Zero Inflation in Small Area Estimation Models: Improving Forest Inventory Estimates

Julian Schmitt*[1], Josh Yamamoto*[2], Kelly McConville[3], Kate Hu[4], Grayson White[5], and George Gaines[6]

[1]Harvard College
[2]Reed College
[3]Department of Statistics, Harvard University
[4]HSPH Department of Biostatistics, Harvard University
[5]Michigan State University
[6]United States Forest Service: Forest Inventory and Analysis Program

October 29, 2022

Abstract

The estimation of forest resources across small domains, such as the county level, is a problem well addressed using small area estimation. Resource variables of interest, such as tree volume and basal area, can only be measured with expensive in-situ ground plot observations. While collecting many measurements on the ground is expensive, remote sensing layers, which include variables like temperature, tree canopy cover, and enhanced vegetation index (EVI), are typically available for free from national satellite systems. Current estimators used to assess mean resources by domain (analogously a resource total) combine these two data streams include direct estimators like post-stratification (PS) which rely only on measurements inside a domain and indirect estimators such as the area level empirical best linear unbiased prediction (EBLUP) and and unit level EBLUP, which borrow strength across domains to generate predictions. However, in a setting where a significant portion of the plot measurements are zero applying current estimators results in model mis-specificity and poor confidence interval coverage and is particularly relevant to studying a wildfire prone region, as wildfires typically generate a lot of zeros. Using data from the US Forest Inventory and Analysis Program (FIA) from a large ecological region in the Northern US Rocky Mountains as an example, we examine how using zero-inflation small area estimators (ZI-SAE) could further improve upon current estimators. By tracking MSE, bias, and variance in a series of simulation studies across 10 domains that had between 2.5% and 49% zeros, we found the zero-inflation SAE model has lower empirical MSE than either the PS or unit or area-EBLUP across all subsections and lower relative bias than most indirect estimators, particularly when the proportion of zero-inflation is highest.

1 Introduction

1.1 Background

The United States Forestry Inventory and Analysis Program (FIA) monitors the nation's forests by collecting data on, and providing estimates for, a wide array of forest attributes. Not only is this work vitally important, but it's essential that it be done accurately and efficiently: "[The] FIA is responsible for reporting on dozens, if not hundreds, of forest attributes relating to merchantable timber and other wood products, fuels and potential fire hazard, condition of wildlife habitats, risk associated with fire, insects or disease, biomass, carbon storage, forest health, and other general characteristics of forest ecosystems." [1].

To assess forest metrics across the United States, the FIA employs a quasi-systematic sampling design to collect data at ground plots across the U.S. The FIA employs stratified sampling approach to selecting these ground plots, first partitioning the entire U.S. into 6000 acre hexagons and then randomly sampling locations from within these hexagons for measurement the FIA catalogs "plot-level" data. In combination with remote sensing data taken from satellite observation, the FIA uses these sparse ground plots to build estimates of forest attributes [2]. The remote sensing data typically includes climate metrics (e.g. temperature and precipitation), geomorphological measures (e.g. elevation and eastness), as well as metrics like tree canopy cover which can be measured from a satellite. Two common forest attributes of interest include the number of tree stems per acre and basal area, a measure of the total area per acre occupied by tree stems. When the areas are spatially large, the current estimators that the FIA employs perform well, however, there has been an increasing demand for accurate and reliable estimates of forest attributes in small areas, defined as sub-populations, typically with few observations per area. The enormity of the nation's forests in combination with the resources required to collect plot-level data means that for small-area estimation, typically only a few plot-level observations are available to build estimates. As mentioned above, collecting data at the plot-level is both labor-intensive and expensive and thus prohibits additional data collection. Instead, the FIA employs statistical methods alongside structure in the dataset to improve forest attribute estimation, referred to as Small Area Estimation (SAE). Here, we study SAE techniques to improve forest inventory estimates in the case where the data is zero-inflated.

Zero-inflated data shows up frequently in a myriad of places, from prairies, to forest fridges, to areas that have been affected by wildfires. A plot level dataset is classified as zero-inflated when the response variable follows a semi-continuous distribution, with a large number of zero observations. In the forest attribute estimation setting, we further restrict our values to be non-negative. This data is challenging to model as when the proportion of zeros is sufficiently large, normality assumption on outcome variables for regression fail, motivating the use of other modeling techniques.

