A Forest for the Trees: Using Random Forests for Small Area Estimation on US Forest Inventory Data

A Forest for the Trees: Using Random Forests for Small Area Estimation on US Forest Inventory Data

by Julian F. Schmitt

Advised by Kelly McConville

A senior thesis presented to the

Department of Applied Mathematics

in partial fulfillment for Honors in

Applied Mathematics in Earth and Planetary

Sciences

Department of Applied Mathematics

Harvard University

Cambridge, Massachusetts

March 24^{th}, 2023

Contents

Abstract

Methods which estimate population parameters of interest across small areas is a growing field of research. These problems arise frequently in election prediction, healthcare monitoring, and environmental studies. The Forest Inventory and Analysis Program (FIA) of the US Forest Service tracks forest metrics, such as basal area and above ground carbon, to ensure sustainable stewardship of the nation's forests and preserve her resources for future generations. Their estimates combine expensive ground plot observations of the variables of interest alongside inexpensive and plentiful auxiliary data collected by remote sensing. Historically, estimators in this setting either rely on means or linear parametric models, such as the post-stratified estimator, area-level empirical best linear unbiased predictor (area-EBLUP), and unit-level empirical best linear unbiased predictor (unit-EBLUP) models. Here, we present the results of a simulation study to compare these standard estimators to a new problem-specific estimator, as well as machine learning models. The problem-specific zero-inflated estimator is introduced to address the overabundance of zero observations in FIA ground plot observations, while machine learning methods, including the random forest and mixed-effects random forest (SMERF) seek to flexibly capture non-linear relationships between the predictors and the response variable to improve performance while also addressing the zero-inflation problem. We track both bias and root mean squared error across the six estimators to assess their performance and find that there is no universal "best model." Instead we find a complex story in which the post-stratified and area-EBLUP models have exceptionally low bias, particularly across areas with low-carbon levels however when examining root mean squared error, the zero-inflation model performs well. Across higher carbon levels model performance is even more complex. We close with implications for these results alongside avenues to improve estimation at scale across the US.

1 Introduction

1.1 Why Study Small Areas?

Understanding local behavior is important. Insights between subgroups, defined either geographically or demographically, is increasingly valuable in research areas including election prediction, healthcare, and the environment. For instance, the makeup of politically representative groups is decided by the behavior of

voters in local electorates. Gerrymandering, a tool used to reshape districts to advantage a candidate or party, takes advantage of the small area phenomenon, and can be used to suppress the influence of racial groups [1]. The effect of gerrymandering on the composition of congress has been simulated by both Friedman & Holden [2] and Chen & Cottrell [3]. Political surveys, or polls, are often used to predict these elections or understand public sentiment, and are similarly affected by the small area phenomenon. Often either samples are collected with a response-bias that is hard to eliminate or estimators that are asymptotically unbiased but perform poorly on small samples are applied; both provide false confidence in the poll's predicted outcome [4]. In pharmaceutical trials, understanding the influence of a drug across strata based on age, sex, and racial groups is critical to developing effective treatments; negative treatment effects can hide in a sample mean [5, 6]. Alternatively, improving the resolution of weather models is similarly essential to solve the fluid dynamics processes that exist on a continuum; people like to know whether it's going to rain on their walk through Harvard Yard after class lets out at 11:45am [7].

While estimation techniques vary across domain, central to each problem is the lack of an ability to collect a population census. In pharmaceutical clinical trials, we observe the drug's effect on only a few members of a population. In polls, we can call only a few hundred people for a political survey. In weather prediction, there are a finite number of stations and satellites collecting observations. Worse still, the data collected are often biased. To extrapolate these findings more accurately from the sampled population, or begin correcting these biases, we need small area estimation (SAE). SAE seeks to do estimation over population sub-groups when the sample sizes in these groups are typically too small for estimators which rely solely on the data in the sub-group. To achieve this, SAE models combine the observations we can make with other information about the larger population to improve the precision and accuracy of estimates.

1.2 The Forestry Setting

In the forestry setting, researchers are interested in variables which indicate the health or size of the forest, like biomass, live carbon, or basal area; the values of these variables are most useful at high granularity across the study domain. However, a census or near-census of trees across a large region, like United States, is impossible. Methodological analyses such as in Spawn *et al.* [8] and Goetz *et al.* [9], rely purely on

abundant satellite imagery with tools such as radar, light detection and ranging (LiDAR), and optical tools to study carbon across the global domain. However, satellite, or "auxiliary," data, such as GEDI [10], is notorious for high estimate variance; Goetz *et al.* [9] found an order of magnitude difference in carbon measurements between a direct remote sensing approach and a "combine and assign" approach. On the other hand, estimates made by researchers on the ground have much lower error than the remote sensing layers; combining the precision of the measurements made by hand alongside the breadth of satellite data using SAE methods has the potential to produce more reliable estimates. In the United States, the US Forest Service Forest Inventory and Analysis Program (FIA) collects ground data across the US using a random sampling design. We conclude that the combination of remote sensing layers and FIA ground plots can significantly improve the accuracy of estimates of variables of interest, like forest carbon.

These observations are particularly useful for estimating forest attributes across small areas. Interest from local government and foresters has increasingly placed emphasis on estimation at increasingly small areas, such as at the county or town level [11]. Further directives from Congress, including the 2014 and 2018 UDSA Farm Bills, order the FIA to: "implement procedures to improve the statistical precision of estimates at the sub-State level" and to do so by integrating "advanced remote sensing technologies to provide estimates for State- and national-level inventories" [12, 13]. The USDA [13] further states that forests should be used to "increase carbon sequestration." These congressional directives further motivate our foray into understanding forest carbon across small areas through the combination of advanced remote sensing layers and FIA-based ground plots both towards improving an inventory of the nations resources and as a means to address climate change.

The directive to examine smaller areas means that models must rely on with fewer available sampled plots to estimate forest metrics. Naturally, a reduction in observations, means the traditional survey estimators will have higher variance. The shift places increasing pressure on the estimators to perform well with fewer observations. Current data used to estimate forest metrics across the United States is gathered using a quasi-systematic sampling design. Under this design, the FIA partitions the continental U.S. into 6000-acre hexagons and then randomly samples a location from within these hexagons for measurement once every 5 or 10 years. The samples are catalogued as "plot-level" data [14]. The remote sensing data typically includes

climate (e.g. temperature and precipitation) and geomorphological (e.g. elevation and terrain ruggedness) measures, as well as satellite measured metrics like tree canopy cover. In combination with remote sensing data, the FIA uses these plot-level observations to build its estimates of forest attributes [14]. Common forest attributes of interest include the number of tree stems per acre, forest carbon, and basal area. We focus on forest carbon, specifically the amount of carbon stored in the above-ground portion of tree (abbreviated "carbon"). Currently the FIA uses the post-stratified estimator to produce official estimates of these metrics [14]. The post-stratified estimator performs well when the small areas contain a large number of samples, however the efficiency of the estimator decreases rapidly when only a few sample plots are available in a small area of interest. For the post-stratified estimator, the model employed to combine the plot and auxiliary data is the simple one-way ANOVA model. The vast size of the nation's forests in combination with the high cost of collecting plot-level data places increasing pressure on the model as the mechanism by which to improve estimates. Here, we study several SAE techniques, ranging from the methods currently used by the FIA, to simple linear models, to problem-specific estimators, to random forests. One notable pattern in the FIA dataset is the high number of zero observations, or zero-inflation. Traditional linear estimators, like the area- and unit-level EBLUP models, don't utilize this problem-specific knowledge; this leave room for improvement with specialized models. We explore two model types to address the zero-inflation in the data: a two-part linear model to specifically incorporate a prediction of whether the observation is zero, and a machine learning model with the flexibility to capture zero observations directly.

1.3 Literature Review

1.3.1 The Importance of Accurate Forest Estimation

Robust estimation of forests is essential to understanding the changing health and resource availability of national forest ecosystems [15–18]. Beginning in 1950, the FIA used aerial photography and human experts to classify forests and auxiliary information [14]. Since the satellite age, aerial photography analyzed by human eyes has been replaced by a automated system of auxiliary data collection. However, variables of interest, like carbon, remain hard to collect accurately via remote sensing and are still collected using ground crews. The zero-inflation problem arises in the data collection as frequently there are no trees at a location

visible on an aerial photo and so crews will simply impute the values as zero to avoid a wasted site visit (Chapter 2 [14]). As these sites, which include deserts, cropland, and human developments, are common, these datasets have a large number of zero observations which violate the assumptions of some linear models if the proportion is sufficiently large. This motivates the development of new methods for estimation which appropriately handle this aspect of the data. Accurate estimation allows the FIA to provide guidance to policy makers, stakeholders, and industry towards its mission 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" [19].

Accurate estimation of forest carbon is motivated by the forests' role in the climate system. Forests store vast amounts of carbon, roughly 400 gigatons of carbon (GtC), or approximately half of the total atmospheric concentration which is about 849 GtC [20, 21]. The large relative size of total stored carbon to atmospheric carbon means that changes to forest inventories can have a significant influence on atmospheric CO_2 concentrations, which are the main driver of global climate change [22]. Tracking net positive or negative changes across forests globally is essential to better constrain climate change projections. An understanding of these changes begins with the ability to perform robust estimation of current tree stocks. To achieve this state-of-the-art understanding, we should also use state-of-the-art statistical models to perform this estimation. By using robust SAE models to improve forest inventories, applications that deal specifically with estimating changes to forest carbon stores can be improved. For example, several studies seek to estimate the carbon capture potential of reforestation in [15–17]. While initial estimates in [15] suggest upwards of 2/3 of to-date anthropogenic emissions could be stored in tree biomass via restoration projects, many responses have since cut that number down to roughly a $1/10^{th}$ [16]. These restoration projects would need to fill a significant portion of the globe in which trees can grow climatically, which will likely require significant human contribution.

1.3.2 Classical Estimation Frameworks for SAE

Currently the FIA uses the post-stratified estimator (PS) to make their estimates of forest attributes, typically at the state-level (Bechtold *et al.* [14] Chapter 4). PS combines a single auxiliary variable with

ground observations within the small area [14]. Recent literature [23–26] has demonstrated the effectiveness of adding structure and more auxiliary variables to improve model performance, through parameterized models and Bayesian frameworks. Traditional estimators include direct estimators, indirect estimators that use generalized linear models or mixed models, and a small subset of studies exploring machine learning in the forest domain [27, 28]. We focus on comparing our machine learning method to the classical post-stratified estimator (PS), the area-level Empirical Best Linear Unbiased Predictor (area-EBLUP, or AE), the unit-level Empirical Best Linear Unbiased Predictor (unit-EBLUP, or UE), and a two-part mixed model to address zero-inflation (zero-inflated estimator, or ZI). Conceptually they can be thought of in the following frameworks:

1. **Direct Estimators:** Direct estimators include the sample mean and PS estimator. These estimators rely only on sampled data from within the small area. PS is the standard estimator that FIA employs and is asymptotically unbiased and also corrects sampling bias which the sample mean does not address [14]. The PS estimator incorporates one categorical auxillary variable, in addition to the response variable. For example, in a political survey, voter preference may be combined with the respondents age bracket. If voting patterns differ across age groups and our sample isn't representative of the population - perhaps older people are more likely to answer a phone call - then the sample mean is an inaccurate estimate of the true population mean. As the proportion of the stratification variable across the population is generally known, we can use PS to correct the bias. Even in cases where the sample is representative, post-stratifying reduces the variability of the estimator, especially if there is a strong association between the post-strata and the response variable.

2. **Indirect Parametric Estimators:** Indirect estimators borrow strength from data collected both inside and outside the small area of interest. Typically, these models assume a linear structure between the response variable and auxiliary data to generate these results. The AE, UE, and ZI estimators are all parametric models that rely on a linear regression component. The difference between the area-level model, the AE, and unit-level models, the UE and ZI, is the level at which the predictions are made:

   (a) *Unit-level:* Unit-level estimators use correlation between population-level information and the individual measurements. In the forestry setting, this means that the model predicts the response

variable, such as basal area or stored carbon, at the level of individual plot measurements before aggregating to the area-level.

(b) *Area-level:* At the area-level, the model is fit and predictions are made at the area level. Area-level models are faster to run computationally and don't require knowledge of individual observations.

When comparing the two classes, direct estimators have the advantage of being more easily interpreted and having low computational cost, however indirect estimators can often lead to a reduction in variance as a result of making use more. However, the reduction in variance relies of the quality of the model fit and a poorly specified model can induce bias into the estimator. Analysis by Goerndt *et al.* [24] compared several different direct and parametric-indirect estimators in the forestry setting, including PS, AE, UE, imputation, and multiple linear regression, finding that the UE model performed the best. However, when the model is misspecified, the UE estimator can be severely biased. For data where the unit-level relationships are noisy and complex, the area-level relationships, built on sample means, tend to be less so, signaling that the AE might be the optimal estimator. Apriori, we expect that in cases where there is sufficient zero-inflation, the ZI model should outperform the UE because it specifically incorporates prediction of these zeros and the AE because it utilizes unit-level observations as opposed to area-level means. Furthermore, while a limited literature exists to assess zero-inflation in inventory estimation problems, zero-inflation models are being used to improve estimates of forest responses to wildfire [18, 29] and climate change [30]. The drawback of using these models is both their statistical and computational complexity. Goerndt *et al.* [24] lends confidence that our study, which uses PS, AE, and UE, compares both models that are currently in use and parametric models that are particularly strong in the forestry setting.

1.3.3 Machine Learning Estimation Frameworks for SAE

Indirect, non-parametric, ensemble estimators including Random Forests (RF) and SAE Mixed-Effects Random Forests (SMERF) among others, relax the requirement of a parametric relationship between the response and predictor variables. This flexibility allows models to more easily capture complex relationships between auxiliary variables and the response than parametric models. Furthermore, ensemble methods, which combine many individual models, are typically more robust and have reduced variability [31]. While using

machine learning (ML) models have only been applied to the forest estimation recently [28, 32], ML models have been used to solve estimation problems for decades. One model that is particularly well suited for the SAE forest estimation problem is the SAE Mixed Effects Random Forest (SMERF) model. The model combines both the flexibility of a random forest to capture features like zero-inflation, with a small-area effect. We posit that this combination will improve estimation in comparison to the current suite of estimation techniques used to model US forests.

1.3.4 Random Forest Literature History

While non-parametric ensemble methods are not new, they have only recently begun to be applied to the forestry setting. Decision and regression trees are popular non-parametric models for categorical and quantitative, respectively, response variables. Trees are structured using the classical definition of a tree in graph theory and each node can have either zero or two children. Random forests are ensembles of decision or regression trees. Random forests greatly decrease variance at the cost of a slight increase in bias [31, 33]. While decision trees have been around for centuries, today's random forest algorithm was first proposed in Gordon *et al.* [33]. In the forestry setting, Freeman *et al.* [28] have shown that random forests remain one of the most effective predictive models for forest attribute estimation, performing better other state-of-the-art models, such as stochastic gradient boosting. This motivated our choice to adapt RFs to forestry estimation in the small area estimation setting. Random forests for clustered data was first addressed by [34] and the small area setting was explored in several papers led by Krennmair which produced the SMERF algorithm [35–37]. SMERF fits a single random forest to the entire dataset while allowing for differences across small areas not accounted for by the predictors. The algorithm, reminiscent of the expectation-maximization algorithm [38], alternates between fitting a random forest on an adjusted dataset to calculate the random effects and regressing on the model's residuals to calculate the small area fixed effects. This method has already been shown to be effective in the forestry setting in a estimation study of tree diameters across Caribbean islands [27].

1.3.5 Literature Conclusions

We find that the forest attribute estimation domain still has a large gap between current practice and state-of-the-art statistical inference models. We hope to close this gap by demonstrate that the added model structure of the ZI, RF, an SMERF models can improve forest inventory estimation in the small area estimation setting. We will focus our simulation on the FIA's dataset of forest measurements in Georgia, Idaho, Iowa, and Oregon. These states' large geographic and ecological differences will allow us to explore how these three models perform across varied terrain when compared with traditional direct and indirect parametric models. As a comparative study, we do not expect the analysis to provide the most accurate estimates of carbon, however we hope the simulation can demonstrate when each of the models might have an advantage over classical methods.

2 Methods

All of the data analysis, model fitting, and production code was written in the `R` Programming Language [39]. We used a myriad of packages to implement these results, in particular Microsoft's `doparallel` [40] and Dowle & Srinivasan [41] were instrumental in implementing efficient, parallelized model fitting. The simulations are run with model-specific packages [42–46].

