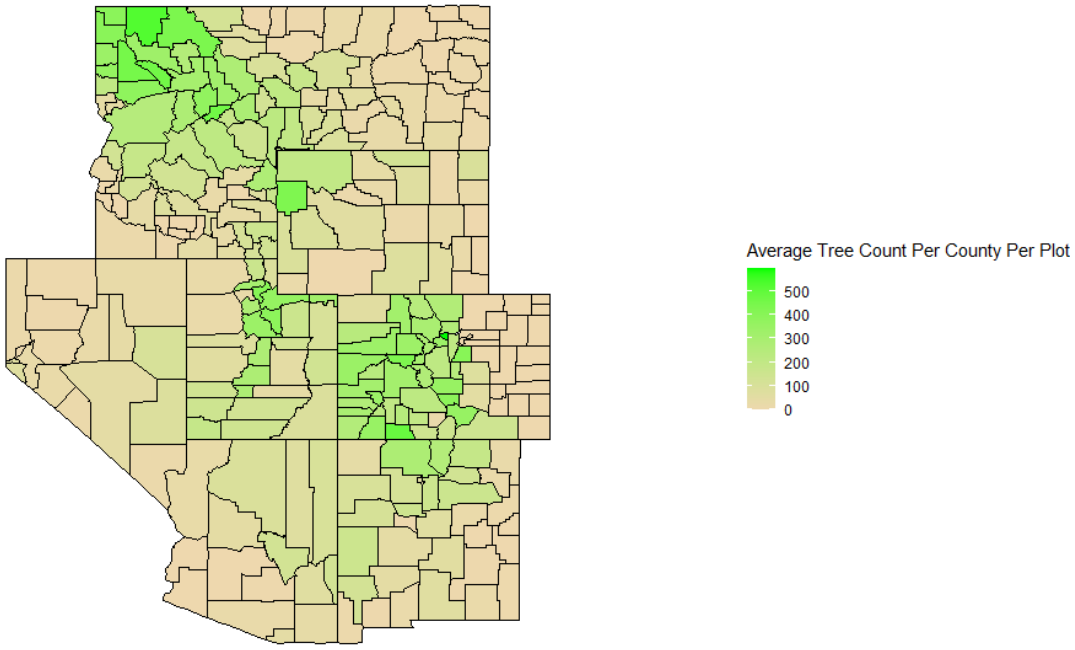

Figure 4: Average Tree Count per Plot by County

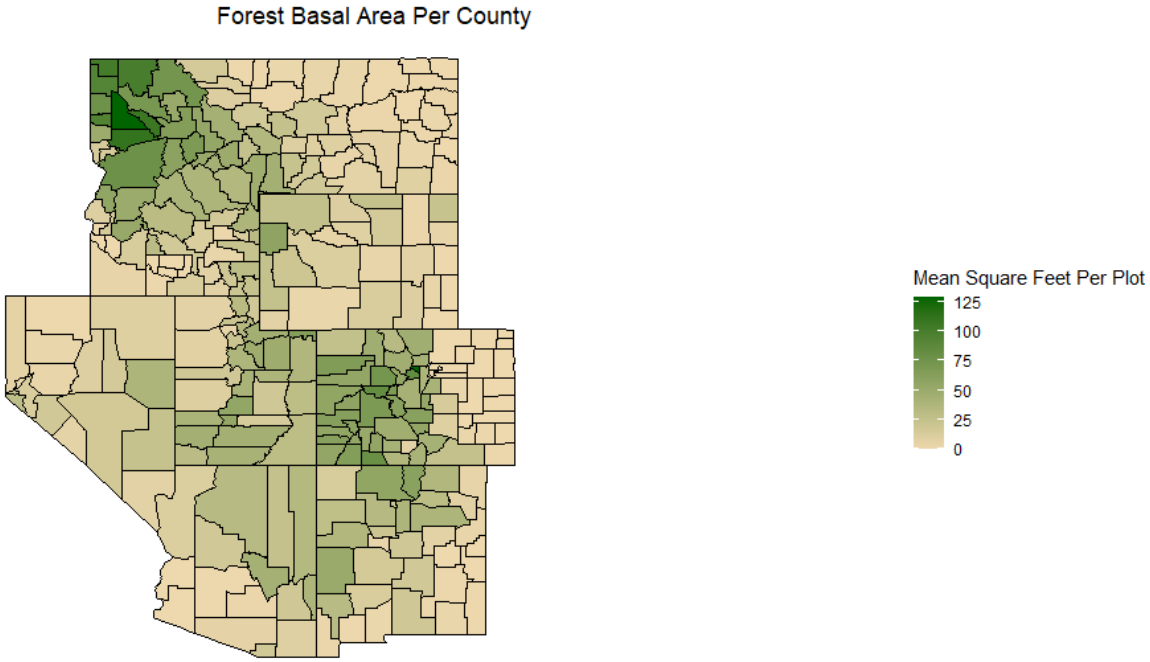

Figure 5: Average Basal Area (square ft) per Plot by County

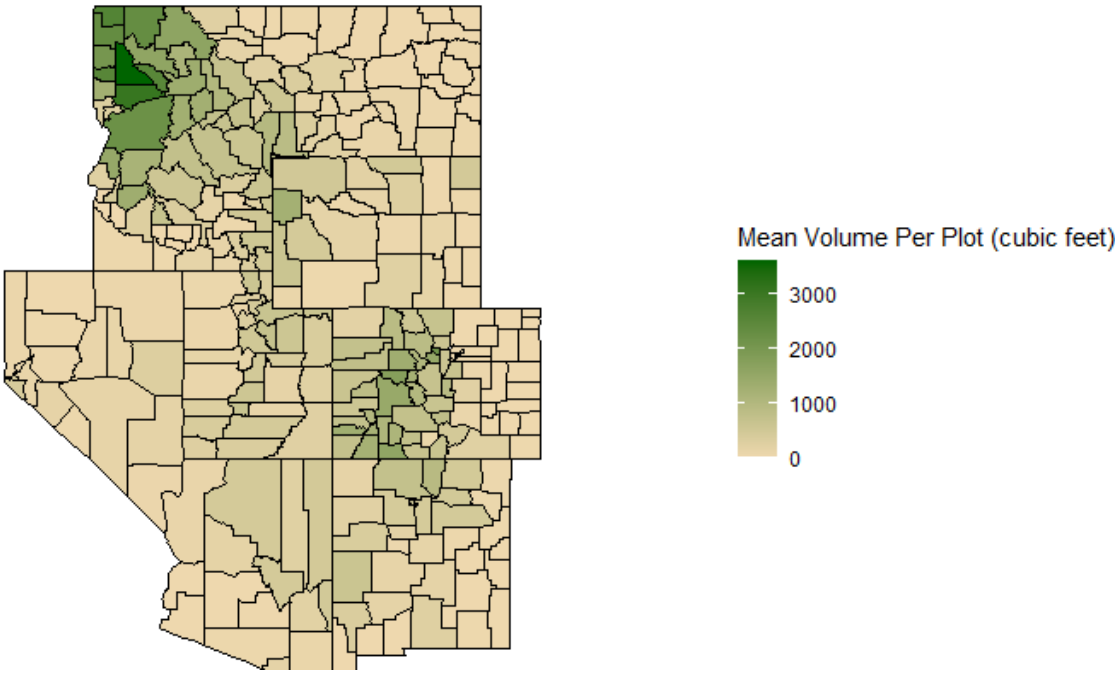

Figure 6: Average Volume of Live Trees (cubic ft) per Plot by County

(decimal degrees, NAD83), Latitude (decimal degrees, NAD83), and Tree Canopy Cover Percentage (NLCD).

We provide additional details on a select few of our predictors. These predictors are further detailed in McConville et al. 2020 and other related forestry literature (McConville, Moisen, and Frescino 2020).

- **Forest Type Groups:** Categorical method of tree classification (Ruefenacht et al. 2008).
- **Forest probability:** An pixel-data based estimate of the probability of forest within a pixel (Blackard et al. 2008).
- **Biomass:** A pixel-data based estimate of above ground biomass (Blackard et al. 2008).
- **Tree canopy cover:** An estimate of percent of plot covered in tree canopy, using spatial data, spatial resolution of 30 m (Homer et al. 2015).

Estimation Levels

We built models at five different resolutions for our variables of interest: Total Interior West (IW), State-level, County-level, Ecology Province, Ecology Section. We will, regardless of resolution, get estimates of the forest parameter (average) at the county level. We included ecological data by utilizing the structure of the ecological codes available at a plot-level (McNab et al. 2007).

3.4 Data Preparation

Prior to creating any estimates, we omitted missing values. Out of the 65860 intial observations in the dataset, 103 contained variables with missing values. However, 99.94% of our original data remain.

Regarding key variables, we concatenated state and county code to denote the unique county for a specified state, commonly known as a fips code. We also binned the unique ecology of each plot, utilizing the structure of the ecological codes (McNab et al. 2007). To make ecologies generalizable, we took