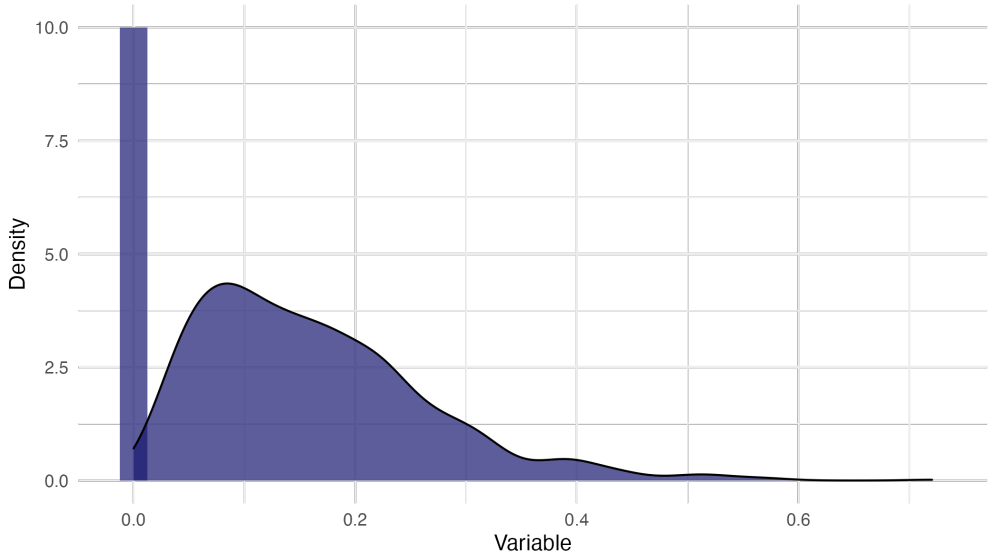

Figure 1: An example of zero-inflated data in forestry. Here, the proportion of tree canopy cover follows a non-negative continuous distribution, with a large number of zeroes.

For FIA applications, a lot of important forest attribute variables, such as basal area, exhibit a strong zero-inflation structure, see Figure 1. For example, when looking at the distribution of basal area we see very clear zero-inflation.
This begs the question of why these key FIA forest attributes are distributed in this way, and the answer is quite simple. Due to the high cost of going out and collecting ground plot data, the FIA will usually look at the remote sensed data near a plot that they are supposed to go out to, and if they have good evidence that

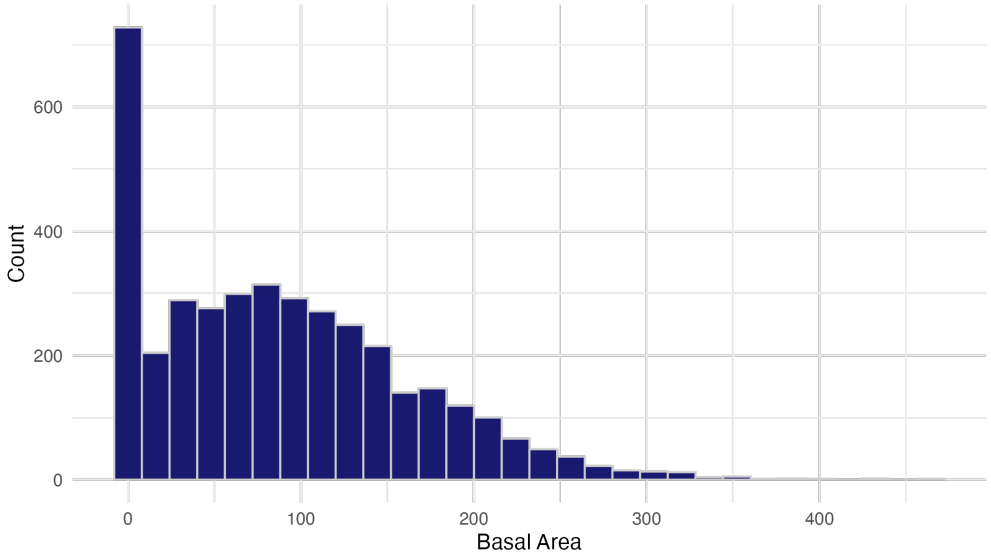

Figure 2: An example of zero-inflation in actual FIA data.

there are no trees there (i.e if it is in a parking lot) they will simply mark all of the forest attributes on that plot as zero. This is what leads to the unique data structure of zero-inflation that we observe in Figure 2.

1.2 Estimator Evaluation

A central process for this project will be to effectively compare the performance of various estimation methods. In particular we will want to assess how our estimation method performs relative to the methods that the FIA currently employs. We will want to argue that our SAE estimation method is somehow "better" than the current methods when our data is zero-inflated, and we need to introduce more technical language in order to do so.