2.1 Data

For the SAE setting, our data consist of a finite population U with N observations. The elements of U are assigned to one of D domains, $i = 1, 2, \ldots, D$ which correspond to the small areas. Typically $D \ll N$ as there are many more observations than small areas; in our simulation, $D = 30$ and $N = 31,073$. Each element of U consists of "pixel-level" data: a pair $(\vec{x}_{ij}, y_{ij})$ encoding the predictor variables $\vec{x}_{ij}$ and response variable y_{ij} of the jth pixel in ith domain. In our simulation, $\vec{x}_{ij} = \{x_{ij}^1, \cdots x_{ij}^p\}^T$ consists of $p = 8$ remotely-sensed variables, which are as follows:

1. `tcc16:` The percent tree canopy cover from the 2016 Nataional Land Cover Dataset (NLCD) [47]. As described later in Section 3.2.4 this variable has the highest correlation with our response variable,

carbon, and so we fit an additional set of simulations without this predictor.

2. `elev`: Elevation (meters) from the LANDFIRE Digital Elevation Model (DEM) Grid [48].

3. `tri`: Terrain Ruggedness Index (index) from the LANDFIRE DEM [48].

4. `def`: Mean annual climatic water deficit (millimeters) from the TOPOFIRE dataset. These means are over 1981-2010 (30 year normal) [49].

5. `ppt`: Mean annual precipitation (100 micrometers) from the PRISM 1981-2010, 30 year normals dataset [50].

6. `tmean`: Mean annual temperature (hundredths of a degree Celsius) from the PRISM 1981-2010, 30 year normals dataset [50].

7. `tmin01`: January mean temperature (hundredths of a degree Celsius) from the PRISM 1981-2010, 30 year normals dataset [50].

8. `tnt`: Boolean variable for the remotely-sensed presence of trees from the LANDFIRE dataset [48].

We denote the number of observations in each domain i be N_i. The predictive variables $\vec{x}_{ij}$ is the set of auxiliary data collected via remote satellite; the response variable y_{ij} is the amount of above-ground carbon in metric tonnes per hectare (carbon). Carbon is only available for the FIA-sampled groundplots for which there are $31,073$ observations. We define the population means as $\mu_i = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}$ and the auxiliary variable population mean for each section as the component-wise average of the $\vec{x}_{ij}$'s, $\vec{\mu}_{x,i} = \frac{1}{N_i}\sum_{j=1}^{N_i} \vec{x}_{ij}$ across each of the small areas.

In SAE the goal is always to estimate μ_i when all of the auxiliary measurements are known (for some models $\vec{\mu}_{x,i}$ suffices) but only a small number of y_{ij} measurements are known. The few y_{ij} measurements that are observed for a domain are called the sample. We denote the response variable and auxiliary variable vector with an asterisk, y^*_{ij} and $\vec{x}^*_{ij}$, respectively. While we have access to all $\vec{x}_{ij}$ (or at least their mean, $\vec{\mu}_{x,i}$), only the sample y^*_{ij} can be used in the model fitting stage. All $\vec{x}_{ij}$ observations are available to generate simulation estimates of μ_i.

For our simulation, we used FIA data from 4 states: Georgia, Idaho, Iowa, and Oregon. They represent a diverse set of ecological regions, locations, and climates. This 4-state dataset contains plot-level observations for $31,073$ sites collected with FIA's quasi-systematic sampling technique. Across the 4 states there are $D = 30$ ecosections each with more than 200 observations, which we consider a minimum acceptable population size for which to perform our estimation. Therefore, our overall goal is to compare the performance of our estimators in estimating μ_i, the density of above-ground carbon, for $i = 1, \dots 30$, across all 30 ecosections in these 4 states. To assess our estimators we generate $K = 2000$ samples, $s_1^n, s_2^n, \cdots, s_k^n, \cdots, s_K^n$ for each of size $n = 4, 8, 16$ and 32. Each sample gives us n observations of the response variable that we do observe and in each simulation this sample, $(\vec{x}_{ij}^*,\ y_{ij}^*)$, for each ecosection $i = 1, 2, \dots, 30$ will be denoted with the asterisk. The sample means are denoted $\bar{x}_i^*$ and $\bar{y}_i^*$ for area i. We fit each model on the sampled pairs for all ecosections and use the full auxiliary dataset to produce our estimate $\hat{\mu}_i$ for each ecosection. Note that we have access to the response variable for all $N = 31,073$ observations, which allows us to calculate the true population parameter, μ_i, but in each simulation repetition we assume all but n are unobserved.

2.2 Current Suite of Estimators

2.2.1 Post-Stratified (PS)

Recall that the post stratified estimator is a direct estimator, meaning that $\hat{\mu}_i$ is calculated using only y_{ij}^* and $\vec{x_{ij}}$ in section i. The Post-Stratified (PS) estimator is calculated using a weighted average of the post-stratified means. The means to be weighted are taken across the y_{ij}^* in each strata of a categorical variable. We let there be H strata, indexed $h = 1, 2, \cdots, H$ and the number of auxiliary observations in each strata be N^h. In our case, the categorical variable is `tnt`, for which $H = 2$ (plot is either recorded as having trees or not), and the number of auxiliary observations in each of the two strata, N^1 and N^2 sum to the total number of auxiliary measurements in section i: $N^1 + N^2 = N_i$. Then the PS estimator is given by:

$$\begin{aligned}\hat{\mu}_i &= \sum_{h=1}^{H} \frac{N^h}{N} \left[\frac{1}{n} \sum_{j=1}^{n} y_{ij}^* \right] \\ &= \sum_{h=1}^{H} \frac{N^h}{N} \bar{y}_i^*\end{aligned}$$

Thus the PS estimator as a weighted mean of means and is a statistic of the data only in section i. For the forestry application, average carbon is first calculated across the samples in each of the forest and non-forest regions and then these measurements are combined proportional to the total number of tree vs non tree observations. PS is an asymptotically unbiased and consistent estimator for the section mean μ_i. PS is implemented in the R package `FIESTA` [26].

2.2.2 Unit-Level EBLUP (UE)

The Unit-level EBLUP (UE) fits pixel-level observations to a linear regression model with random effects to account for differences between sections and fixed effects to capture the relationship between the auxiliary variables and carbon. We denote the section-level effects as v_i and $\boldsymbol{\beta}$, respectively. The relationship between each entry $(\vec{x}_{ij}, y_{ij}) \in U$ is given by:

$$y_{ij} = \vec{x}_{ij}^T \boldsymbol{\beta} + v_i + \varepsilon_{ij}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of the coefficients associated with the fixed effects, v_i, is the random effect associated with area i, and ε_{ij} is the individual random effect for observation j in area i. Here we assume that both random effects are distributed normally:

$$v_i \sim \mathcal{N}(0, \sigma_v^2) \qquad \text{and} \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_e^2)$$

In this case σ_v^2 is the between-area variance parameter and σ_e^2 is the within-area variance parameter. These parameters are obtained using either method of moments or restricted maximum likelihood (REML); we use REML. From this, $\hat{\boldsymbol{\beta}}$ and $\hat{v}_i$ are estimated as laid out in Rao [51]. Once the model has been fit we use the model to predict the small area estimates $\hat{\mu}_i$. We do so as follows:

$$\hat{\mu}_i = \frac{1}{N_i} \sum_{i=1}^{N_i} \left[\vec{x}_{ij}^T \hat{\boldsymbol{\beta}} + \hat{v}_i \right]$$

We implemented the UE model using the `hbsae` package written by Boonstra [43].

2.2.3 Area-Level EBLUP (AE)

Similar to the UE, the area-EBLUP (AE) assumes a linear relationship between the auxiliary and response variables, but at the area-level. While the AE only uses means of the auxiliary population to predict over, it still requires sample pairs $(\vec{x}^*_{ij}, y^*_{ij})$ to estimate the variance of the small area and individual random errors. The AE is fit using $\bar{x}^*_i$ and $\bar{y}^*_i$ for all 30 ecosections. We again use v_i and ε_i to distinguish the small-area effects for each subsection and individual random errors. Our equation becomes:

$$\bar{y}^*_i = \bar{x}^{*^T}_i \boldsymbol{\beta} + v_i + \varepsilon_i$$

Note that the random errors are now errors of a sample mean as opposed to a singular sample value. We again assume the errors are distributed normally as follows:

$$v_i \sim \mathcal{N}(0, \sigma^2_v) \qquad \text{and} \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2_e)$$

The parameters σ^2_v and σ^2_e are again estimated using either method of moments or REML, and $\hat{\boldsymbol{\beta}}$ and $\hat{v}_i$ are estimated as in Rao (2015) [51]. Note that since this is an *area-level* estimator, the prediction is made using the mean of the auxiliary data, $\vec{\mu}_{x,i}$. Our prediction for the small area mean is then:

$$\hat{\mu}_i = \vec{\mu}_{x,i}\hat{\boldsymbol{\beta}} + \hat{v}_i$$

2.3 The Unit-Level Zero-Inflation Estimator (ZI)

The ZI model builds on UE, extending the estimator to account for a high number of zeros in the data. In the forestry setting, our carbon estimates have non-negative observations which are frequently zero, a setting explored by Pfeffermann *et al.* [52]. To extend the linear framework of the UE described in 2.2.2 to address zero-inflation we first consider the variables that an SAE model relies on. For this model, we assume that the covariates between each of the measurements are fully independent once the small area effects are known

and so the covariate matrix is the identity. For a given model we can think of the prediction as:

$$\hat{y}_{ij} = \mathbb{E}[y_{ij}|\vec{x}_{ij}, v_i] \tag{1}$$

We see that our prediction of y_{ij} depends on the auxiliary measurements at that location, $\vec{x}_{ij}$ and the corresponding small area mean, v_i. By adding extra conditioning on whether or not the observation itself is zero, we can expand 1 as follows:

$$\begin{aligned}\mathbb{E}[y_{ij}|\vec{x}_{ij}, v_i] &= \underbrace{\mathbb{E}[y_{ij}|\vec{x}_{ij}, v_i, y_{ij} = 0]}_{\text{evaluates to zero}} \mathbb{P}(y_{ij} = 0|\vec{x}_{ij}, v_i) + \mathbb{E}[y_{ij}|\vec{x}_{ij}, v_i, y > 0]\mathbb{P}(y_{ij} > 0|\vec{x}_{ij}, v_i) \\ &= \mathbb{E}[y_{ij}|\vec{x}_{ij}, v_i, y_{ij} > 0]\mathbb{P}(y_{ij} > 0|\vec{x}_{ij}, v_i) \end{aligned} \tag{2}$$

Out of equation 2 comes a wonderful intuition for the structure of a zero-inflation model. The conditional expectation yields an estimator in which one model is fit to the nonzero data which is then weighted by the probability that that point is not-zero which can be modeled separately. With the source of the model mis-specificity in the zero-inflation removed from the expectation, we can now fit a linear model to $\mathbb{E}[y_{ij} \ |\vec{x}_{ij}, v_i, y_{ij} > 0]$ without violating the assumptions of random normal error. To model second term, $\mathbb{P}(y_{ij} > 0 \ |\vec{x}_{ij}, v_i)$, we choose a logistic function.

2.3.1 Linear Mixed Model (LMM)

We model $\mathbb{E}[y_{ij}|\vec{x}_{ij}, v_i, y_{ij} > 0]$ as a linear mixed model with random intercepts fit to the nonzero portion of the sample data. The subscript nz specifies that we fit this model on all of the original observations that are non-zero. We superscript our response with a star ($\star$) to denote the prediction is from our linear model and is not our final estimate for y_{ij}.

$$y^{\star}_{ij,nz} = \mathbf{x}^T_{ij,nz}\boldsymbol{\beta}_{nz} + v_{i,nz} + \varepsilon_{ij,nz} \qquad \text{where} \qquad v_{i,nz} \sim \mathcal{N}(0, \sigma^2_{v,nz}) \ , \ \varepsilon_{ij,nz} \sim \mathcal{N}(0, \sigma^2_{e,nz}) \tag{3}$$

Again $\vec{x}_{ij,nz} = (x^1_{ij,nz}, ..., x^p_{ij,nz})^T$ is a $p \times 1$ vector of covariates, however corresponding only to variables which are nonzero while $\boldsymbol{\beta}_{nz}$ is a $p \times 1$ vector of fixed effects. Furthermore $\sigma^2_{v,nz}$ is the between area variance

parameter and $\sigma^2_{e,nz}$ is the within area variance parameter. Note the similar construction to the UE model.

2.3.2 Logistic Mixed Model (GLM)

We model $\mathbb{P}(y_{ij} > 0 \mid R = r_{ij})$ as a logistic mixed model. We distinguish the small area effect in the GLM from that in the LMM with a star as they are different effects and is an important distinction for when we combine the two models in 2.3.3.

$$p_{ij} = \frac{\exp\left(\vec{x}_{ij}^T\boldsymbol{\gamma} + v_i^\star\right)}{1 + \exp\left(\vec{x}_{ij}^T\boldsymbol{\gamma} + v_i^\star\right)} \qquad \text{where} \qquad v_i^\star \sim \mathcal{N}(0, \sigma^2_{v^\star}) \tag{4}$$

where again, $\vec{x}_{ij}$ is a $p \times 1$ vector of covariates, $\boldsymbol{\gamma}$ is a $p \times 1$ vector of fixed effects, and $\sigma^2_{v^\star}$ is the value of the between area variance parameter. We don't include the subscript nz here as the logistic model is built on the entire sample data set.

2.3.3 Combining the Models

We then combine these two estimators to create estimates at the pixel-level:

$$y_{ij} = y^\star_{ij,nz} p_{ij} = \left[\vec{x}^T_{ij,nz}\boldsymbol{\beta}_{nz} + v_{i,nz} + \varepsilon_{ij,nz}\right] \cdot \left[\frac{\exp\left(\vec{x}_{ij}^T\boldsymbol{\gamma} + v_i^\star\right)}{1 + \exp\left(\vec{x}_{ij}^T\boldsymbol{\gamma} + v_i^\star\right)}\right] \tag{5}$$

For simplicity we assume that the random effects between the two parts of the model are uncorrelated:

$$\text{Corr}\left(v_{i,nz}, v_i^\star\right) = 0 \qquad \forall j \in 1, ...J$$

In general this assumption does not hold, however Pfeffermann *et al.* [52] showed that in a similar setting, including the correlations improved the accuracy of their estimates only marginally, while introducing further complexity to the model itself. We fit the LMM an GLM models in Equations 3 and 4 to get estimates for $\hat{\boldsymbol{\beta}}_{nz}, \hat{\boldsymbol{\gamma}}, \hat{v}_{i,nz}$, and $\hat{v}_i^\star$. Using the R package `lmer` [44] these parameters are estimated using restricted maximum likelihood (REML). The parameters are then used in Equation 5.

With these equations we produce our estimate $\hat{\mu}_i$ by first estimating $\hat{y}_{ij}$ for each observation $\vec{x}_{ij}$ in the auxiliary dataset:

$$\hat{y}^{\star}_{ij,nz} = \vec{x}^{T}_{ij}\hat{\boldsymbol{\beta}}_{nz} + \hat{v}_{i,nz}$$
$$\hat{p}_{ij} = \frac{\exp\left(\vec{x}^{T}_{ij}\hat{\boldsymbol{\gamma}} + \hat{v}^{\star}_{i}\right)}{1 + \exp\left(\vec{x}^{T}_{ij}\hat{\boldsymbol{\gamma}} + \hat{v}^{\star}_{i}\right)}$$

While the two part model is fit on the plot-level sample dataset, it is applied to the pixel-level data set. An estimate for our final model, at the individual plot level is taken to be the product of the two estimates as discussed in section 2.3:

$$\hat{y}_{ij} = \hat{y}^{\star}_{ij,nz}\hat{p}_{ij}$$

$\hat{\mu}_i$ is calculated by aggregating to the domain level:

$$\hat{\mu}_i = \frac{1}{N_i}\sum_{i=1}^{N_i}\hat{y}^{\star}_{ij,nz}\hat{p}_{ij} \tag{6}$$

Equation 6 gives the full form of the zero-inflation model estimator. Notice that since we are summing over the total number of pixel-level data points in area i, we must predict both of our models in our two part estimator on the entire pixel-level data set. This is more computationally expensive than fitting the UE and AE models.

2.4 Random Forest Models

Random Forests (RFs), first introduced by Breiman [31] utilize an ensemble of decision trees to make predictions. We focus on RFs because they are adept at capturing non-linear relationships between predictors, handle high-dimensional data well, and do so with minimal tuning. To give intuition for the RF models, we build intuition by describing regression trees as their building blocks, exploring how to combine these trees into a RF model, and then extending RFs to the small-area estimation setting.