the string denoting a plot's ecology and used the first 4 to 5 elements to denote the section and the first 3 to 4 elements to denote the province. The range of elements depends on whether an ecology is an Alpine Meadow Province, and an Alpine Meadow Province has one additional element—an M—at the beginning of its code.

Ecology Example:

Each ecological code can be broken down into section and province, using a subset of the ecological code. For example, for an Alpine Meadow Province, we have:

Subsection, most granular: M332Ba: Bitterroot Glaciated Canyons Subsection

Section: M332B: Northern Rockies and Bitterroot Valley Section

Province, least granular: M332: Middle Rocky Mountain Steppe – Coniferous Forest - Alpine Meadow Province

Additionally, we split our data into a training and test set, and selected 20% of counties at random for the test set and the remaining 80% for training purposes. We verified that the contents of the test set included every unique ecological provinces. We split the data to further explore and validate predictor selection after analyzing what resolution to build our estimates.

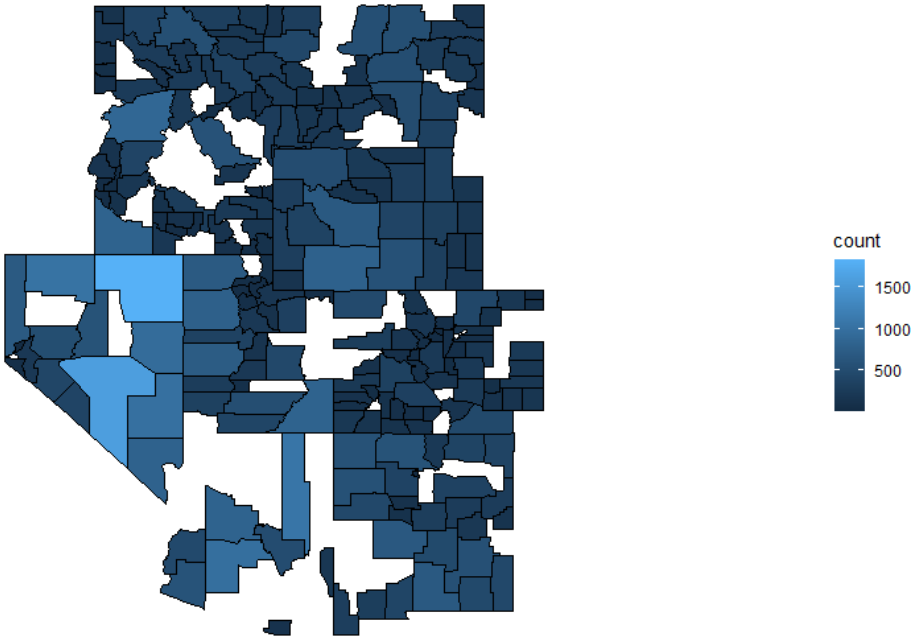

Figure 7: Counties Selected to Training Dataset

4 Results

Our primary objective was to determine the relative efficiency of the GREG in comparison to the PS for estimating forest attributes at the county-level in the IW for the four forest inventory variables: biomass, count of live trees, volume of live trees, and basal area. Our objective was to improve the estimation of these variables by using the GREG estimator and comparing it to the typically used post-stratification estimator. In the linear model used in the GREG, we determine which auxilliary variables to include, along with which resolution the linear model should be constructed at, choosing from the IW, state, province, ecosection, and county level. We then determined which estimator was more precise by comparing the bootstrap standard errors across county-level estimates.