By far the most common metric used to asses a statistical model is the Mean Squared Error (MSE). When working with an estimator $\hat{\mu}$ for a parameter μ, the MSE is calculated as

$$MSE(\hat{\mu}) = E\Big[\Big(\hat{\mu} - \mu\Big)^2\Big] \tag{1}$$

In other words it represents the the average squared deviation from the true value of the parameter value. Equation 1 can further be broken down to show that

$$MSE(\hat{\mu}) = E\Big[\Big(\hat{\mu} - E[\hat{\mu}]\Big)^2\Big] + \Big(E[\hat{\mu}] - \mu\Big)^2 \tag{2}$$

Thus minimizing the MSE boils down to simultaneously minimizing these two terms. The first is referred to as the variance and the second the squared bias. It makes sense why we would want to minimize the bias as we want our estimator to be close to the true value on average. Similarly it's intuitive why we would want our estimator to be less variable. Unfortunately, there is a tradeoff between bias and variance due to the fact that they are inversely related.

Quite often an estimators are evaluated by how small their MSE is. This is indeed an important metric and when we look at our results we will compare the MSEs of our estimators, but we will also compare the bias and the variance of our estimators on their own to get a deeper sense for how they perform. Importantly, the FIA is oftentimes most concerned with estimators that have the lowest bias, but not at any cost. We will carry these ideas with us when we assess and compare our estimators.

1.3 Application Area

Before describing our exact application area, we'll take a moment to describe the FIA's data structure. In previous sections we described the way that the FIA collects plot-level data as well as remote sensed data, but so far we have not discussed a very important way that the data is structured.

The FIA breaks the United States down into smaller areas in a hierarchical manner. The areas in each level are created with the goal of maintaining some level of ecological homogeneity within that area. Thus the prefix "Eco" is prepended to the names of these areas to stress that they are ecologically defined. At the smallest level are Eco-Subsections, which are nested inside of Eco-Sections, which are in turn nested inside of Eco-Provinces. Each plot-level data point lies within a combination of these three levels of hierarchy and it's this data structure that allows us to define how some of our estimators utilize data from outside of a given small area of interest.

$$\text{Eco-Subsections} \subset \text{Eco-Sections} \subset \text{Eco-Provinces}$$

For this particular project, we are using data from the Eco-Section M333A which is nested inside of the Eco-Province M333, a region in the Northwestern United States containing portions of Montana, Idaho, and Washington State. The larger province is characterized by forest-steppe, coniferous forest, and alpine meadow. M333A and M333 as well as the other Eco-Sections nested in M333 can be seen in Figure 3 [3]. Eco-Section M333A contains 8 Eco-Subsections and $1,204$ total plot-level data points. The Eco-Subsections within M333A are labeled by the letters a-i (excluding f, for reasons unknown to us). We were given the plot-level data that the FIA had collected for the Eco-Section, as well as the remotely sensed variables *aggregated to the eco-subsection level*. Due to the enormous size of the raw pixel-level remote sensed data files, this aggregated form is the most common format in which these files are used.

This is important because some of our estimators required pixel-level, that is for each 30×30 meter plot in M333A, data for our auxiliary variables which we were able to eventually aquire. The size of this data was part of the reason that we only used M333A instead of the whole Eco-Province. Importantly, this meant that we also needed pixel-level data for the forest attribute variables that we treated as our response.

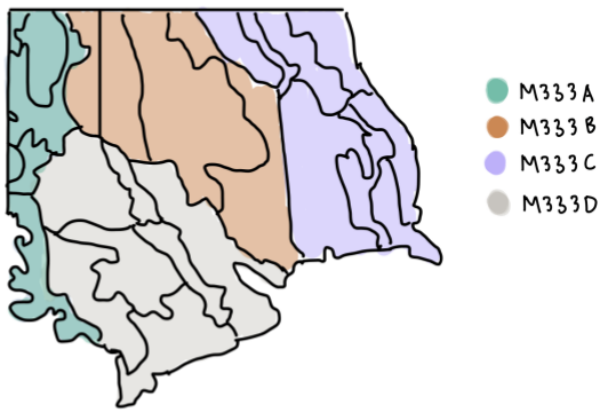

Figure 3: Eco-Province M333 is located in the Northwestern United States and is classified by Eco-Sections A-D (shaded), which can be further broken down into Eco-Subsections which are outlined.

One of the reasons that Eco-Section M333A was chosen, was because it is a well studied region which would allow our findings to be compared more easily with the findings of other SAE estimation papers on FIA data. But more importantly, the forest attributes of interest in M333A exhibit a nice range of zero-inflation across the Eco-Subsections, which are the small areas that we will be estimating these attributes on. Although no strict rules exist for what fraction of zeros is required for the data to be considered zero-inflated, we feel comfortable that the M333A basal area variable can be considered zero-inflated, see 2. This variation in percent zero is extremely valuable as it will give us additional insight into how our estimator performs across