2.4.1 Regression Trees

Regression trees are a supervised learning technique which builds predictions by clustering data. As RFs don't deal with small areas, we introduce slightly simplified notation that ignores the small area labels. Assume we have training data $(\boldsymbol{X}, \vec{y})$, in $\mathbb{R}^{n\times p}$ and $\mathbb{R}^{n\times 1}$, respectively, which are comprised of individual observations $\{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \cdots (\vec{x}_n, y_n)\}$. The regression tree on $\boldsymbol{X}$ is a partitioning of the hyperplane in $\mathbb{R}^p$ that gives a rule for predicting any new $\vec{x}_i \in \mathbb{R}^p$ based on the conditions found at each node in the tree. The tree is comprised of a set of nodes $\mathcal{P}$ which have either zero or two children nodes. If the node has no child nodes, we call it a leaf of the tree. Together the leaves of the regression tree form the partition of $\mathbb{R}^p$. Each node, $n_k \in \mathcal{P}$ is associated with a region $s_k \in \mathbb{R}^p$, and if the node is a parent node, it's children partition s_k into two regions based on the splitting criteria.

The tree is parameterized by its structure, the termination criteria, and the splitting criteria. We use the CART algorithm as described by Breiman [53] to determine splits until the termination criteria is reached. The termination criteria we chose for our simulation allows the trees to grow to a maximum size of ten nodes; another common criteria limits the depth of the tree. Until the termination criteria is reached, the regression tree continues to grow based on the splitting criteria: At each node, n_k, the space, s_k, is partitioned into s_{k+1} and s_{k+2}. The partition is done using the CART algorithm which splits on a single variable exhaustively checking all possible splits to find the one that minimizes the sum of squared errors (SSE) across the two groups of the response variable. SSE is calculated as:

$$SSE = \sum_{i:\vec{x_i}\in s_{k+1}} \left(y_i - \bar{y}_{s_{k+1}}\right)^2 + \sum_{i:\vec{x_i}\in s_{k+2}} \left(y_i - \bar{y}_{s_{k+2}}\right)^2 \tag{7}$$

where $\bar{y}_{s_{k+1}}$ denotes the mean y-value of observations in s_{k+1} and $\bar{y}_{s_{k+2}}$ is the mean y-value in s_{k+2}. The best split is chosen by brute force by evaluating all possible splits on the predictor variables available and choosing the split which minimizes SSE in 7. In the case of building a RF, to ensure ensemble trees form with sufficient variance, each split typically is only chosen across a random fraction of the available predictor variables. We utilize the default split of $\lceil M/3 \rceil$ implemented in the R package `randomForest` [45].

After the stopping criteria is reached, the decision tree assigns a value to each leaf node by averaging the y

values of the points in the region. Formally, for leaf n_k, the estimate of any new $\vec{x}_i \in s_k$, which we denote $\hat{y_k}$, is:

$$\hat{y_k} = \frac{\sum_{i=1}^{n} y_i \mathbb{I}(\vec{x_i} \in s_k)}{\sum_{i=1}^{n} \mathbb{I}(\vec{x_i} \in s_k)}$$

Figure 1 gives an example decision tree that uses FIA auxiliary variables tree canopy cover (`tcc16`) and enhanced vegetation index (evi) to predict basal area, or the average number of square feet per acre covered in tree stems. In addition to demonstrating the regression tree structure and splits, this tree shows that the decision tree is able to quickly separate a significant number of zeros from the dataset in node 4 as the data in that split have an average basal area of 2. As basal area is strictly positive, we conclude that most of these observations are small or zero. Notably, if the tree is allowed to pick the best split based on SSE across all of the predictors, then for any given dataset, the tree will be purely deterministic.

2.4.2 Bootstrap Aggregation towards the Random Forest

To avoid the problem of creating identical trees, randomness is introduced between trees using bootstrapping and random variable selection for splits. First, each regression tree is fit on a bootstrapped sample; a technique which resamples the data with replacement, often to estimate variance. Thus no tree observes the same dataset. Second, for each split the regression tree is restricted to choosing splits on only a subset of predictors. More intuition is described in Breiman [31]. In combination, choosing a subset of the candidate splits at each nodes and bootstrapping significantly reduces the correlation. To produce an ensemble estimate, we perform bootstrap aggregation, or bagging, which combines the estimates for an x from each tree; this bootstrap aggregation of many regression trees is a random forest. The random forest makes predictions by taking an unweighted average of decision tree predictions. To prediction a new $\vec{x}_i$ using the forest of regression trees, $F = \{f_1, f_2, \cdots, f_B\}$, we take:

$$\hat{y}_i = \frac{1}{B} \sum_{k=1}^{B} f_k(\vec{x}_i) \tag{8}$$

Equation 8 gives the prediction for an individual pixel. In order to compute the small area means in our simulation, we first must reintroduce the concept of small areas to RFs, which we do below in Section 2.5.1.

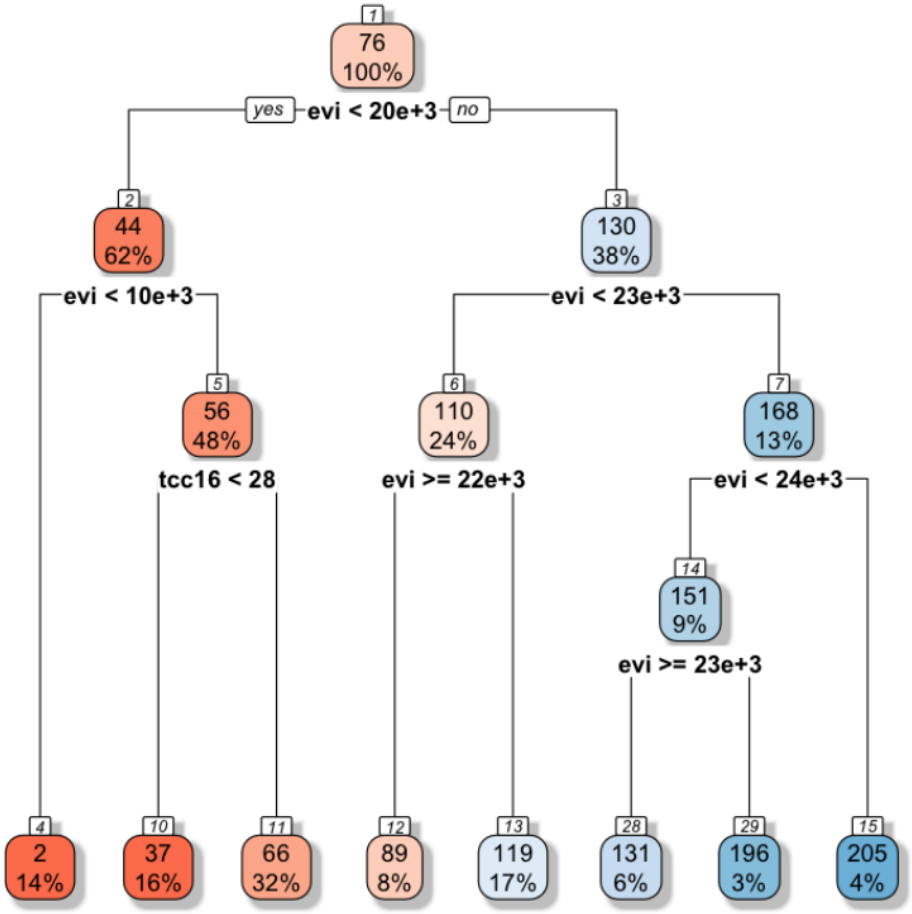


Figure 1: An example regression tree used to predict basal area from tree canopy cover (`tcc16`) and enhanced vegetation index (evi). Note the tree structure contains a single root node (labeled 1) which has an average basal area of 76 and contains 100% of the data. The splitting criteria $evi < 20e+3$ was chosen based on the SSE across all possible splits. It evaluates to a Boolean for each observation and sample observations are split into nodes 2 and 3, where the process repeats. Eventually the leaf nodes are reached which are used to generate the final prediction.

2.5 Random Forests models in the Small Area Estimation Setting

2.5.1 The Random Forest (RF)

To evaluate a RF we fit a random forest regressor on the sample. For our simulation, we fit each sample with a RF comprised of 500 trees, splitting at each node was done on the the default number of parameters, $\lceil p/3 \rceil$, and trees were allowed to grow to a maximum of 10 nodes. Unlike the PS, AE, UE, and ZI models, the RF is not deterministic. Returning to our previously defined SAE notation, our estimates for the small area means are as follows:

$$\hat{\mu}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \hat{y}_{ij} = \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{1}{B} \sum_{k=1}^{B} f_k(\vec{x}_{ij})$$

where we replace $\hat{y}_i$, as defined in equation 8, with the doubly indexed $\hat{y}_{ij}$. We see that the RF model is small-area agnostic, meaning that it assumes all of the variability in the small area means can be accounted for by differences in the auxiliary data. As such, if the small area effects are in fact significant, we would expect the RF model to do quite poorly. To address this problem we turn to an extension of the RF which includes small area effects.

2.5.2 Mixed Effects Random Forests for Small Area Estimation (SMERF)

The SMERF model combines the strength and flexibility of a random forest to capture the relationship between the auxiliary data and the response variable, with a framework that incorporates small area effects. To implement the SMERF, we follow Krennmair *et al.* [36] and Krennmair [35] and use the R package `MixRF` to run our simulations [46]. Thus far, our analysis has focused on linear models, however, small area estimation more broadly, includes just the assumption that there is an effect from the small areas that is not accounted for by the predictors. For models like the AE and UE, the fixed effects are included as follows:

$$y_{ij} = \vec{x}_{ij}^T \boldsymbol{\beta} + v_i + \varepsilon_{ij}$$

Recall that we assume normality of the distribution of domain-specific random intercepts, $v_i \sim \mathcal{N}(0, \sigma_v^2)$, and the individual random errors, $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. However, linear assumptions are restrictive. In the forestry setting, perhaps a specific combination of temperature, precipitation, and sunlight benefits one tree species significantly more than others. Furthermore, carbon values are often severely right skewed; some regions have high tree density but none have negative observations, and so a normality assumption may not be most representative. We can generalize the above linear term, $\vec{x}_{ij}^T\boldsymbol{\beta}$, to a general $f(\vec{x}_{ij})$:

$$y_{ij} = f(\vec{x}_{ij}) + v_i + \varepsilon_{ij}$$

Letting f be a random forest, we have now defined the SMERF model [35]! With this framework in hand, we have estimates of $\hat{f}$, fit using the random forest algorithm [45], and the random forest variance estimates $\hat{\sigma}_v^2$ and $\hat{\sigma}_\varepsilon^2$, which are calculated using REML in Step 6 of the SMERF algorithm. To calculate our SMERF predictions we predict over the auxiliary data for a given sample s_k^n using:

$$\hat{y}_{ij} = \hat{f}(\vec{x}_{ij}) + \hat{v}_i = \hat{f}(\vec{x}_{ij}) + \left(\frac{\hat{\sigma_v^2}}{\hat{\sigma_v^2} + \hat{\sigma_\varepsilon^2}/\pi_{k,i}^n}\right)\left(\frac{1}{\pi_{k,i}^n}\sum_{j \in s_{k,i}^n}\left(y_{ij} - \hat{f}^{OOB}(\vec{x}_{ij})\right)\right)$$

as described in Krennmair [35]. The first term, $\hat{f}(x_{ij})$, is subsection agnostic and is a prediction from the random forest, while the second term, is a re-weighted measure of the average error of the model for sampled observations in subsection i. Here, $\hat{f}^{OOB}$ predicts $\vec{x}_{ij}$ using out-of-bag (OOB) predictions, meaning only the decision trees of $\hat{f}$ which did not train on $(\vec{x}_{ij}, y_{ij})$ are used; this is standard practice for evaluating RF models [31]. We note that as the subsection sample size, $\pi_{k,i}^n$, grows, the coefficient on the second term will approach 1. For our simulation, $\pi_{k,i}^n = n$ because we fix the sample size to $n = 4, 8, 16, 32$. For smaller sample sizes we down-weight the maximum likelihood estimate of the subsection mean provided by the sample. This is a form of shrinkage and is reminiscent of the James-Stein estimator. It makes sense to apply shrinkage in the situation because assuming there was no difference between subsections, we would still expect a difference in the sample predictions due to randomness. The difference is particularly pronounced for small sample sizes. To derive estimates for our subsection means, we take a mean over the pixel-level

predictions of auxiliary dataset:

$$\begin{aligned}\hat{\mu}_i &= \frac{1}{N_i}\sum_{j=1}^{N_i}\hat{y}_{ij} \\ &= \frac{1}{N_i}\sum_{j\in U_i}\hat{f}(\vec{x}_{ij}) + \left(\frac{\hat{\sigma_v^2}}{\hat{\sigma_v^2}+\hat{\sigma_\varepsilon^2}/\pi_{k,i}^n}\right)\left(\frac{1}{\pi_{k,i}^n}\sum_{j\in s_{k,i}^n}\left(y_{ij}-\hat{f}^{OOB}(\vec{x}_{ij})\right)\right)\end{aligned}$$

2.5.3 The SMERF Algorithm

Fitting a SMERF model to a sample, s, is done using an iterative approach that alternates between fitting a random forest on adjusted y_{ij}^{adj} and estimating the mixed effects from the residuals of the random forest model. Any iterative algorithm requires a stopping criteria and for SMERF the condition is defined when either the small areas are deemed sufficiently converged or a maximum number of iterations is reached. We define the stopping criteria as reached as when either a maximum number of iterations is exceeded, $\tau > 0$, or the algorithm has converged sufficiently to below some tolerance, $\lambda > 0$, as measured by the difference in log likelihood of a linear regression on the residuals between subsequent iterations. The algorithm is as follows:

1. Choose a maximum number of iterations, τ, and a likelihood error tolerance, λ. Initialize the number of iterations $t = 0$, and log likelihood error $\ell^0 = \infty$.

2. (Optional) Choose initial random effects, $\vec{V}^0 = (v_1^0, \cdots, v_D^0)$ based on prior knowledge of the problem to improve convergence speed. Without prior knowledge or assuming the effects are zero, set $\vec{V}^0 = \vec{0}$.

3. Update the previous $y_{ij}^{adj,t-1}$ in s based on the random effects from the previous iteration or from step 2 if this is the first iteration:

$$y_{ij}^{adj,t} \leftarrow y_{ij}^{adj,t-1} - v_i^t$$

4. Fit a random forest model, $\hat{f}^t$, on the sample $\vec{x}_{ij}^*$ and adjusted $y_{ij}^{adj,t}$ values. Specify the number of trees and number of nodes to split over at each step.

5. Compute the new residuals, $\vec{r}^t$, where each component is calculated using the difference of the predic-

tion:

$$r_{ij}^t \leftarrow y_{ij} - \hat{f}^t(\vec{x}_{ij})$$

6. Fit a linear fixed effects model, $\mathcal{M}^t$, on the residuals with an intercept and fixed effects across the domains. We use the `lmer` package in `R` [44] to fit $\mathcal{M}^t$ using the formula `'residuals ~ -1 + (1| domain)'`.

7. Update the random effects, $\vec{V}^t$, to the fixed effects from $\mathcal{M}^t$ for all small areas.

8. Calculate the new data log likelihood from the random effects model:

$$\ell^t(\vec{r}^t|\mathcal{M}^t) = \sum_{r_{ij} \in \vec{r}} \log \mathbb{P}(r_{ij}^t|\mathcal{M}^t)$$

9. Evaluate the stopping criteria. If $t \geq \tau$ or $|\ell^t - \ell^{t-1}| < \lambda$ stop the algorithm, return the SMERF model, and continue. Otherwise update t: $t \leftarrow t + 1$ and return to Step 3.

Now that we have fit the SMERF algorithm, we can evaluate it's performance and compute the small area means as described in 2.6.

2.6 Model Evaluation

When evaluating model performance, we focus on two factors: bias and root mean squared error (RMSE). Bias indicates whether the estimator tends to over- or under-predict on average; formally it is the difference between the estimators expected value and μ_i, which is defined in Section 2. RMSE gives an indication of how far away, on average, a model is from the true population parameter μ_i. Also, recall from Section 2 that the simulation study is conducted over $K = 2000$ samples for each sample size $n = 4, 8, 16, 32$. We fit each sample will all 6 models: PS, UE, AE, ZI, RF, and SMERF and then compute the bias and RMSE for a given model and sample size over each ecosection which we will use to compare model performance.

1. **Bias:** To calculate empirical bias we first define our test statistic for each ecosection i, $T = \frac{1}{K}\sum_{k=1}^{K} \hat{\mu}_i^k$ where $\hat{\mu}_i^k$ is the estimate of μ_i produced by fitting the model to sample s_k^n. The bias is written as

follows:

$$bias(T_i, \mu_i) = \mathbb{E}(T_i) - \mu_i = \frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_i - \mu_i$$

2. **RMSE:** The root mean squared error represents the average distance of an estimate from the true population parameter, μ_i and is computed as:

$$RMSE(\hat{\mu}_i) = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{\mu}_i^k - \mu_i\right)^2}$$

To assess model performance we compare bias and RMSE amongst the models within a sample size and section pairing. Comparisons between different sample sizes and sections doesn't tell us how models comparatively perform, although can sometimes provide insight into the data structure.