We began with 7 predictors of interest, and evaluated all possible combinations of these 7 predictors on a training, test, and total data set. In the following section, we walk-through

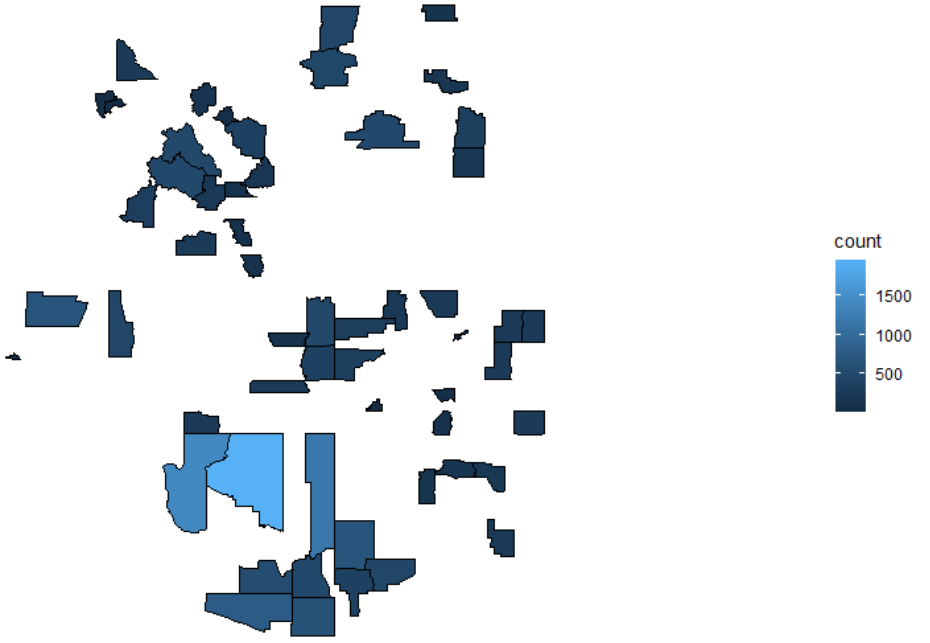

Figure 8: Counties Selected to Test Dataset

our process specifically for biomass. We focus on biomass solely for brevity, though we found the results held across our four variables of interest. This was relieving but unsurprisingly, given our four variables of interest are all positively correlated with one another.

4.1 Linear Model Testing

Once we had decided on our 7 predictors of interest—forest probability, biomass, forest group, percent canopy cover, elevation, latitude, and longitude—we evaluated what combination of predictors to use for GREG estimation. In addition to predictor selection, we were also interested in what resolution level to construct our predictions. Our resolution levels of interest were: state, county, ecological section, and ecological province. To compare resolution levels, we analyzed the variability of model fit—Adjusted R^2—across units of each resolution. In doing so, we analyzed how well our linear models estimated biomass across counties, and if those estimates were more consistent (had less variance) than models estimating biomass across ecological provinces.

At this stage of the analysis, we split the data into a test and training set. In the following analysis, we were considering models specifically estimating the test data.

To analyze resolution level, we considered all possible combinations of predictors—a total of 127 predictor combinations. We ran the set of 127 models across all units of all resolution levels, collecting the Adjusted R^2 values for each model. We then looked at the distribution of the Adjusted R^2 values grouped by resolution level, as shown in 9. We also looked at the spread of Adjusted R^2 values, as shown in 10. From this, we see estimates at the state-level

had the lowest standard deviation of adjusted R^2s, and they tend to have comparatively higher Adjusted R^2 compared to other resolutions. However, as our ultimate objective was to create county-level estimates, we wanted more granularity, and thus decided to use the ecological province resolution level due to its comparatively high Adjusted R^2 values and low standard deviation of Adjusted R^2s.

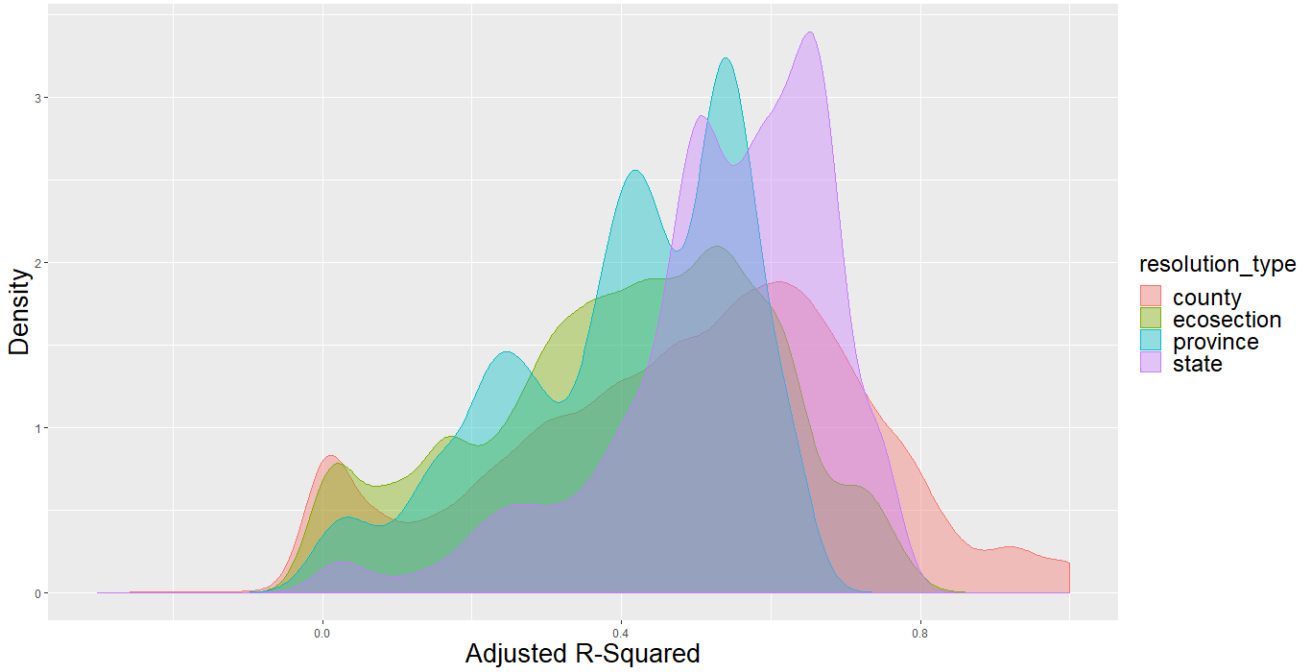

Figure 9: Distribution of Adj. R2 by Resolution for Biomass

Once we decided to build estimates at the province-level, we then evaluated the best set of predictors to use for our province estimates. We looked at the Adjusted R^2 values by size (number of predictors), in addition to the best models by size that include latitude, longitude, or elevation. We then applied this methodology across all variables of interest. Though 11, and our analysis in this section is specific to estimating Biomass, we found the 17 models outlined in 11 applied for all variables of interest. Additionally, given only one single predictor model—using canopy cover—was included at this stage, we decided to add one more single predictor model—forest probability.

Once we decided on our 18 simple linear models, we then added interaction terms for latitude, longitude, and elevation. However, these interaction terms were only added to models that had the initial predictors already in them. For example, for our five variable model that includes latitude and elevation, we consider this model in addition to the same model with the interaction of latitude and elevation, but would not add an interaction term to any of the single-predictor models.

By including interaction terms, we then considered 22 unique linear models. We ran these 22 models across each province, looking at Adjusted R^2 and BIC to determine predictor selection for the GREG. At this stage. we repeat the same process on our test set and total

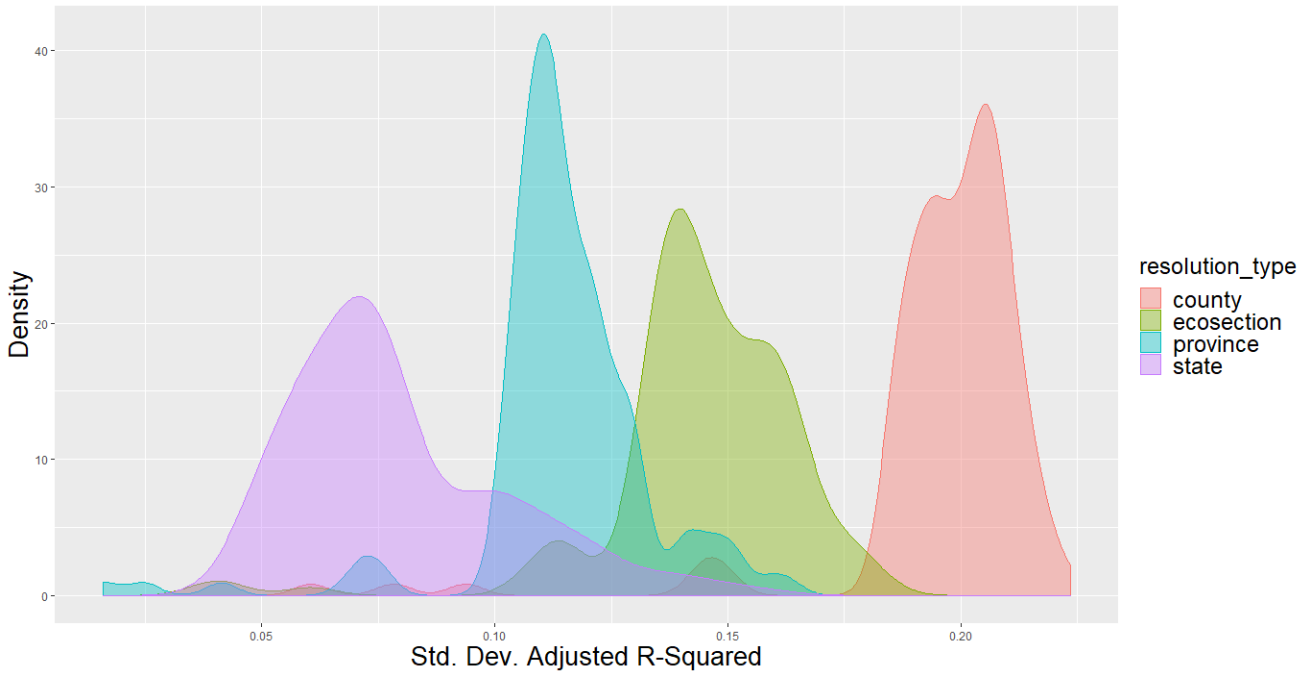

Figure 10: Std. Dev. of Adj. R2 by Resolution for Biomass

data set, comparing values across dataset. Of note, *'NAs'* were removed from the test set and are omitted from 12 and 13.