3 Results

3.1 Framing the Results Section

As the results of any simulation study are heavily influenced by the dataset used, it was essential that the chosen dataset be representative of the system we hoped to model; in this case, the forest system. We deemed choosing a dataset which accurately captured the relationships between the auxiliary variables and our response variable of interest of particular importance for this study so that the results would be generalizable. A major challenge in a simulation study is access to the true population parameters of interest as these are required to calculate the bias and mean squared error of simulation runs. In the forestry setting, pixel level data exists for the entire US at 30×30m resolution, however it does not include carbon, the parameter of interest, which is needed to calculate the population parameter across the small areas. We considered two solutions:

1. Impute carbon across all 30×30m pixels using an approach like k-nearest neighbors from the auxiliary data to the pixel level. This approach would allow us to utilize all auxiliary-level data and predict carbon at high-resolution. However this method is dependent on the imputation method.

2. Treat the samples for which y_{ij} is observed as the population and the sample is drawn from these observations. This is the strategy we ultimately chose.

The first approach benefits from a larger dataset, with pixel densities approximately 1000 higher than samples, and is more representative of forests because the pixels completely cover the study region. However as there is not population-level carbon information, we cannot compute a true μ_i against which to compute the models' bias and RMSE in order to assess the models as the response variable must be imputed. The second approach, using the samples as the population, does not have the drawback of needing to imputing the variable of interest, however has a much smaller population size which is spatially disconnected.

Initially we ran our simulation with the first approach, imputing basal area across section M333A at the pixel level. With over 3 million pixels, this method required the use of a research cluster to run the compute as over 50 billion pixel-level predictions were required, see Appendix A. However, as our imputation method used KNN, we were concerned that the binning method used by KNN was too similar to that used by the RF model, and could lead to a simulation study that favored both the RF and SMERF models. Other imputation methods, such as linear models would certainly benefit the UE, AE, and ZI models which share a common underlying linear assumption and structure. While we include the first set of results that imputed data in Appendix A, the rest of results section is focused on the second approach.

Broadly, we hope that the results will underscore where each of the 6 different models will perform best. For example, if there is strong linear relationships between the auxiliary and response variables, we would expect the linear models (AE, UE, and ZI) to perform comparatively well, while we expect more complex and non-linear relationships between the variables to preference the ML models. In particular we are curious to see the effect of section zero-inflation on relative model performance, hypothesizing that higher levels of zero-inflation will preference ZI, RF, and SMERF. Furthermore, by examining relationships between bias or RMSE and the underlying data we can uncover how sensitive the models are to regional effects.

3.2 Insights from the 4-State FIA dataset

In order to contextualize the simulation results, we perform an exploratory data analysis (EDA) of the 4-state FIA dataset which explores the distribution of predictors, variable correlations, and assesses the proportion

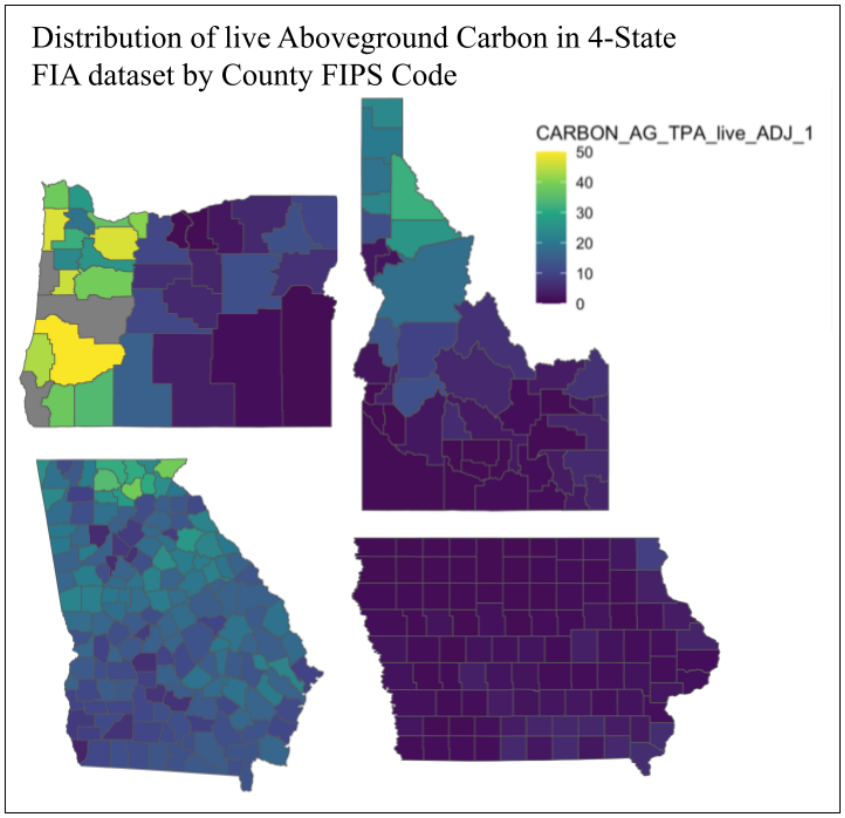
of zero-inflation. The EDA is essential to assess the conditions under which estimators perform best. For our specific simulation study we hope to show the estimators we are introducing to the SAE forestry setting, ZI, RF, and SMERF, perform well across a wide range of environments. In particular, we are interested in how these models perform across different levels of zero-inflation of the response variable. Furthermore, we are curious to see how SMERF compares to the RF, as based on the synthetic simulation study described in Section 3.3, we expect SMERF to outperform RF when large differences in the carbon are not correlated with any of the auxiliary predictors. Performing exploratory data analysis (EDA) is standard practice for any simulation study, as the analysis often reveals hidden variable relationships and distributions which can influence procedure and results.

3.2.1 Response Variable Distribution

We begin by examining the distribution of the response variable, carbon, across the 4 state dataset both at the county level and by section, shown in Figure 2. We see that there is a wide range of average county carbon which which exhibits strong geographic clustering. Georgia has consistent carbon across all counties while Iowa has very little to no carbon in trees across all counties. Oregon and Idaho are more variable, with some counties that have a moderate or large amount of carbon while others have very little. The large spread will allow us to evaluate how flexible the estimators are in their prediction across a broad set of climates and regions, and could help inform which estimator is most reliable in all situations. Notably, when aggregating to the section level, we find some sections with high carbon: M221D (GA), M242A (OR), M261G (OR), and M333D (ID). Several regions with almost no carbon include: 342B (OR), 342C (OR/ID), 342D (ID), and 342I (OR). Notably, despite being infamous for its scarce tree population, the sections in Iowa do not make the bottom 4.

3.2.2 Auxiliary Variable Distributions

We also examined the distribution of the auxiliary variables in Figure 3 which exhibit a wide range of environmental conditions. For example, average section elevation (`elev`) ranges from near 0 to over 2000 meters, while mean temperatures (`tmean`) ranges between 3 and 19°C. Furthermore, we see that certain sections have just over the cutoff of 200 observations, while a few have over 2000. With such a wealth of



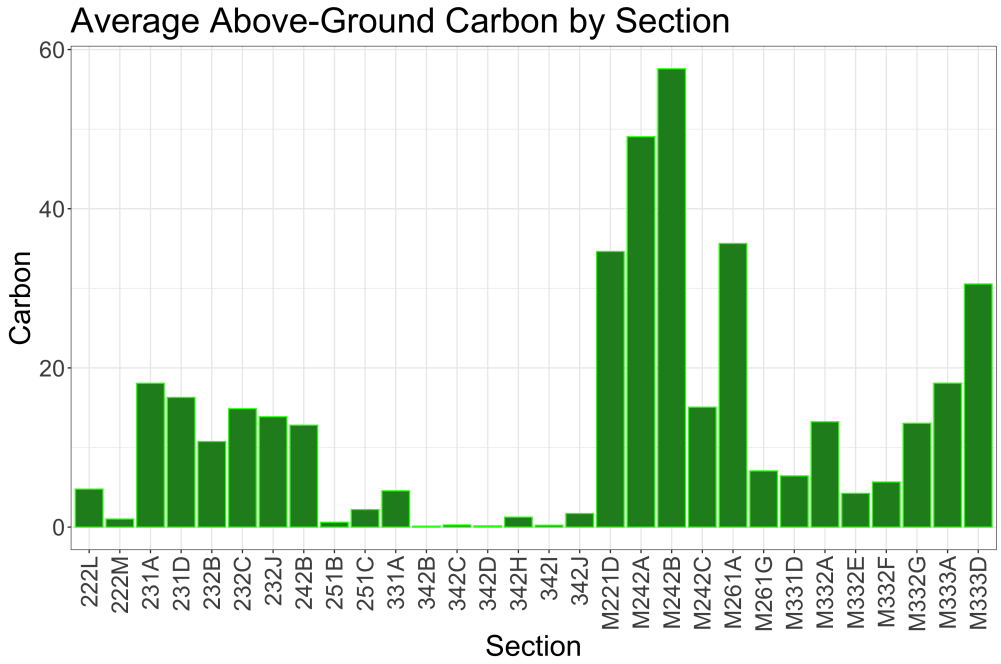


Figure 2: Top: distribution of above ground carbon (Metric tonnes per hectare) for FIA dataset used in the simulation. States shown are (clockwise from top left): Oregon, Idaho, Iowa, and Georgia. We limit our color scale from 0-50 metric tonnes per hectare as a few counties had significantly higher average carbon. Gray shaded counties had no observations in the FIA dataset. Bottom: Average above ground carbon by section.

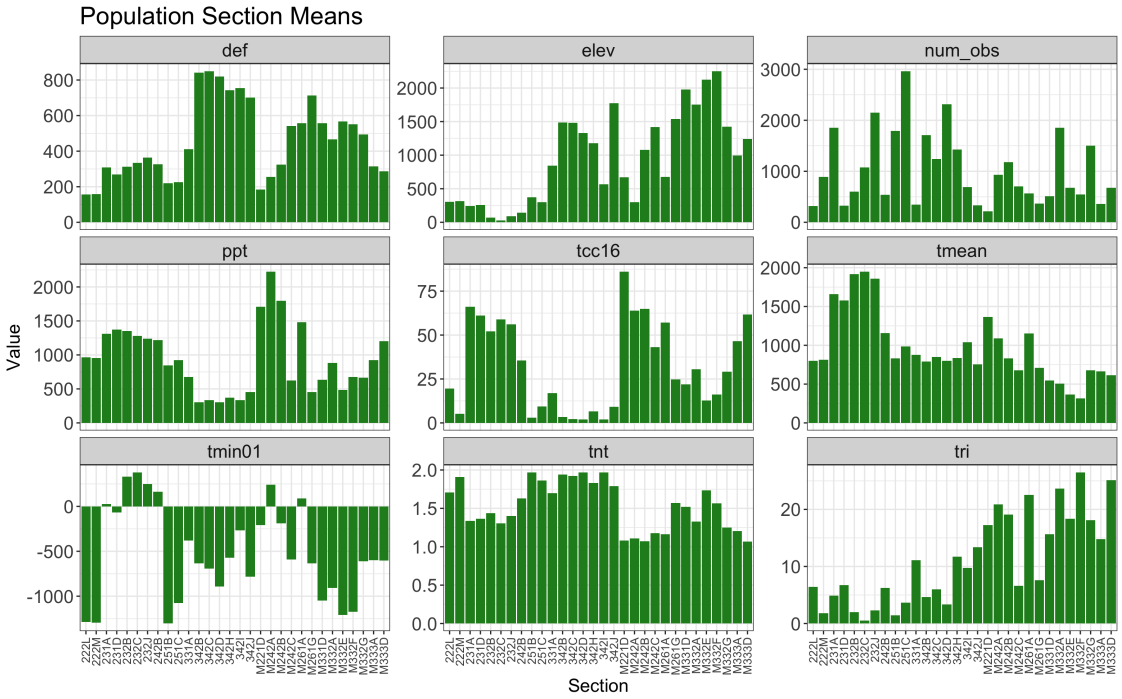


Figure 3: Faceted bar plot of the population section means for all 8 auxiliary variables used in the simulation as described in Section 2.1 the number of observations (`num_obs`) per section.

variables over which to analyze the results, we could have chosen to explore any number of orderings between the results. Ultimately, we chose to order the sections for the FIA results by average section carbon. The variety in terrain, sample size, and environmental factors underscore that models which do well across the majority of regions must be be flexible enough to encode broad relationships between the auxiliary data and response.

3.2.3 Zero-Inflation

As most of our estimators, with the exception of the AE, make predictions at the unit level, we are also particularly interested in how estimators perform against different levels of zero-inflation. Figure 4 reveals significant differences in the proportion of zero-inflation across the 4 states. For example, the majority of counties in Iowa have at least 80% zeros across all counties, while most counties in Georgia have less than 50% zero-inflation. Idaho and Oregon have a wide range of zero-inflation percentages, with observations across high ($> 90\%$) and low ($< 20\%$) zero-inflation proportions, perhaps because these states are relatively large and encompass more diverse climatology than either Iowa or Georgia. We expect the ZI and SMERF

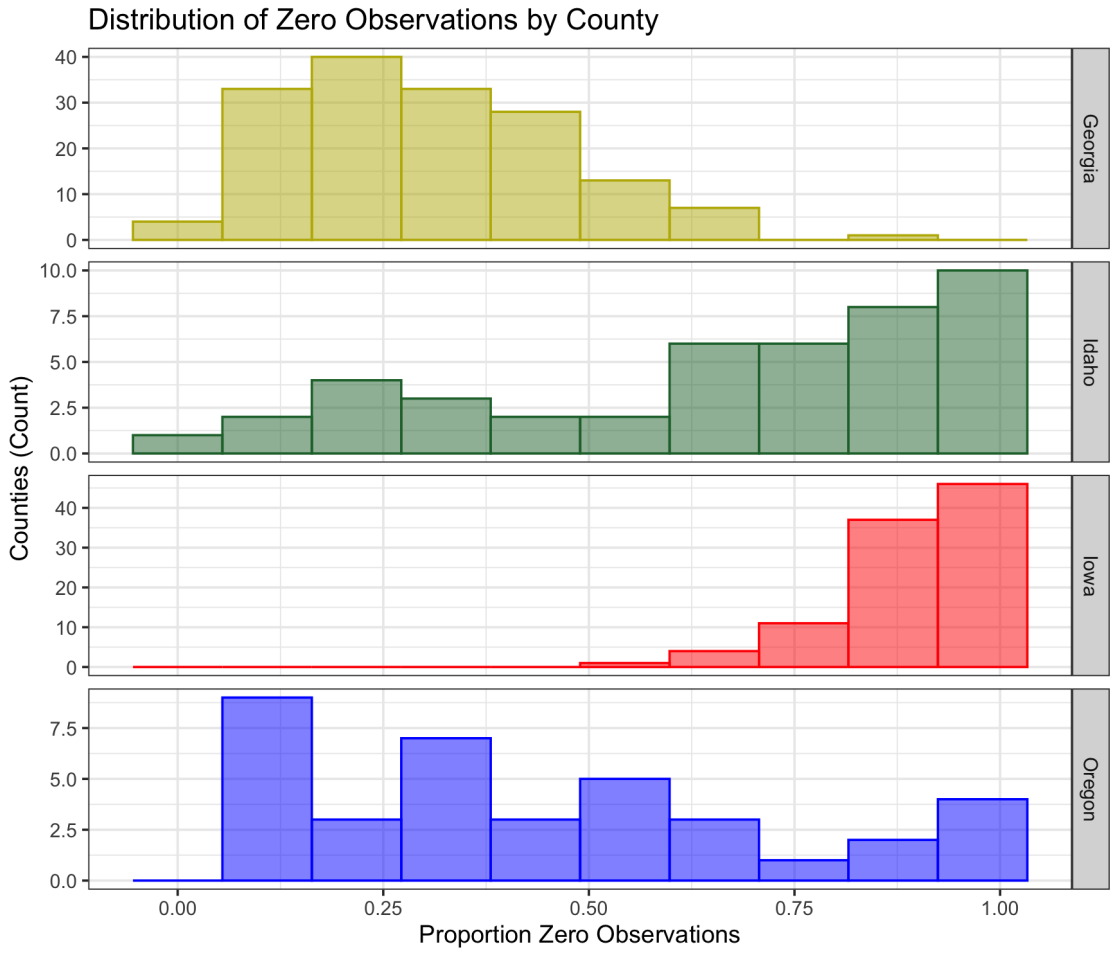


Figure 4: The distribution of the proportion of county observations of above ground carbon which are zero by state. The county-level observations are broken into 10 evenly spaced bins from 0-100% zero-inflation. Note that some states have significantly more counties than others.

estimators to perform well when compared to other estimators in the regions with highest zero-inflation, such as Iowa, as their model structures can specifically account for zeros.