From 12 and 13, we evaluated our linear models, looking at which models maximized Adjusted R^2 and minimized BIC.

Motivated by the results of 12 and 13, we ultimately decided on 5 models to use for the GREG:

1. Forest Probability, Biomass, Percent Canopy Cover, Elevation, Latitude, Longitude, & All Elevation-Latitude-Longitude Interactions
2. Forest Probability, Biomass, Percent Canopy Cover, Elevation, Latitude, & Longitude
3. Forest Probability, Biomass, & Percent Canopy Cover
4. Percent Canopy Cover
5. Forest Probability

We were surprised the model with all predictors, the *'Kitchen Sink'*, continued to outperform simpler models in terms of Adjusted R^2 and BIC, statistics constructed to penalize the addition of predictors. To further test these results, we decided on including three simpler models—the single predictor models and the 3 predictor model. Additionally, due to availability of pixel-level data, we decided on models that did not utilize the forest group variable. We hope the inclusion of this predictor will be considered in future research.

Model	Size	Mean Adj. R2
6 Variable Model (No Forest Group)	6	0.52
Kitchen Sink	7	0.52
5 Variable Model (No Forest Group, Latitude)	5	0.52
5 Variable Model (No Forest Group, Longitude)	5	0.52
4 Variable Model (No Forest Group, Longitude, Latitude)	4	0.51
5 Variable Model (No Forest Bio., Group)	5	0.51
4 Variable Model (No Forest Bio., Group, Longitude)	4	0.51
Forest Prob., Elevation, Canopy Cover	3	0.51
4 Variable Model (No Forest Bio., Group, Elevation)	4	0.50
Forest Prob., Canopy Cover	2	0.49
4 Variable Model (No Forest Prob., Bio., Group)	4	0.45
Elevation, Latitude, Canopy Cover	3	0.44
Longitude, Latitude, Canopy Cover	3	0.42
Canopy Cover	1	0.41
Elevation, Longitude, Latitude	3	0.16
Elevation, Latitude	2	0.14
Longitude, Latitude	2	0.05

Figure 11: Best Simple Linear Models for Biomass

Model	Size	Adj. R2 Total	Adj. R2 Train	Adj. R2 Test
6 Variable Model With Elevation, Latitude Interaction	11	0.66	0.63	0.69
Kitchen Sink (With Int.)	10	0.66	0.63	0.69
4 Variable Model (No Forest Group, Elevation, Longitude)	7	0.65	0.62	0.68
4 Variable Model (No Forest Group, Longitude, Latitude)	6	0.65	0.62	0.68
5 Variable Model with Elevation, Latitude Interaction	6	0.65	0.62	0.68
5 Variable Model With Elevation, Latitude Interaction	7	0.65	0.62	0.68
Forest Prob., Group, Elevation, Canopy Cover	6	0.65	0.62	0.68
Forest Prob., Bio., Canopy Cover	6	0.65	0.62	0.68
5 Variable Model (No Forest Bio., Longitude)	5	0.65	0.62	0.68
Forest Bio., Canopy Cover	5	0.65	0.62	0.68
Forest Prob., Elevation, Canopy Cover	4	0.65	0.62	0.70
Forest Prob., Canopy Cover	4	0.65	0.62	0.70
Forest Group, Canopy Cover	3	0.64	0.62	0.67
6 Variable Model, All Interactions	6	0.64	0.62	0.67
Canopy Cover	5	0.64	0.62	0.67
Forest Prob.	4	0.63	0.60	0.68
Kitchen Sink (No Int.)	3	0.63	0.60	0.65
6 Variable Model (No Forest Group)	2	0.63	0.60	0.66
5 Variable Model With Elevation, Longitude Interaction	2	0.60	0.57	0.61
6 Variable Model (No Longitude)	2	0.57	0.55	0.58
5 Variable Model (No Forest Group, Longitude)	1	0.54	0.52	0.55
5 Variable Model (No Forest Group, Latitude)	1	0.53	0.51	0.53

Figure 12: Linear Model Comparison by Adj R2 for Biomass

Model	Size	BIC Total	BIC Train	BIC Test
6 Variable Model With Elevation, Latitude Interaction	11	55118.96	41888.10	14079.74
Kitchen Sink (With Int.)	10	55133.33	41894.97	14082.54
4 Variable Model (No Forest Group, Elevation, Longitude)	7	55146.86	41903.21	14082.91
4 Variable Model (No Forest Group, Longitude, Latitude)	6	55160.89	41910.65	14085.27
5 Variable Model with Elevation, Latitude Interaction	6	55179.46	41925.66	14085.19
5 Variable Model With Elevation, Latitude Interaction	7	55154.11	41913.87	14078.88
Forest Prob., Group, Elevation, Canopy Cover	6	55169.96	41922.01	14082.00
Forest Prob., Bio., Canopy Cover	6	55180.35	41933.91	14082.56
5 Variable Model (No Forest Bio., Longitude)	5	55195.24	41941.45	14085.54
Forest Bio., Canopy Cover	5	55195.37	41942.27	14084.92
Forest Prob., Elevation, Canopy Cover	4	55236.72	41969.90	13086.02
Forest Prob., Canopy Cover	4	55231.20	41966.43	13085.57
Forest Group, Canopy Cover	3	55236.08	41968.77	13085.98
6 Variable Model, All Interactions	6	55258.46	41959.52	14142.37
Canopy Cover	5	55294.35	41989.17	14143.26
Forest Prob.	4	55351.97	42025.35	13147.50
Kitchen Sink (No Int.)	3	55368.15	42036.34	13151.77
6 Variable Model (No Forest Group)	2	55368.74	42034.95	13149.56
5 Variable Model With Elevation, Longitude Interaction	2	55935.54	42444.99	13291.06
6 Variable Model (No Longitude)	2	56313.19	42641.35	13455.66
5 Variable Model (No Forest Group, Longitude)	1	56586.30	42803.10	13574.59
5 Variable Model (No Forest Group, Latitude)	1	56817.69	43188.79	13489.01

Figure 13: Linear Model Comparison by BIC for Biomass

4.2 Predictor Selection

We found, as shown in 9, that using a state-level resolution had the lowest standard deviation of adjusted R^2s, but as we wanted more granularity for the purposes of estimation, we decided to use the second lowest resolution level: Ecological Province.

Using the procedure outlined in The GREG estimator, we developed GREG using a combination of the following predictors:

1. Forest Probability, Biomass, Percent Canopy Cover, Elevation, Latitude, Longitude, & All Elevation-Latitude-Longitude Interactions
2. Forest Probability, Biomass, Percent Canopy Cover, Elevation, Latitude, & Longitude
3. Forest Probability, Biomass, & Percent Canopy Cover
4. Percent Canopy Cover
5. Forest Probability

We found that generally, the GREG estimator had a higher level of precision in estimating our four variables of interest in the majority of counties in the IW. Furthermore, the counties in which there was not improvement were counties with a miniscule amount of the variable of interest.

As the conclusions drawn from tables 14 - 16 are similar to one another, for the remainder of the paper future tables and figures focus solely on one variable, Basal Area.

	Canopy Cover	Forest Probability	Three Variable	Kitchen Sink (No Int.)
Mean Post Strat Bootstrap SE	0.84	0.84	0.84	0.84
Mean GREG SE	0.69	0.80	0.67	0.81
Mean Relative Efficiency	1.02	0.91	1.08	0.84
Fraction of Counties Improved	0.66	0.56	0.71	0.46

Figure 14: Mean Standard Errors, Relative Efficiencies, and Fraction of Counties Improved for Biomass

	Canopy Cover	Forest Probability	Three Variable	Kitchen Sink (No Int.)
Mean Post Strat Bootstrap SE	2.31	2.31	2.31	2.31
Mean GREG SE	1.98	2.21	1.90	2.35
Mean Relative Efficiency	1.02	0.93	1.07	0.85
Fraction of Counties Improved	0.65	0.60	0.76	0.44

Figure 15: Mean Standard Errors, Relative Efficiencies, and Fraction of Counties Improved for Basal Area

	Canopy Cover	Forest Probability	Three Variable	Kitchen Sink (No Int.)
Mean Post Strat Bootstrap SE	16.08	16.08	16.08	16.08
GREG Mean SE	15.62	15.97	15.53	18.81
Mean Relative Efficiency	0.90	0.88	0.91	0.70
Fraction of Counties Improved	0.53	0.47	0.54	0.20

Figure 16: Mean Standard Errors, Relative Efficiencies, and Fraction of Counties Improved for Count of Live Trees

	Canopy Cover	Forest Probability	Three Variable	Kitchen Sink (No Int.)
Mean Post Strat Bootstrap SE	49.52	49.52	49.52	49.52
Mean GREG SE	41.87	47.57	40.19	48.44
Mean Relative Efficiency	0.99	0.91	1.05	0.82
Fraction of Counties Improved	0.68	0.55	0.71	0.44

Figure 17: Mean Standard Errors, Relative Efficiencies, and Fraction of Counties Improved for Volume of Live Trees

	Canopy Cover	Forest Probability	Three Variable	Kitchen Sink (No Int.)
Mean Post Strat Bootstrap SE	2.93	2.93	2.93	2.93
Mean GREG SE	2.44	2.77	2.34	2.82
Mean Relative Efficiency	1.17	1.06	1.23	1.00
Fraction of Counties Improved	0.79	0.71	0.92	0.54

Figure 18: Mean Standard Errors, Relative Efficiencies, and Fraction of Counties Improved for Basal Area, excluding counties in the bottom 20 percent of Basal Area

Tables 14 - 17 displays the mean bootstrap standard errors for the different estimators and the percent of counties whose estimates showed a increased level of precision when switching from the PS to the GREG. All forms of the GREG, except for the one that includes all of the predictors have a lower mean standard error than the PS estimator for each variable. The best form of the GREG that has the lowest mean standard errors in each of our variables of interest was the GREG that included Forest Probability, Biomass, & Percent Canopy Cover. The variables that experiences the highest gains from utitlizing the GREG are Basal Area, Biomass, and Volume of Live Trees, over 70% of counties in the IW experiences decreased variances in their predictions.

When we remove the bottom 20% of counties with the lowest counts for each variable of interest (using plot data), and shown in table 18, our GREG outperforms the PS to a greater extent. Implementing this county-level adjustment removes counties with 0 counts for each variable of interest, instances where the PS would have SE's of 0.

Note: When counties were removed due to being in the bottom 20% of a variable of interest, this resulted in approximately 23% of all counties considered to be removed.