3.2.4 Variable Correlation

Finally, we expected many of the predictor variables to be correlated. For example, intuitively we expect a strong correlation between average and minimum temperatures. Figure 5 shows the marginal distribution for each variable, joint distribution for each pair, and correlation between each of the study variables across all $N = 31,073$ observations, ignoring section boundaries. We note that almost all of the variables have significant correlation, which is unsurprising because similar environments tend to have similar conditions. For example, higher elevation (`elev`) has a strong negative correlation with average temperature (`tmean`) of -0.795, which can be explained by adiabatic cooling. We note that `tcc16` is the auxiliary variable with the the strongest correlation with `carbon`. This is logical because `tcc16` measures tree canopy cover which should intuitively correlate with the amount of tree at a given location. This finding motivated us to run our simulation in Section 3.4 with and without `tcc16` as a predictor. Alongside large correlation coefficients, we notice that the joint densities seem to cluster data in several distinct groups, usually 3 or 4 are visible, which we expect to be from the states. The marginal histograms follow a similar pattern often with several modes. Perhaps the more flexible models, like RF and SMERF, will be able to separate the state eco-sections and result in improved predictive power.

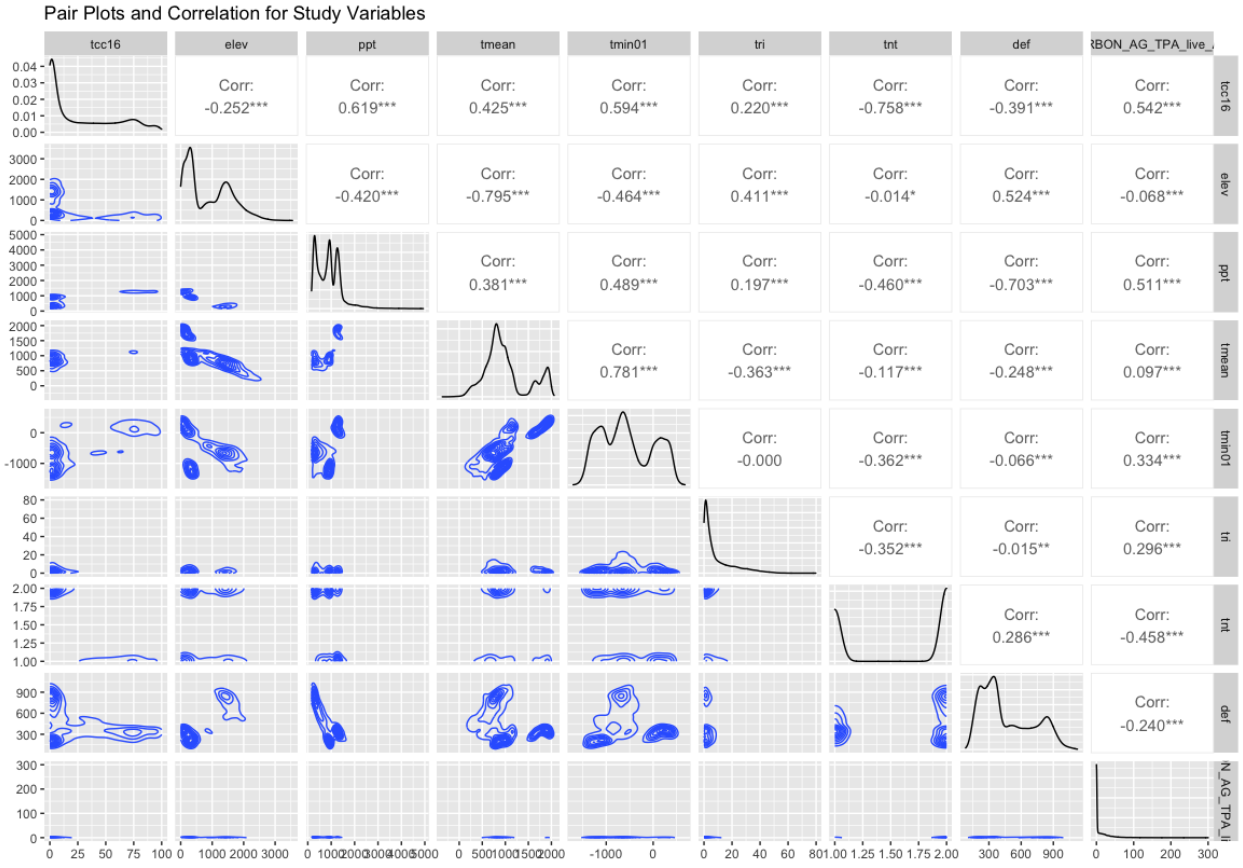


Figure 5: Joint density plots, marginal density plots, and correlations between all predictor variables and the response variable. The variables are listed at the top and right of each column of plots. The histograms on the variables themselves are located along the diagonal in black, the pairwise density plots are below the diagonal in blue, and the the correlations, along with asterisks to denote significance are in the upper diagonal.

3.3 Synthetic Simulation Study

3.3.1 Motivation

To better understand how the estimators behaved, we generated a synthetic dataset and ran an identical simulation study design to the one we planned to use for the FIA dataset. By using synthetic data we could control the relationship between the predictor variables and response variables and tune the influence of the small areas. As we chose to not generate zero-inflated data for this study, we omitted the zero-inflation two-part mixed model (as in this setting it would be a less efficient UE), fitting on the remaining 5 models: PS, AE, UE, RF, and SMERF.

3.3.2 Data Generation

For our simulation design we generated $N = 2,000$ observations across $D = 8$ small areas, nominally labeled A, B, C, D, E, F, G, H. We began by generating 8 small area means, μ_i, for $i = 1, \cdots, D$ using a beta distribution to ensure non-negative small area effects and then scaled the means to be between 0 and 100:

$$\mu_i \sim 100 * beta(3, 3)$$

We chose a $beta(3, 3)$ distribution as it resembles a normal which we'd expect for small area effects however is restricted to be positive like our response variable. To generate our 2000 auxiliary measurements, $\vec{x}_{ij}$ and response variables y_{ij} we used the following distribution:

$$\vec{x}_{ij} \sim \begin{pmatrix} \mathcal{N}(100, 100) \\ \mathcal{N}(50, 400) \\ bern(0.5) \end{pmatrix} \qquad y_{ij} \sim \mathcal{N}(0, 25) + 5 \cdot \vec{x}_{ij}^3 + \mu_i$$

where $\vec{x}_{ij}^3$ is the third element of $\vec{x}_{ij}$ and μ_i is the small area mean. We choose to assign pairs $(\vec{x}_{ij}, y_{ij})$ randomly with equal probability. Now because y_{ij} is only correlated with $\vec{x}_{ij}^3$ the other two predictors should have no impact, as they are pure random noise. This simple structure, without correlation between predictors or zero-inflation, will allow us to see the effects of just the sample size and model on bias and empirical mean

squared error.

3.3.3 Simulation Design

With our N simulation observations, we next generated 500 simulation samples for each value of $n = 8, 16, 32$ across each of the small areas. To do this we grouped the $2,000$ simulation observations by small area and uniformly at random selected n observations for each area. With our samples in hand, we then fit each of the 5 models to each sample, and then calculated the bias and MSE across all 500 samples for each n.

3.3.4 Synthetic Simulation Results

The first observation to be made from Figure 6 is that the RF model performs significantly worse than all others as measured by both percent bias and root mean squared error. The RFs poor performance is not surprising when we consider it's the only model that does not allow for small area effects. As our synthetic data is generated using large small area differences between the regions, which are not predicted by any of the response variables, we'd expect any model without the flexibility to encode differences between the small areas to perform extremely poorly. These initial finding highlights the importance of using small area models in settings where there are differences between small areas that are not captured by the predictors. In the forestry setting these differences could be a result of human logging practices, different tree species and growth patterns, or the presence of an invasive species among others.

Amongst PS, UE, AE, and SMERF, we make the following three observations in Figure 6:

1. In general with increasing n, the RMSE decreases. This makes sense as larger sample sizes typically reduce variance. We see a decrease in bias with increasing s, however the trend is less consistent across the small areas.

2. The PS estimator, the only direct estimator, performs worse than AE, UE, and SMERF, particularly on RMSE for small sample sizes ($n = 8$). As PS only takes one stratified variable we fed it the 3rd component of, x, $x_i[3]$, which is a binary stratification variable and the only true predictor. So in fact, unlike the Simulation with FIA data discussed in Section 3.4, the PS estimator received the same amount of predictive information as all other estimators. This finding underscores that for small

sample sizes, indirect estimators are typically able to make better predictions across regions.

3. There is no clear winner between the AE, UE, and SMERF models. We note that the bias percentages are extremely low, on the order of 0.5-2%, while RMSE values are mostly less than 3, indicating estimates are typically close to the true small area mean. This suggests the models are effective at capturing the relationship we encoded between the predictors and response.

We conclude that additional data complexity will have to be added to separate the performance of the AE, UE, and SMERF models. Perhaps this complexity will be present in the 4-state FIA dataset.

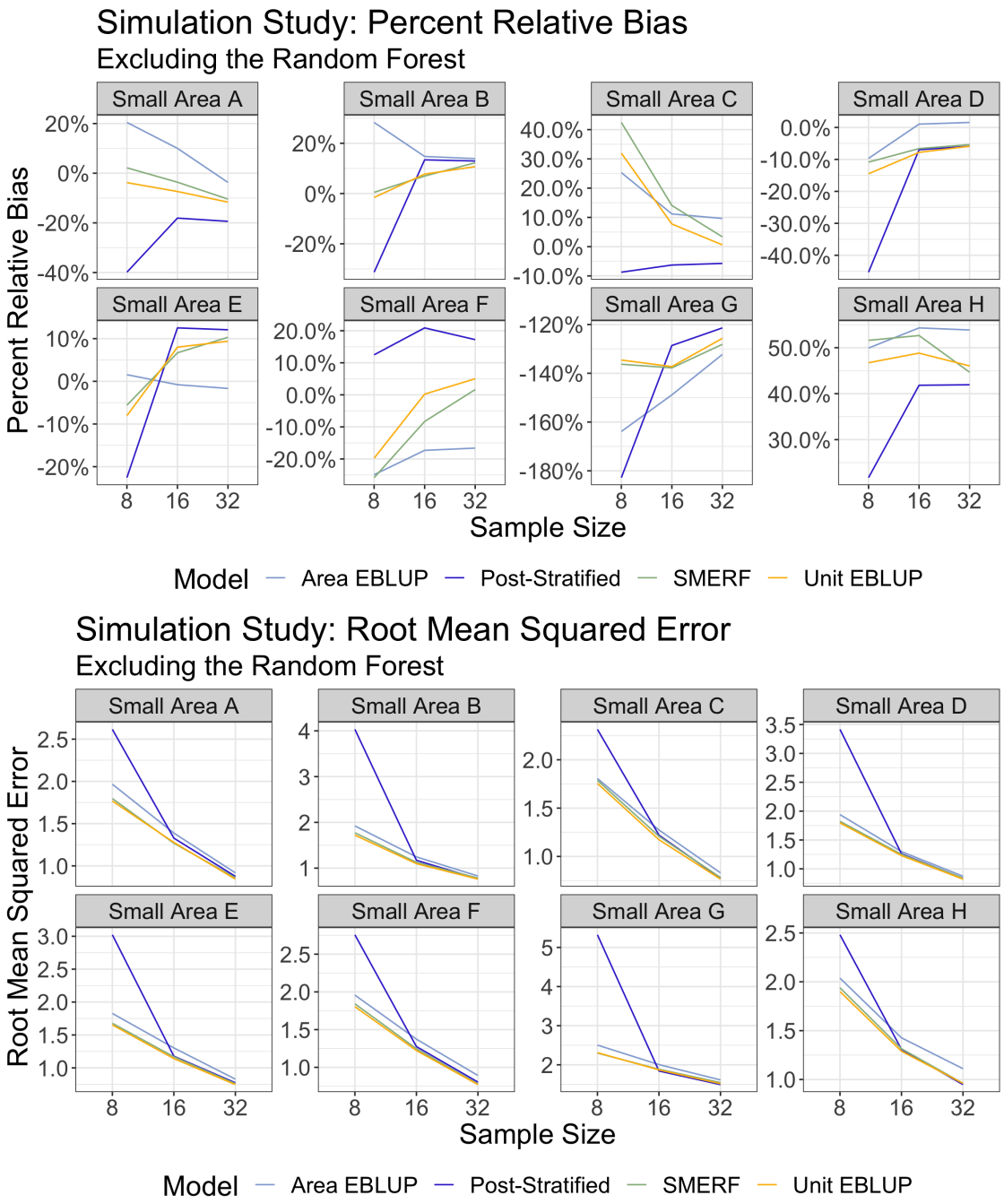


Figure 6: Synthetic Simulation Results. Top: empirical percent bias excluding the RF model across all 8 small areas in simulation study. Bottom: root mean squared error for all 4 models (excluding RF) across all study regions. Within each panel, each line represents a specific model and is colored accordingly and tracks the change in bias or RMSE across simulation sample sizes. The results including the RF model are included in Appendix C

3.4 FIA Simulation Results

3.4.1 Simulation Design with FIA data

As there was a wide range of the total samples in each group, we limited our study to sections which had more than 200 total observations so that the sampled sizes (4, 8, 16, 32) would remain a small fraction of the total observations. After applying this threshold, 30 sections remained across the four-state dataset. We then generated 2,000 simulation samples as described in Section 2.1 for each sample size and fit each of the six study models; PS, AE, UE, ZI, RF, and SMERF, to each simulation sample as described in Section 2. After fitting each model we predicted across all 30,073 observations. We calculate the model bias and root mean squared error across each model and simulation sample size.

3.4.2 Accuracy and Precision at the Section Level

By evaluating bias and RMSE across all sample sizes, models, and sections; a total of 1440 statistics, we can find which model performs best. Figures 7 and 8 contains the full simulation results which immediately reveal that the RF and SMERF models do not outperform the other models across the section samples. We had hoped that the SMERF model, which combines the flexibility to separate zero observations (see Figure 1) and section level random effects, would shine both in accuracy and precision.

As a visual inspection did not reveal a clear best-performing model, we re-ordered the facets by amount of carbon to begin finding patterns. Carbon, as our response variable, is important to model accurately in high carbon sections because this is more to the logging industry and because changes in high-carbon regions have a larger impact on terrestrial carbon storage and ultimately atmospheric CO_2. Modeling carbon in low carbon regions, particularly those affected by fires, is important for understanding forest regeneration among other reasons.

We first examine the sections with low carbon, examining the top 2 rows of both the percent relative empirical bias and root mean squared error plots in Figures 7 and 8. In the top row of Figure 7, we see that most models have a positive bias and that the PS and AE models perform the best both across both bias and RMSE (in Figure 8). The positive bias is intuitive because when carbon levels are very close to 0, it's harder to underestimate, particularly for models which can only produce positive predictions of

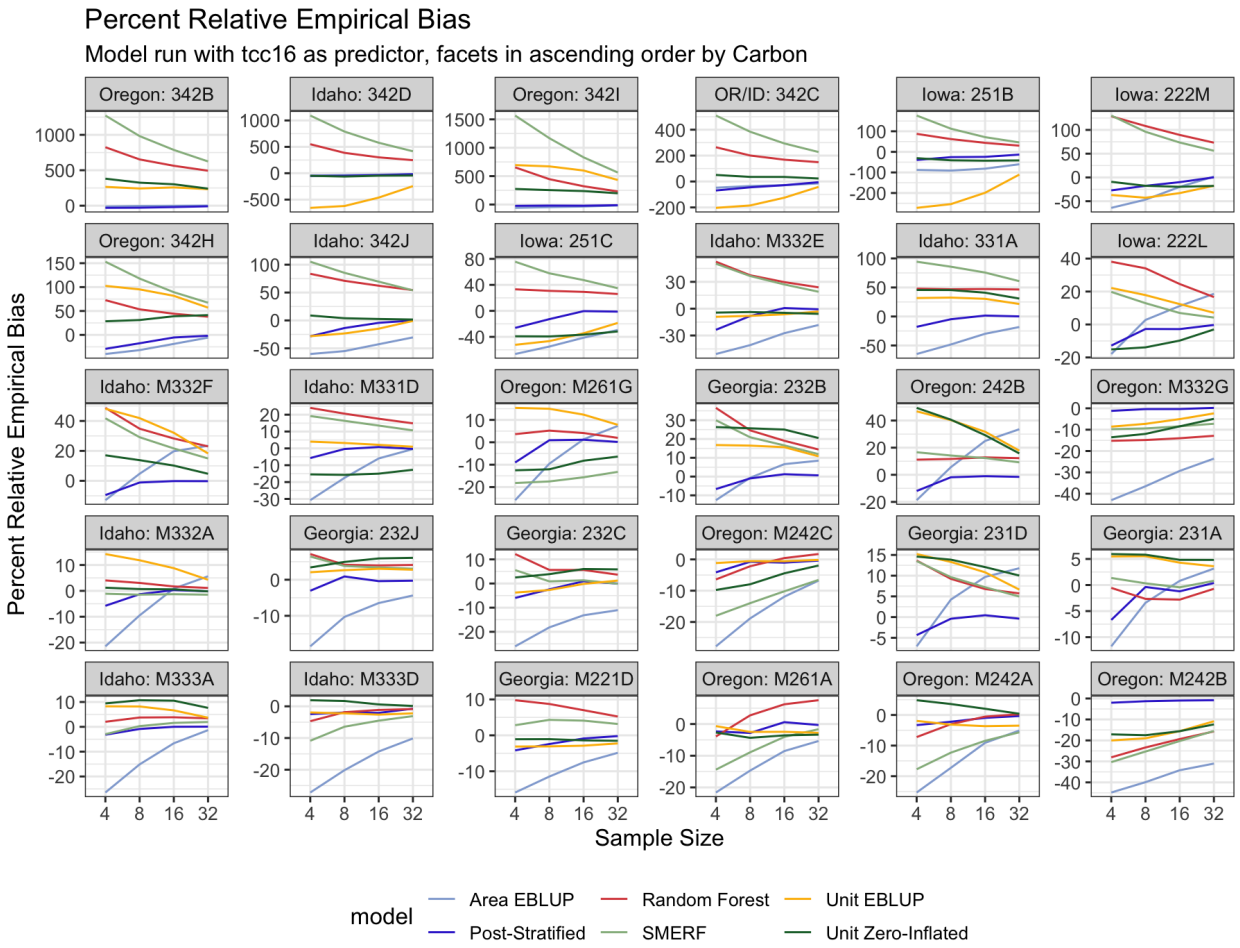


Figure 7: Bias results from the 4-state simulation. Each of the 30 sections is represented separately in a facet, which are ordered by increasing Carbon (left to right, top to bottom). Each model is denoted by a color and tracks performance across sample sizes. The y-axis reports the percent relative bias.