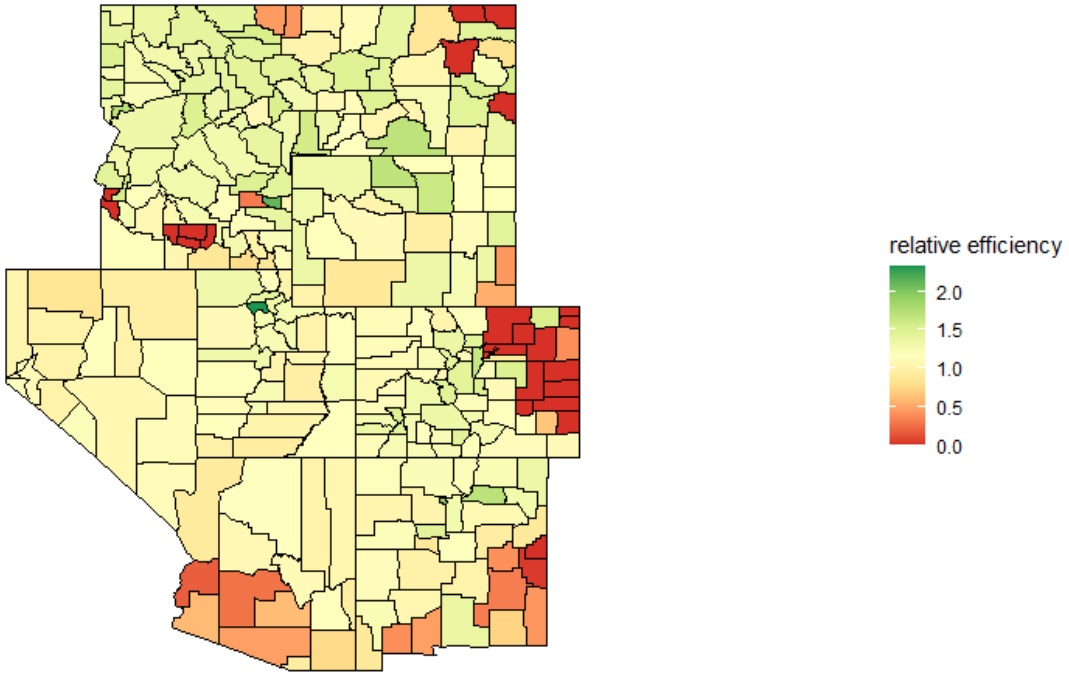

Figure 19: Relative Efficiency of the 3 variable GREG by County for Basal Area

From Figures 19 and 21, we display the Relative Efficiency (RE) of the GREG with Forest Probability, Biomass, & Percent Canopy Cover. As noted earlier, we see improvements in precision in roughly 70% of counties for every variable except for Count of Live Trees. In roughly 30% of counties, however, there was no improvement over the post-stratification method. From the maps above, we see that the counties that show no improvement are generally the counties with 0 counts for their variables of interest. Additionally, the GREG with Canopy Cover Percentage had a majority of estimates with greater precision than the

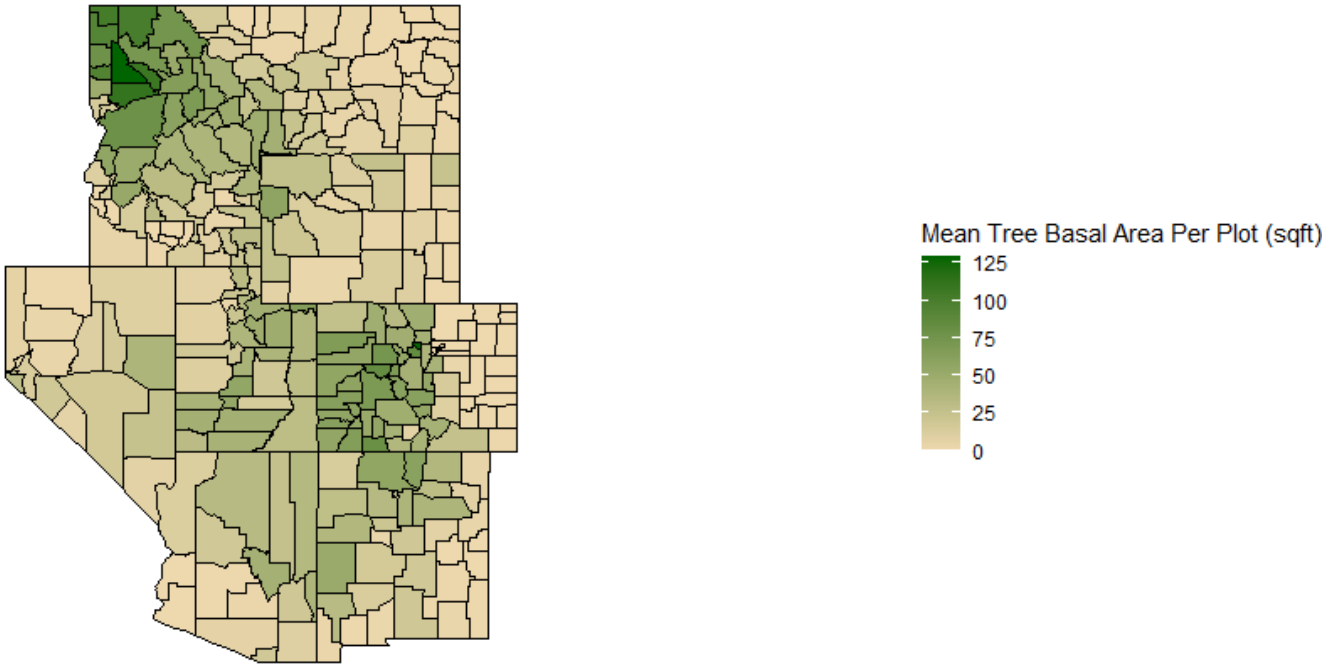

Figure 20: Average Basal Area (square ft) per Plot by County

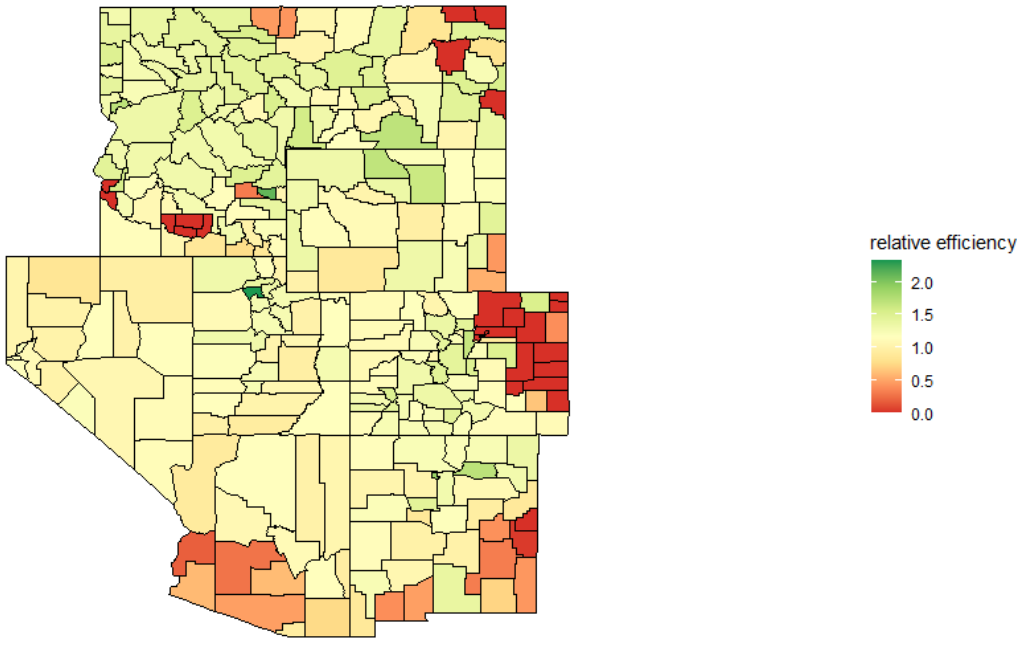

Figure 21: Relative Efficiency of the 3 variable GREG by County for Basal Area

post-stratification estimator for every variable of interest, though this model was not as precise as the 3 variable GREG.

5 Conclusion

We found that using province-level estimates to create county-level GREG estimates were generally more efficient in predicting forest attributes compared to the PS method utilized by the FIA. GREG estimates were particularly more efficient in estimating counties with forested areas. Though we found the GREG estimator to be more efficient than the PS estimator in terms of number of counties, there were some counties where PS was better—namely counties where we had less data and counties with none to little amount of our forest attributes of interest.

In the future, we hope to explore additional aspects of our research. As our GREG estimates were relatively inefficient at estimating smaller counties—both in terms of plot-level data and pixel-level data—we hope to explore the possibility of creating conglomerate counties. There were also a number of GREG estimates that were negative. As our variables of interest have a lower bound at zero, we hope to create a methodology for consistently dealing with these cases.

There is also an opportunity to assess county-level weights on their impact on model selection, as our initial model validation methods—BIC and Adjusted R^2—led us to initially include a model that produced poor GREG estimates, compared to other GREG models as well as the PS. In addition to the impact of predictor interactions, there are further opportunities to consider other predictive variables of interest that we've omitted, such as forest group.

Our work outlined here is only one part of an investigation of estimating forest attributes, with application of the PS and GREG estimators. Beyond the estimators explored in this paper, our work implies that estimators of all kinds deserve further exploration to find the best for estimating forest characteristics and assessing under what conditions to use and improve them.

References

Bechtold, William A., and Paul L. Patterson. 2015. "The Enhanced Forest Inventory and Analysis Program National Sampling Design and Estimation Procedures." SRS-GTR-80. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. https://doi.org/10.2737/SRS-GTR-80.

Blackard, J, M Finco, E Helmer, G Holden, M Hoppus, D Jacobs, A Lister, G Moisen, M Nelson, and R Riemann. 2008. "Mapping U.S. Forest Biomass Using Nationwide Forest Inventory Data and Moderate Resolution Information." *Remote Sensing of Environment* 112 (4): 1658–77. https://doi.org/10.1016/j.rse.2007.08.021.

"Forest Inventory and Analysis National Program - About Us." n.d. Accessed May 9, 2020. https://www.fia.fs.fed.us/about/about_us/.

Homer, Collin, Jon Dewitz, Limin Yang, Suming Jin, Patrick Danielson, John Coulston, Nathaniel Herold, James Wickham, and Kevin Megown. 2015. "Completion of the 2011 National Land Cover Database for the Conterminous United States – Representing a Decade of Land Cover Change Information." *PHOTOGRAMMETRIC ENGINEERING*, 10.

"Interior West Forest Inventory & Analysis - About Us." n.d. Accessed May 9, 2020. https://www.fs.fed.us/rm/ogden/about/index.shtml.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning: With Applications in R*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/.

Magnussen, Steen, and Lutz Fehrmann. 2019. "In Search of a Variance Estimator for Systematic Sampling." *Scandinavian Journal of Forest Research* 34 (4): 300–312. https://doi.org/10.1080/02827581.2019.1599063.

McConville, Kelly S., Gretchen G. Moisen, and Tracey S. Frescino. 2020. "A Tutorial on Model-Assisted Estimation with Application to Forest Inventory." *Forests* 11 (2): 244. https://doi.org/10.3390/f11020244.

McConville, Kelly, Becky Tang, George Zhu, Sida Li, Shirley Chueng, and Daniell Toth (Author and copyright holder of treeDesignMatrix helper function). 2018. "Mase: Model-Assisted Survey Estimators." https://CRAN.R-project.org/package=mase.

McNab, W. H., D. T. Cleland, J. A. Freeouf, J. E. Keys, G. J. Nowacki, and C. A. Carpenter. 2007. "Description of Ecological Subregions: Sections of the Conterminous United States." WO-GTR-76B. Washington, DC: U.S. Department of Agriculture, Forest Service. https://doi.org/10.2737/WO-GTR-76B.

Ruefenacht, B., M. V. Finco, M. D. Nelson, R. Czaplewski, E. H. Helmer, J. A. Blackard, G. R. Holden, et al. 2008. "Conterminous U.S. And Alaska Forest Type Mapping Using Forest Inventory and Analysis Data." *Photogrammetric Engineering & Remote Sensing* 74 (11): 1379–88. https://doi.org/10.14358/PERS.74.11.1379.