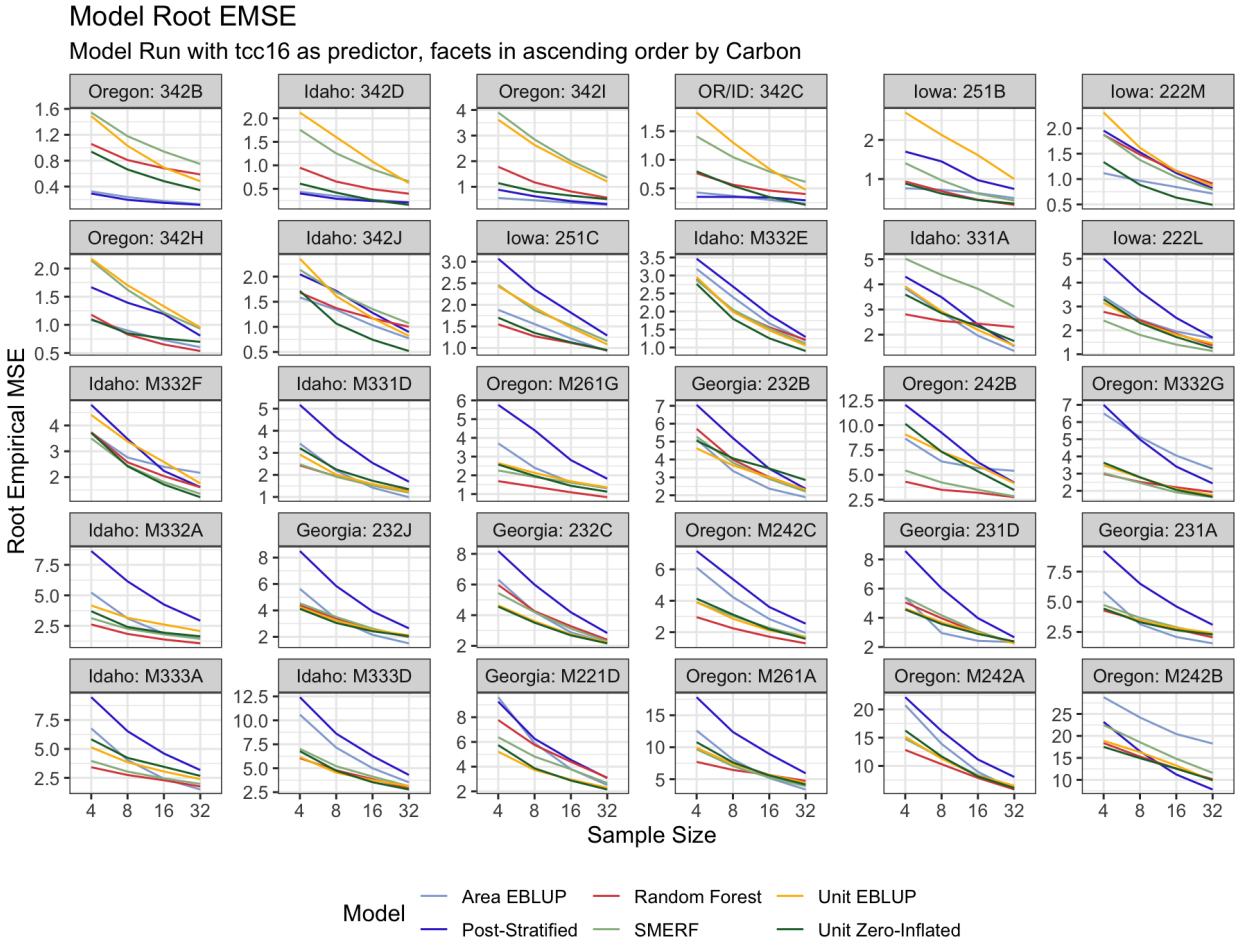


Figure 8: RMSE results from the 4-state dataset simulation run. The plot is broken into 30 facets, one for each section in the state data with more than 200 total observations and are ordered by increasing Carbon (left to right, top to bottom). The axis records the sample size per section used to fit each of the six models differentiated by color. The y-axis reports the root mean squared error.

section carbon, including the PS, RF, and SMERF models. Perhaps this explains why RF and SMERF perform particularly poorly in these sections. Another key story is the comparison between the ZI model and UE in regions where there is low carbon, which typically have higher zero-inflation. We find that across most sections with relatively low carbon, the ZI model has lower bias and much lower RMSE across these sections, lending evidence that in these settings the ZI model should be applied. These results agree with the ZI model's theoretical intuition, which says that the model is not model mis-specified because of its two-part construction, see Section 2.3. Notably, at higher carbon, and correspondingly lower proportion of zero observations, the zero-inflated model performs roughly equivalently to the UE, suggesting that performance gains from this model only occur at high levels of zero-inflation. We conclude that in low carbon settings, the zero-inflated model is the strongest unit-level model, while both the PS and AE are also strong candidates. Next, we turn our attention to the story where average section carbon is high. In these regions, the bias corresponds to a much higher error in carbon; thus in the sense of absolute error, bias and RMSE matter more in the high carbon regime. We first examine bias, finding that the AE performs particularly poorly in comparison to the other models which tend to cluster together and has a large negative bias. The AE sees significant improvement as the sample size increases, sometimes converging on the performance of the other models when the sample size was 32. We find that the PS estimator typically has significantly higher RMSE than the other models, however from a visual inspection we conclude that aside from the AE and PS estimators, the other models perform similarly at high carbon.

3.4.3 Pairwise Model Comparison: Who Wins?

Going beyond visual inspection, we hoped to evaluate the model's overall performance across all regions and sample sizes to be able to provide a broad sense of model performance. To do so we compared each pair of models' bias and RMSE against each other. We then defined a model to "win" at a given sample size and section if it had lower bias or RMSE. We then tallied the total number of wins for each model. After comparing all of the models, we calculated the total number of wins each model had over the others as a proxy for overall model performance relative to the rest of the estimator suite.

Table 1 contains the results of the bias pairwise comparison. In the model runs with and without `tcc16` as

Pairwise Bias Comparison

Number of wins of row model over column model (without tcc16)

Model	Area-EBLUP	Post-Stratified	Random-Forest	SMERF	Unit-EBLUP	Unit-ZI	Total
Area-EBLUP	0	14	64	62	49	56	245
Post-Stratified	106	0	112	109	108	113	548
Random-Forest	56	8	0	52	41	41	198
SMERF	58	11	68	0	41	46	224
Unit-EBLUP	71	12	79	79	0	64	305
Unit-ZI	64	7	79	74	56	0	280

Pairwise Bias Comparison

Number of wins of row model over column model (with tcc16)

Model	Area-EBLUP	Post-Stratified	Random-Forest	SMERF	Unit-EBLUP	Unit-ZI	Total
Area-EBLUP	0	12	55	59	54	45	225
Post-Stratified	108	0	109	111	103	98	529
Random-Forest	65	11	0	65	51	42	234
SMERF	61	9	55	0	36	34	195
Unit-EBLUP	66	17	69	84	0	57	293
Unit-ZI	75	22	78	86	63	0	324

Table 1: Between model comparisons across all 30 subsections and sample sizes. Each entry represents the total number of times the bias was lower in the row model than the column model out of the 120 different section and sample size comparisons. For example, the Post-Stratified estimator had lower bias 106 times when compared with the area-EBLUP when the area-EBLUP was run without `tcc16`. We total the row model's overall number of wins in the last column which we then treat as a proxy for model performance compared to the ensemble in general.

a predictor, the PS estimator performs extremely well scoring 548 and 529 total wins out of 600 possible, respectively. We expect the slight drop in the total wins for PS in the scenario with `tcc16` is due to the added to the strength of `tcc16` as a predictor gained by the other models; recall that PS is fit only using the categorical tnt (tree/no tree) predictor. With `tcc16`, we also see that the ZI model wins more times against the UE model, suggesting that `tcc16` is a powerful predictor for the probability that an observation is zero that can be captured by the GLM in the ZI model. This is intuitive in the problem setting because if `tcc16` is low there is little tree cover, which likely corresponds with lower carbon. Alongside the AE, the machine learning models, RF and SMERF, perform poorly compared to the other models. Interestingly, the SMERF performs comparitively worse when `tcc16` is added as a predictor, even when compared with the PS estimator which did not observe `tcc16`. It is unclear why adding a strong predictor would make SMERF decrease in relative predictive power to PS.

Turning our attention to the results of the RMSE pairwise comparison in table 2, we find that in both simulations, the ZI model has the most total wins, and wins against every other model more than 50% of the time (> 60 wins). The second place overall winner is the RF in the setting with `tcc16` and the AE model in the setting without `tcc16`. Notably, the PS estimator has the fewest total wins, and all models outperform this estimator, and by a particularly wide margin when `tcc16` is included as a predictor. Examining the results in Table 1 and 2 we find that, particularly when `tcc16` is used as a predictor, the ZI model is the strongest; it has the best MSE across all regions and scores second best on bias. While PS scores best on bias, it has the lowest precision of all 6 models. We conclude that the ZI model is the best balance between accuracy and precision.

Pairwise Empirical MSE Comparison

Number of wins of row model over column model (without tcc16)

Model	Area-EBLUP	Post-Stratified	Random-Forest	SMERF	Unit-EBLUP	Unit-ZI	Total
Area-EBLUP	0	78	62	63	69	58	330
Post-Stratified	42	0	43	49	47	35	216
Random-Forest	58	77	0	66	76	43	320
SMERF	57	71	54	0	64	49	295
Unit-EBLUP	51	73	44	56	0	37	261
Unit-ZI	62	85	77	71	83	0	378

Pairwise Empirical MSE Comparison

Number of wins of row model over column model (with tcc16)

Model	Area-EBLUP	Post-Stratified	Random-Forest	SMERF	Unit-EBLUP	Unit-ZI	Total
Area-EBLUP	0	99	54	62	69	46	330
Post-Stratified	21	0	24	30	34	19	128
Random-Forest	66	96	0	85	84	53	384
SMERF	58	90	35	0	65	41	289
Unit-EBLUP	51	86	36	55	0	33	261
Unit-ZI	74	101	67	79	87	0	408

Table 2: Between model comparisons of MSE / RMSE for all 6 models across sample sizes and sections. Entries denote the number of times the row model has lower RMSE than the column model. The final column gives the total number of wins the row model had over all the other models.

4 Discussion

4.1 Future Work

4.1.1 Why do our Machine Learning Models Fall Short?

The goal of applying ML models to the forestry setting was to improve estimation. The RF and SMERF models fell short of this objective as they did not perform better along any dimension we explored: high/low carbon, small/large sample sizes, or high/low variable correlation, See Appendix D. There are several possible explanations for this shortcoming:

1. **The data has an inherent linear structure that can be effectively captured by linear models.** For example, higher tree canopy cover (`tcc16`) indicates that more trees are present, and thus more carbon. Thus, despite being less expressive, the UE and ZI in particular, have a structure built into them which results in better performance than a model that attempts to make predictions simply by clustering the data (e.g. a decision tree).

2. **The tuning performed on the models was insufficient.** Before running our simulation study we examined several decision trees trained on samples to determine how deep to split the data, and assumed that a forest with 500 trees was sufficient. Retrospectively, before running the simulation, performing cross validation on these parameters across a subset of the data could have helped choose a combination of parameters which would have performed better. In general, however, random forests tend to require the least tuning across common machine learning models to perform well. The lower amount of hyper-parameter tuning was one of the advantages of the model and why we thought it was particularly well suited to the problem, all puns aside.

3. **Different machine learning frameworks could do better.** Perhaps other models, such as neural networks or gradient boosted trees, would be better able to capture the relationship between the predictors and our response, carbon. We focused on random forest models because the natural extension, the SMERF, falls into the small area estimation framework as it explicitly accounts for variation between domains not explained by the predictors. Interestingly, in Section 3.4.2 we found that the Random

Forest often outperformed the SMERF, suggesting that at least for the dataset we focused on, perhaps the need for small area effects is overstated. Perhaps the random forest is able to effectively capture the complexity of each ecosystem in the predictors, negating the benefits of including small areas as in the SMERF. Future comparative studies could assess how to tune the small area effects differently.

4.1.2 Improving Estimation Beyond our Simulation Study

Ultimately, the goal of any comparative simulation study is to improve prediction on the domain of interest. In the forestry setting, our goal was to improve estimation of forest attributes across small areas. Current discussions [11–13] suggest forestry is increasingly interested in areas smaller than the eco-sections we studied, with interest in estimation going all the way to the county or town level. As our simulation study focused only on the subset of pixels in the 4-state dataset that had corresponding y_{ij} measurements of above ground carbon, the estimates are not intended to provide estimates of true average carbon in these regions. To make competitive predictions with our models we would fit a model on *all* of these available carbon measurements and then predict across the US-wide 30×30 meter grid of auxiliary variables. In effect, the population of this simulation study would become the sample. The computational cost of doing this in a simulation study was too high and also does not permit a comparison of estimators as the ground truth, μ_i is unknown. Beyond using a pixel-level dataset to improve prediction accuracy by using the 30×30, we might also look to improve accuracy by incorporating more predictors, such as those from the GEDI satellite.

4.2 Conclusions

Motivated by congressional directives [12, 13], the carbon storage potential of forests [9, 15–17], and the ever-looming threat of climate change [22], this project set out to address the need for modern estimators in small area estimation for forest inventory projections. With access to meticulously recorded estimates of live above-ground carbon and a dearth of remotely-sensed observations across 4 ecologically-heterogeneous states, we designed a simulation study to explore how best to improve forest carbon prediction. The estimators we carefully chose to compare consisted of three traditional estimators: the post-stratified, area-EBLUP, and unit-EBLUP models, and three modern estimators: the zero-inflated, random forest, and mixed effects

random forest estimators. The post-stratified estimator was chosen as our baseline because it currently produces all of the FIA's inventory estimates. While it performs well when applied to large domains, such as at the state-level, we found that it's performance drops considerably more than other models when the sample size is very small ($n = 4, 8$). The broadly linear structure of forestry data allows models like the area- and unit-EBLUPs to have higher precision than PS. However, the zero-inflated model, which is a problem-specific estimator built to address the high frequency of zero-observations in forestry data, outstrips them when applied to the forestry setting.

While we had hoped that a minimally-tuned random forest or mixed-effect random forest would outperform the other estimators, the story ended up being far more complex. A myriad of reasons could explain why (see Section 4.1.1) and only serve to motivate exploration into future modeling efforts. For example, image-based learning, such convolutional neural networks, or architectures which build explicit connections and distances between observations, such as graph neural networks, have yet to reach the forestry domain and could revolutionize prediction. Such architectures, which incorporate an explicit spatial component could perhaps generalize the small area estimation framework to examine effects along a continuum as opposed to discrete small areas. Forest estimation, an interdisciplinary problem requiring advanced satellite imagery, computing power, forest measurements taken by hand, and robust statistical estimators, is motivated by a need to understand our natural resources and an uncertain future. It's easy to get lost in the methodological details of a simulation study, but its important to remain aware of the larger impact and look towards new avenues for improvement because, as they say, you might just miss the forest for the trees.

5 Acknowledgements

I would like to thank my thesis advisor, Kelly McConville, for guiding this project from its inception. Her forest of connections made this project possible. Jonathan Thompson, a researcher at the Harvard Forest provided guidance and suggestions for how to best contextualize the application area that we are working on. His "forestry-sense" was invaluable. Other connections included FIA researchers Tracey Frescino, Gretchen Moisen, and Grayson White who got us access to the data sets we needed and answered my forestry questions like: "What are the units on this variable?"

I would also like to acknowledge Josh Yamamoto's contributions to the project. As part of the Undergraduate Forestry Data Science Lab Summer 2022 research group, we collaboratively worked to build and run simulations on a different dataset to evaluate the zero-inflation model against the the post-stratified estimator and linear models. Josh contributed towards this project by turning these models (AE, UE, and ZI) into production code.

Also, to my parents for their unwavering confidence and support throughout all my endeavors.

Finally, to the trees for their physical support, inspiration, and oxygen. Without them, this project, and likely humanity, would not exist.

References

1. Swain, C. *Black faces, black interests : the representation of African Americans in Congress* eng. ISBN: 0-674-07615-X (Harvard University Press, Cambridge, Mass., 1993).

2. Friedman, J. N. & Holden, R. T. The Rising Incumbent Reelection Rate: What's Gerrymandering Got to Do With It? *The Journal of Politics* **71.** Publisher: The University of Chicago Press, 593–611. ISSN: 0022-3816. https://www.journals.uchicago.edu/doi/full/10.1017/S0022381609090483 (2023) (Apr. 2009).

3. Chen, J. & Cottrell, D. Evaluating partisan gains from Congressional gerrymandering: Using computer simulations to estimate the effect of gerrymandering in the U.S. House. en. *Electoral Studies* **44,** 329–340. ISSN: 0261-3794. https://www.sciencedirect.com/science/article/pii/S0261379416303201 (2023) (Dec. 2016).

4. Kennedy, C. *et al.* An Evaluation of the 2016 Election Polls in the United States. *Public Opinion Quarterly* **82,** 1–33. ISSN: 0033-362X. https://doi.org/10.1093/poq/nfx047 (2023) (Mar. 2018).

5. Berry, S. M., Broglio, K. R., Groshen, S. & Berry, D. A. Bayesian hierarchical modeling of patient subpopulations: Efficient designs of Phase II oncology clinical trials. *Clinical Trials* **10.** Publisher: SAGE Publications, 720–734. ISSN: 1740-7745. https://doi.org/10.1177/1740774513497539 (2023) (Oct. 2013).

6. Riordan, D. O. *et al.* Prevalence of potentially inappropriate prescribing in a subpopulation of older European clinical trial participants: a cross-sectional study. en. *BMJ Open* **8.** Publisher: British Medical Journal Publishing Group Section: Pharmacology and therapeutics, e019003. ISSN: 2044-6055, 2044-6055. https://bmjopen.bmj.com/content/8/3/e019003 (2023) (Mar. 2018).

7. Tabari, H. *et al.* Local impact analysis of climate change on precipitation extremes: are high-resolution climate models needed for realistic simulations? English. *Hydrology and Earth System Sciences* **20.** Publisher: Copernicus GmbH, 3843–3857. ISSN: 1027-5606. https://hess.copernicus.org/articles/20/3843/2016/ (2023) (Sept. 2016).

8. Spawn, S. A., Sullivan, C. C., Lark, T. J. & Gibbs, H. K. *Harmonized global maps of above and belowground biomass carbon density in the year 2010* en. Number: 1 Publisher: Nature Publishing Group. https://www.nature.com/articles/s41597-020-0444-4 (2023) (Wiley, 2003).

9. Goetz, S. J. *et al.* Mapping and monitoring carbon stocks with satellite observations: a comparison of methods. *Carbon Balance and Management* **4,** 2. ISSN: 1750-0680. https://doi.org/10.1186/1750-0680-4-2 (2023) (Mar. 2009).

10. Dubayah, R. *et al.* Global Ecosystem Dynamics Investigation (GEDI)GEDI L4B Gridded Aboveground Biomass Density, Version 2. en. Artwork Size: 0 MB Medium: GTiff Publisher: ORNL Distributed Active Archive Center Version Number: 2, 0 MB. https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=2017 (2023) (2022).

11. Prisley, S. *et al.* Needs for Small Area Estimation: Perspectives From the US Private Forest Sector. *Frontiers in Forests and Global Change* **4.** ISSN: 2624-893X. https://www.frontiersin.org/articles/10.3389/ffgc.2021.746439 (2023) (2021).

12. USDA, E. *Agricultural Act of 2014: Highlights and Implications* 2014. https://www.ers.usda.gov/agricultural-act-of-2014-highlights-and-implications/ (2023).

13. USDA, E. *Agriculture Improvement Act of 2018: Highlights and Implications* 2018. https://www.ers.usda.gov/agriculture-improvement-act-of-2018-highlights-and-implications/ (2023).

14. Bechtold, W. A., Patterson, P. L. & Editors. The enhanced forest inventory and analysis program - national sampling design and estimation procedures. en. *Gen. Tech. Rep. SRS-80. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 85 p.* **080.** https://www.fs.usda.gov/research/treesearch/20371 (2022) (2005).

15. Bastin, J.-F. *et al.* The global tree restoration potential. en. *Science* **365,** 76–79. ISSN: 0036-8075, 1095-9203. https://www.science.org/doi/10.1126/science.aax0848 (2022) (July 2019).

16. Veldman, J. W. *et al.* Comment on "The global tree restoration potential". *Science* **366.** Publisher: American Association for the Advancement of Science, eaay7976. https://www.science.org/doi/10.1126/science.aay7976 (2022) (Oct. 2019).

17. Waring, B. *et al.* Forests and Decarbonization – Roles of Natural and Planted Forests. *Frontiers in Forests and Global Change* **3.** ISSN: 2624-893X. https://www.frontiersin.org/articles/10.3389/ffgc.2020.00058 (2023) (2020).

18. Hawbaker, T. J. *et al.* Changes in wildfire occurrence and risk to homes from 1990 through 2019 in the Southern Rocky Mountains, USA. en. *Ecosphere* **14.** Number: 2, e4403. https://www.fs.usda.gov/research/treesearch/65836 (2023) (2023).

19. Forest Inventory Data & Tools (FIA). en. https://www.fs.usda.gov/research/products/dataandtools/forestinventorydata (2023) (Jan. 2022).

20. Kayler, Z., Janowiak, M. & Swanston, C. Global Carbon. en-us. *U.S. Department of Agriculture, Forest Service, Climate Change Resource Center*. https://www.fs.usda.gov/ccrc/topics/global-carbon (2017).

21. Köchy, M., Hiederer, R. & Freibauer, A. Global distribution of soil organic carbon – Part 1: Masses and frequency distributions of SOC stocks for the tropics, permafrost regions, wetlands, and the world. English. *SOIL* **1.** Publisher: Copernicus GmbH, 351–365. ISSN: 2199-3971. https://soil.copernicus.org/articles/1/351/2015/soil-1-351-2015.html (2023) (Apr. 2015).

22. Masson-Delmotte, V., Zhai, P. & Pirani, A. IPCC, 2021: Summary for Policymakers. In: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. *IPCC*. https://www.ipcc.ch/report/ar6/wg1/chapter/summary-for-policymakers/ (2021).

23. McConville, K. S., Moisen, G. G. & Frescino, T. S. A Tutorial on Model-Assisted Estimation with Application to Forest Inventory. en. *Forests* **11.** Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, 244. ISSN: 1999-4907. https://www.mdpi.com/1999-4907/11/2/244 (2022) (Feb. 2020).

24. Goerndt, M., Monleon, V. & Temesgen, H. A comparison of small-area estimation techniques to estimate selected stand attributes using LiDAR-derived auxiliary variables. *Canadian Journal of Forestry Research. 41: 1189-1201* **41,** 1189–1201. https://www.fs.usda.gov/research/treesearch/46164 (2022) (2011).

25. White, G. W., McConville, K. S., Moisen, G. G. & Frescino, T. S. Hierarchical Bayesian Small Area Estimation Using Weakly Informative Priors in Ecologically Homogeneous Areas of the Interior Western Forests. *Frontiers in Forests and Global Change* **4.** ISSN: 2624-893X. https://www.frontiersin.org/articles/10.3389/ffgc.2021.752911 (2022) (2021).

26. Frescino, T. S., Patterson, P., Gretchen, M., Toney, C. & Elizabeth, F. *FIESTA: A Forest Inventory Estimation and Analysis R Package* 507 25th street, Ogden, UT, USA, 2009.

27. Yang, S.-I. *et al.* Characterizing height-diameter relationships for Caribbean trees using mixed-effects random forest algorithm. en. *Forest Ecology and Management* **524,** 120507. ISSN: 0378-1127. https://www.sciencedirect.com/science/article/pii/S0378112722005011 (2023) (Nov. 2022).

28. Freeman, E. A., Moisen, G. G., Coulston, J. W. & Wilson, B. T. Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance. *Canadian Journal of Forest Research* **46.** Publisher: NRC Research Press, 323–339. ISSN: 0045-5067. https://cdnsciencepub.com/doi/10.1139/cjfr-2014-0562 (2022) (Mar. 2016).

29. Young, J. D. *et al.* Effects of policy change on wildland fire management strategies: Evidence for a paradigm shift in the western US? en. *International Journal of Wildland Fire. 29: 857-877.* **29,** 857–877. https://www.fs.usda.gov/research/treesearch/61274 (2022) (2020).

30. Thompson, J. R., Canham, C. D., Morreale, L., Kittredge, D. B. & Butler, B. Social and biophysical variation in regional timber harvest regimes. en. *Ecological Applications* **27,** 942–955. ISSN: 1051-0761, 1939-5582. https://onlinelibrary.wiley.com/doi/10.1002/eap.1497 (2022) (Apr. 2017).

31. Breiman, L. Random Forests. *Machine Learning* **45,** 5–32. ISSN: 1573-0565. https://doi.org/10.1023/A:1010933404324 (Oct. 1, 2001).

32. Healey, S. P. *et al.* Mapping forest change using stacked generalization: An ensemble approach. en. *Remote Sensing of Environment. 204: 717-728.* **204,** 717–728. https://www.fs.usda.gov/research/treesearch/56397 (2023) (2018).

33. Gordon, A. D., Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification and Regression Trees.* in *Biometrics* **40.** ISSN: 0006341X Issue: 3 Journal Abbreviation: Biometrics (Sept. 1984), 874. https://www.jstor.org/stable/2530946?origin=crossref (2023).

34. Hajjem, A., Bellavance, F. & Larocque, D. Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation* **84.** Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00949655.2012.7 1313–1328. ISSN: 0094-9655. https://doi.org/10.1080/00949655.2012.741599 (2022) (June 2014).

35. Krennmair, P. Tree-Based Machine Learning in Small Area Estimation, 10 (2022).

36. Krennmair, P., Würz, N. & Schmid, T. *Analysing Opportunity Cost of Care Work using Mixed Effects Random Forests under Aggregated Census Data* Apr. 22, 2022. arXiv: 2204.10736[stat]. http://arxiv.org/abs/2204.10736 (2022).

37. Krennmair, P. & Schmid, T. Flexible domain prediction using mixed effects random forests. *Journal of the Royal Statistical Society: Series C (Applied Statistics).* arXiv:2201.10933 [stat], rssc.12600. ISSN: 0035-9254, 1467-9876. http://arxiv.org/abs/2201.10933 (2022) (Oct. 2022).

38. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39.** Publisher: [Royal Statistical Society, Wiley], 1–38. ISSN: 0035-9246. https://www.jstor.org/stable/2984875 (2023) (1977).

39. R Core Team. *R: A Language and Environment for Statistical Computing* https://www.R-project.org/ (R Foundation for Statistical Computing, Vienna, Austria, 2021).

40. Microsoft, C. & Weston, S. *doParallel: Foreach Parallel Adaptor for the 'parallel' Package* https://CRAN.R-project.org/package=doParallel (2020).

41. Dowle, M. & Srinivasan, A. *data.table: Extension of 'data.frame'* https://CRAN.R-project.org/package=data.table (2021).

42. Molina, I. & Marhuenda, Y. sae: An R Package for Small Area Estimation. *The R Journal* **7,** 81–98. https://journal.r-project.org/archive/2015/RJ-2015-007/RJ-2015-007.pdf (June 2015).

43. Boonstra, H. J. *hbsae: Hierarchical Bayesian Small Area Estimation* https://CRAN.R-project.org/package=hbsae (2022).

44. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67,** 1–48 (2015).

45. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2,** 18–22. https://CRAN.R-project.org/doc/Rnews/ (2002).

46. Wang, J. & Chen, L. S. *MixRF: A Random-Forest-Based Approach for Imputing Clustered Incomplete Data* https://CRAN.R-project.org/package=MixRF (2016).

47. USDA/USFS, M. *Multi-Resolution Land Characteristics (MRLC) Consortium — Multi-Resolution Land Characteristics (MRLC) Consortium* 2016. https://www.mrlc.gov/ (2023).

48. UDSA/USFS, L. *LANDFIRE Program: Data Products - Overview - LF 2010 (LF 2010 - LF_1.2.0)* 2010. https://landfire.gov/lf_120.php (2023).

49. Holden, Z. A. *et al.* TOPOFIRE: A Topographically Resolved Wildfire Danger and Drought Monitoring System for the Conterminous United States. EN. *Bulletin of the American Meteorological Society* **100.** Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society, 1607–1613. ISSN: 0003-0007, 1520-0477. https://journals.ametsoc.org/view/journals/bams/100/9/bams-d-18-0178.1.xml (2023) (Sept. 2019).

50. Climate Group, P. *PRISM Climate Group, Oregon State U* 2010. https://prism.oregonstate.edu/ (2023).

51. Rao, J. *Small Area Estimation, 2nd Edition* en-us. ISBN: 978-1-118-73578-7. https://www.wiley.com/en-us/Small+Area+Estimation%2C+2nd+Edition-p-9781118735787 (2023) (Wiley, 2003).

52. Pfeffermann, D., TERRYN, B. & Moura, F. Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries. *Survey Methodology* **34** (Dec. 2008).

53. Breiman, L. *Classification and Regression Trees* en. ISBN: 978-1-315-13947-0. https://www.taylorfrancis.com/books/mono/10.1201/9781315139470/classification-regression-trees-leo-breiman (2022) (Routledge, Oct. 2017).

A Pixel-level model and Cloud Computing

Our initial approach to this problem did involve prediction across this 30×30 meter gridded data for a eco-section in the Northwestern US, M333A which consisted of over 3 million pixel-level observations. Computationally this problem was much more challenging as roughly 150 billion predictions had to be done to produce estimates across all of the models. Particularly for the more complicated ZI and SMERF models, the runs would take close to one day to compute on a laptop computer. To avoid this problem, we designed an architecture to run the models on the Harvard Research Computing cluster which allowed us to submit multiple jobs on much larger machines. Ultimately we decided to work on the 4-state dataset which was 2.5 orders of magnitude smaller, negating the need for the cluster. Part of the reason for this decision was the need to accurately assess how our estimators performed as in order to have ground "truth" values for each section, our dataset needed to have y_{ij} values for each $\vec{x_{ij}}$. We explored several means of imputation, such as K-nearest neighbors, however we felt that evaluating our models against a synthetic dataset would introduce different biases across models, potentially altering the results. I produced a video detailing how to implement a simulation study on YouTube at: https://youtu.be/iYJke2XkSho.

B Implementation of the Zero-Inflated Model

All models, except the zero-inflation model, have already been implemented in packages in the R programming language [39, 42, 44–46]. As such, we include the production code we used to run the zero inflated model. Note that as the zero-inflated model is a two part model, with one component fitting the probability that the observations are zero and the second fitting a regression on the non-zero function observations. We fit both components with the `lme4` package [44]. The implementation of the zero-inflated model is as follows:

```
unit_zi <- function(samp_dat, pop_dat, formula, domain_level = "SECTION"){

  if (!rlang::is_formula(formula)) {
    formula <- as.formula(formula)
    message("model formula was converted to class 'formula'")
  }

  # creating strings of original X, Y names
  Y <- deparse(formula[[2]])
  X <- stringr::str_extract_all(deparse(formula[[3]]), "\\w+")[[1]]

  # function to fit a zero-inflation model
```

```r
  # this function pulls in domain_level from it's parent function input
  fit_zi <- function(samp_dat, pop_dat, formula) {

    Y <- deparse(formula[[2]])
    X <- stringr::str_extract_all(deparse(formula[[3]]), "\\w+")[[1]]

    # function will always treat domain_level as the random intercept
    rand_intercept <- paste0("( 1 | ", domain_level, " )")

    # of form y ~ x_1 + ... + x_n + (1 | domain_level)
    lin_reg_formula <- as.formula(
      paste0(deparse(formula[[2]]), " ~ ",
             deparse(formula[[3]]), " + ",
             rand_intercept)
    )

    # of form y != 0 ~ x_1 + ... + x_n + (1 | domain_level)
    log_reg_formula <- as.formula(
      paste0(deparse(formula[[2]]), " != 0 ~ ",
             deparse(formula[[3]]), " + ",
             rand_intercept)
    )

    # creating nonzero version of our sample data set
    nz <- samp_dat[samp_dat[ , Y] > 0, ]

    # fit linear mixed model on nonzero data
    lmer_nz <- suppressMessages(lme4::lmer(lin_reg_formula, data = nz))

    # Fit logistic mixed effects on ALL data
    glmer_z <- suppressMessages(
      lme4::glmer(log_reg_formula, data = samp_dat, family = "binomial")
    )

    lin_pred <- predict(lmer_nz, pop_dat, allow.new.levels = TRUE)
    log_pred <- predict(glmer_z, pop_dat, type = "response", allow.new.levels = TRUE)

    unit_level_preds <- lin_pred*log_pred

    # d x 2 dataframe
    # where d = # of domains
    zi_domain_preds <- data.frame(
      domain = pop_dat[ , domain_level, drop = T],
      unit_level_preds = unit_level_preds) %>%
      dplyr::group_by(domain) %>%
      dplyr::summarise(Y_hat_j = mean(unit_level_preds)) %>%
      ungroup()

    return(list(lmer = lmer_nz, glmer = glmer_z, res = zi_domain_preds))
  }
  # get model estimates for user supplied data
  original_res <- fit_zi(samp_dat, pop_dat, formula)

  return(original_res$res)
}
```

C Synthetic Simulation Study Random Forest Results

In Section 3.3.4, the results for the RF model were excluded from the full result plots because the RF's performance distorted the bias and RMSE plots. The RF performed particularly badly because the only causal effect on the response variables y_{ij} came from the small area effects. As the RF is not a small area model it could not capture these effects. As such, its performance suffered. The study showed that in general the RF can do quite badly if small area effects are comparatively large, although in the simulation study performed with real FIA data this was not the case. For completeness, we include a plot of the synthetic simulation study that displays the RF results.

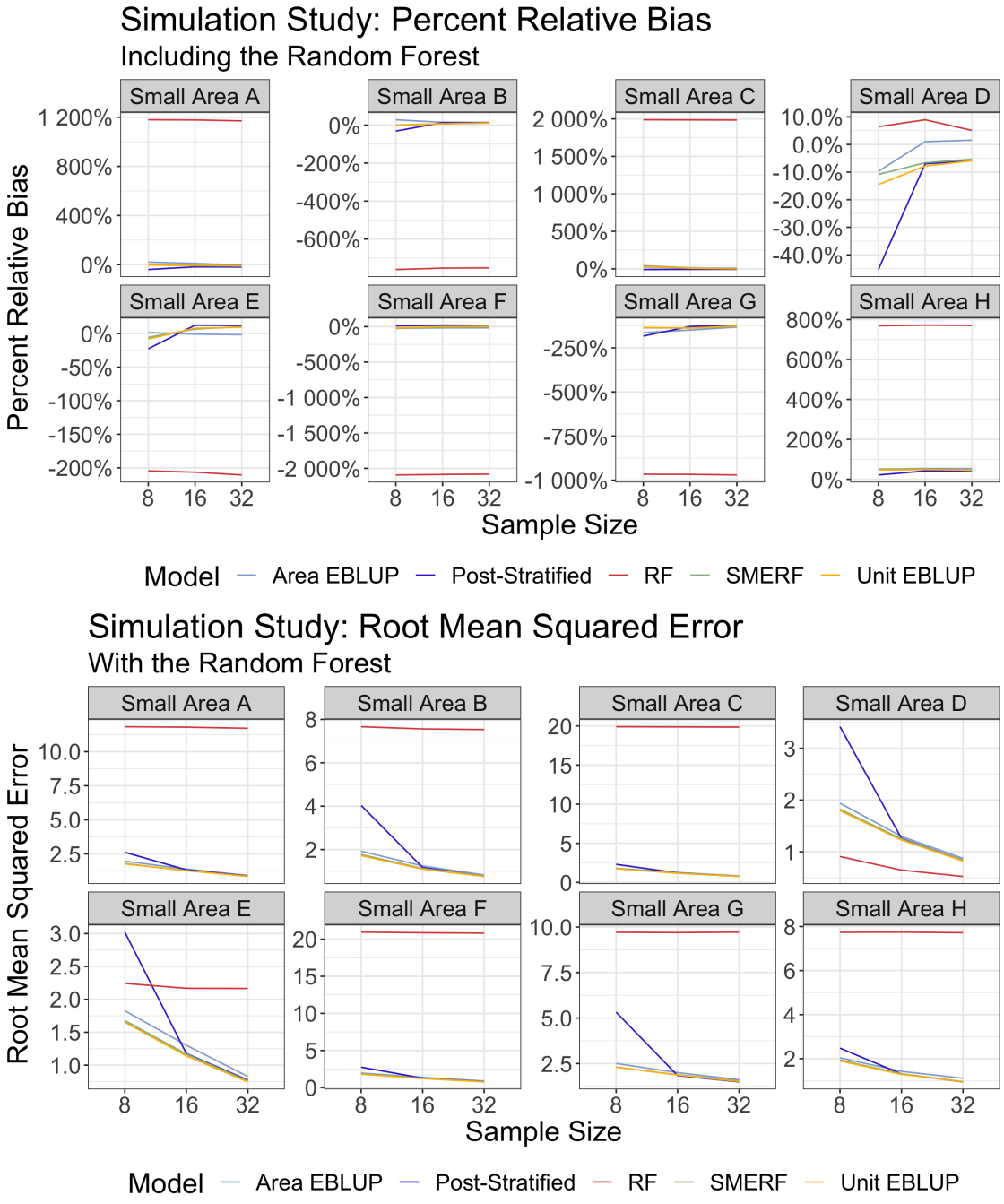


Figure 9: Synthetic Simulation Results. Top: empirical percent bias with the RF model across all 8 small areas in simulation study. Bottom: root mean squared error for all 4 models (with RF) across all study regions. Within each panel, each line represents a specific model and is colored accordingly and tracks the change in bias or RMSE across simulation sample sizes.

D Model Performance by Variable Correlation

In the analysis presented in Section 3.4.2 we focused on the relation between bias and RMSE as a function of average section carbon. This choice to consider the sections as a function of carbon was not an obvious choice. In addition to looking at the results ordered by carbon, we also examined the data as a function of the section percent zero and of the section R^2 correlation coefficient between `tcc16` and carbon. We plot both the bias and RMSE as a function of the section percent in Figure 10. Examining the top graphic in Figure 10 we notice several trends. Unsurprisingly, the average section carbon has a strong negative correlation with the section percent zero, high zero-inflation typically corresponds with low average section carbon. Furthermore, we see that near 100% zero-inflation the variance of the models percent relative bias explodes, which is consistent with the fact that the denominator (e.g. section average carbon) is close to zero. When comparing model performance, we note that the PS, AE, UE, and ZI models cluster around the zero-trend/zero-bias, while the RF and SMERF models are, on average, positive biased when the proportion of zero-inflation is very high. The bottom panel of Figure 10 shows that in general RMSE decreases with increasing section percent zero, however we attribute this decrease of the decrease in total carbon, thus the average deviations tend to be smaller. We note that the PS and AE estimators tend to have higher RMSE values at lower section percent zero-inflation levels, and see improvements when the value increases.

We also used the section Pearson correlation coefficient (R^2) between `tcc16`, x_{ij}, (no longer a vector) and carbon, y_{ij}, to order our plots. We ran a regression for each section, i, of the form:

$$y_{ij} = \beta_i \cdot x_{ij} + \epsilon_{ij}$$

By fully decoupling the sections we the extracted the R^2 correlation coefficient from each of the 30 linear regressions. The hope was to show that the linear models perform better when the correlation coefficient is high, because the structure of the data and model is more similar. Figure 11 shows the model order and error magnitude remains fairly static across the range of R^2 values. We conclude that the correlation coefficient in this simulation setting has little impact on model performance. However, we do observe that there is some correlation between carbon and the R^2 value.

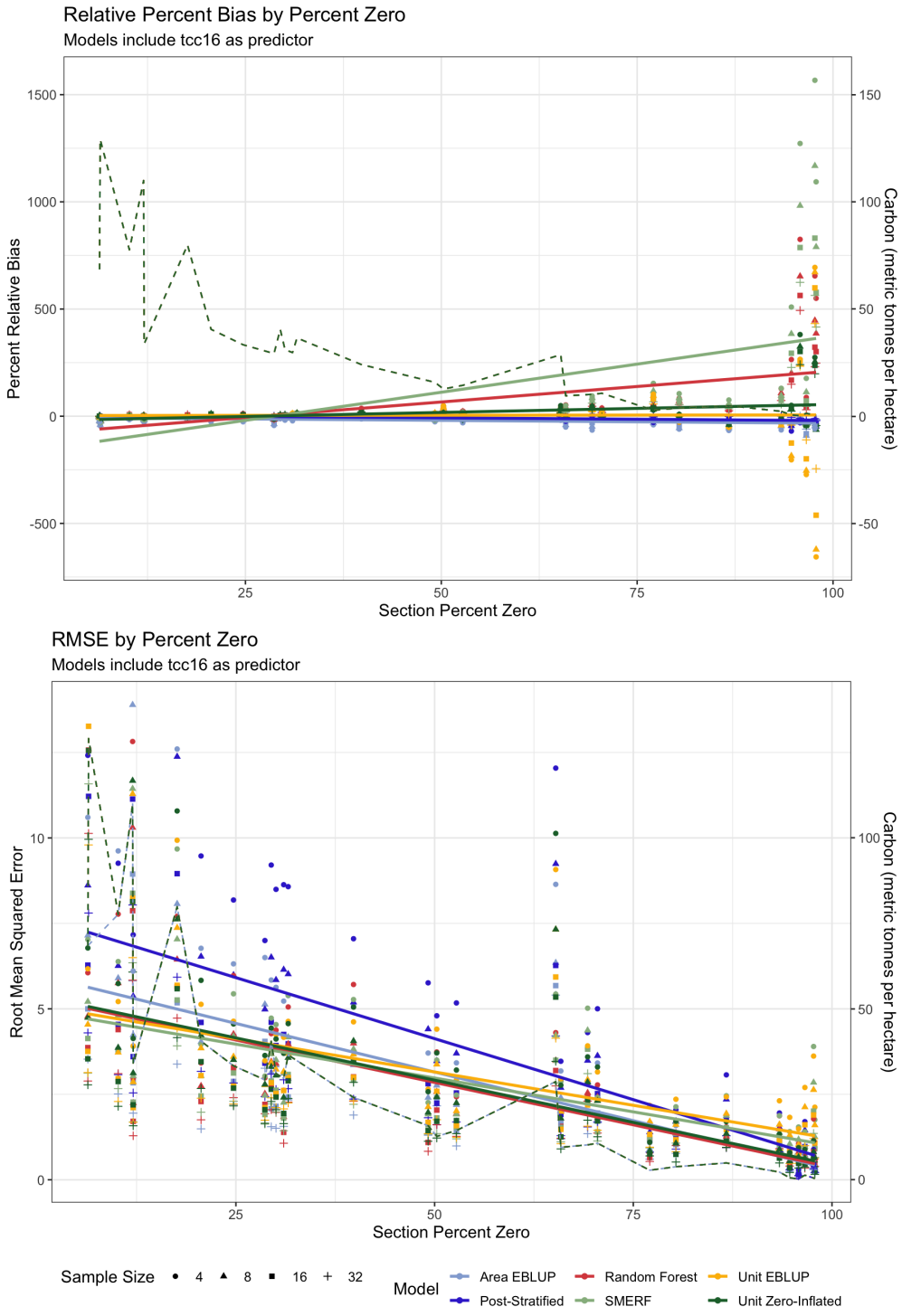


Figure 10: Percent relative bias (top) and RMSE (bottom) across all 30 sections as a function of the section percent zero-inflation. The dashed green line represents the average section carbon by section. The markers represent a particular combination of section, model, and sample size and correspond to the results presented in Figure 1. The model and sample size are indicated in the legend. The solid lines, colored by model, are linear regressions on all markers belonging to the model. Note that because some sections had particularly high bias, the plot is cropped.

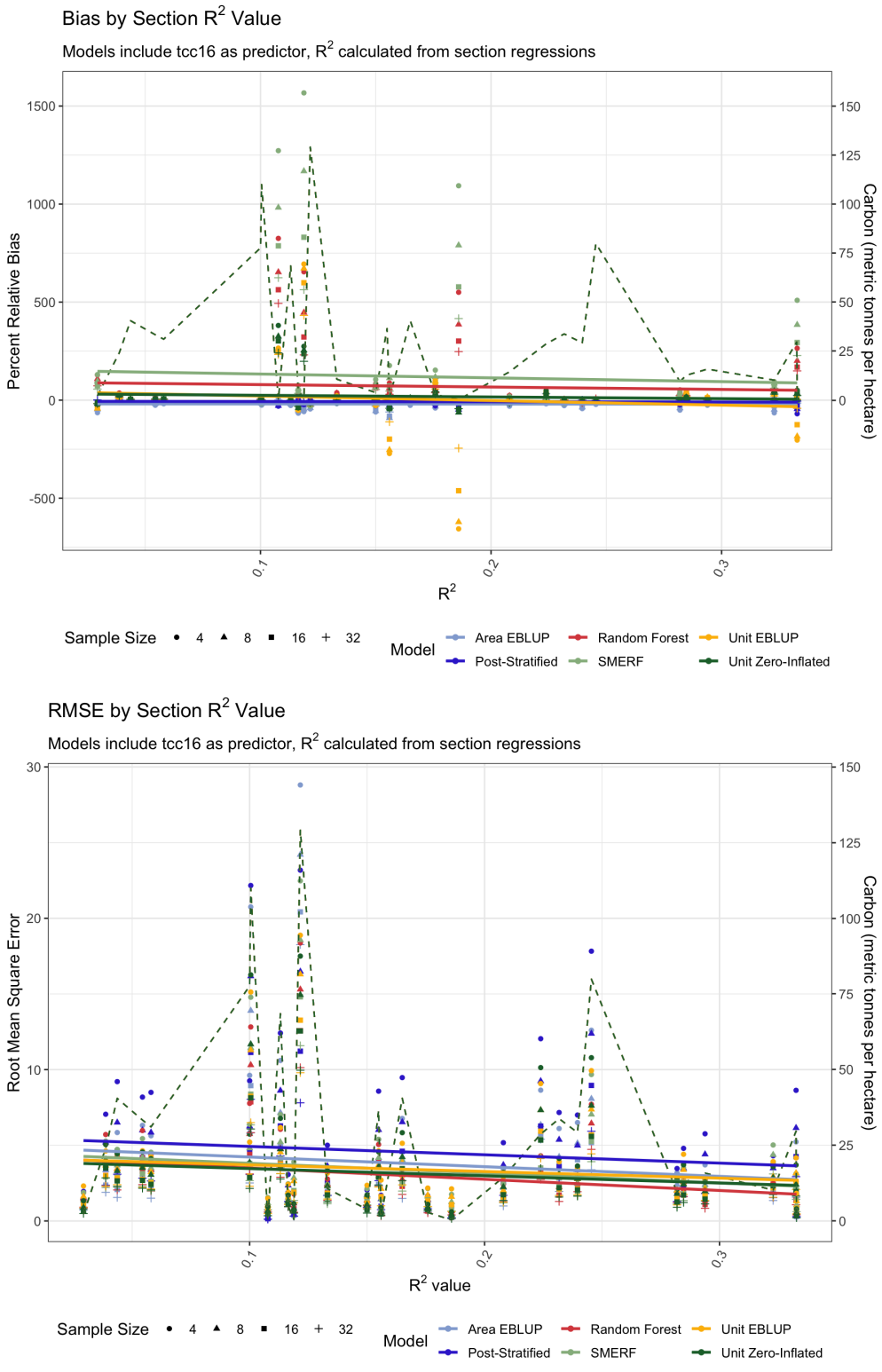


Figure 11: Percent relative bias (top) and RMSE (bottom) across all 30 studied sections ordered as a function of the R^2 correlation coefficient derived from regressing `tcc16` against carbon in each section. The markers represent a particular combination of section, model, and sample size and correspond to the results presented in Figure 1. The model and sample size are indicated in the legend. The solid lines, colored by model, are linear regressions on all markers belonging to the model.