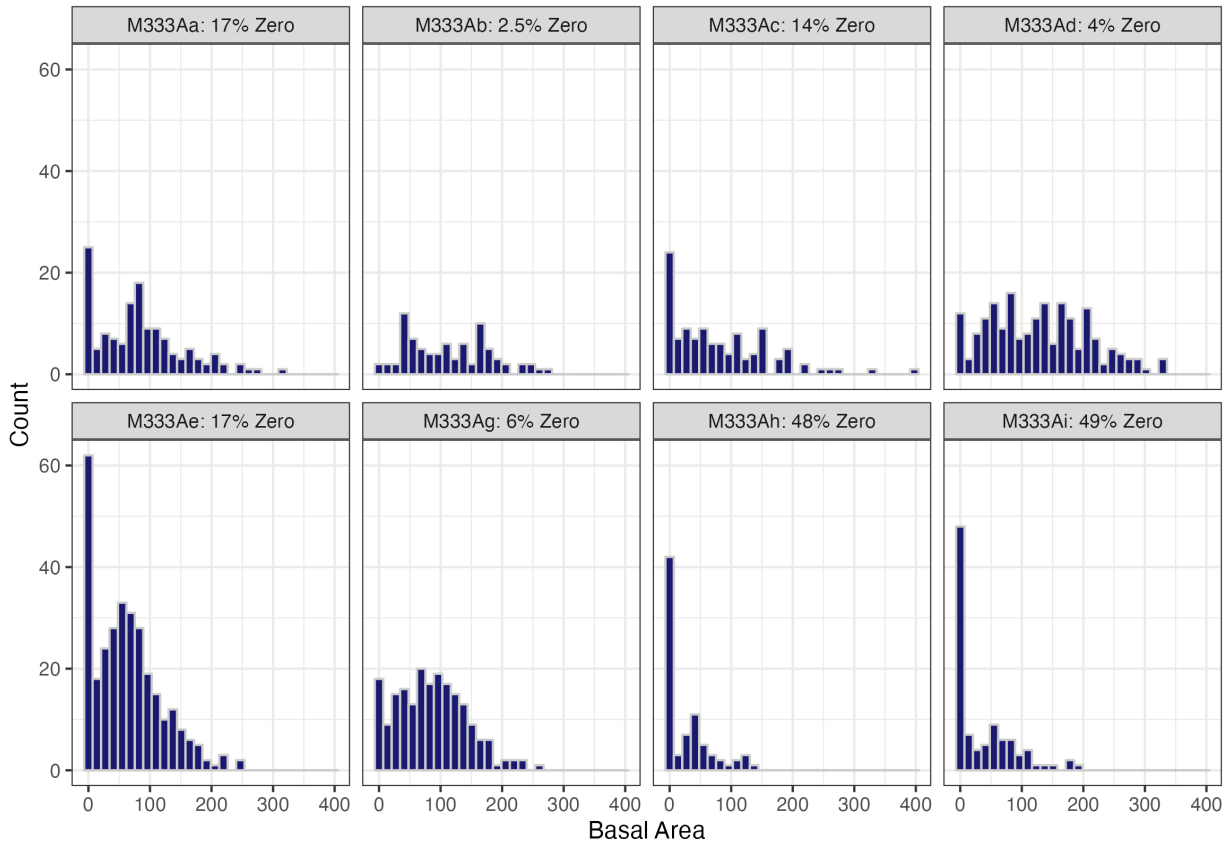

Figure 4: The distribution of the plot-level Basal Area data across the Eco-Subsections of M333A. The facets are labeled both by the Eco-Subsection name, but also by what percent of the Basal Area in that Subsection is zero.

various levels of zero-inflation. Our argument in this paper is certainly not that our estimator should be used in all cases. Rather, we want to develop an understanding for when our estimator will actually be useful, and this data structure will help us do that. Figure 4 shows this phenomenon across the forest attribute of Basal Area.

2 Methods

2.1 Notation

Let U denote a finite population with N elements, in our example U is M333A and $N = 3,792,678$ pixels. We break U into J domains U_j, $j = 1, 2, \cdots J$, these denote eco-subsections M333Aa - M333Ai ($d = 8$ domains). Each subsection U_j is defined as having N_j pixels, which range from $N_2 = 252,506$ to $N_5 = 931,693$. The sample size, where we have observed basal area, of U is denoted $n = 1204$ for U, and indexed n_j for each eco-subsection. Each unit-level observation, i in domain j has auxiliary information x_{ij} that is collected from satellites which include information like percent tree canopy cover (tcc16), enhanced vegetation index (evi), alongside measures of temperature: mean, minimum and maximum. Observations like basal area (BA) must be collected by hand and are therefore not observable at every pixel, we index these as y_{ij}. FIA's target is typically the domain means, μ_j, which is the average of the response variable across the *pixel population*. In eco-province M333, there are a total of 23 domains, called subsections, which contain between 28 and 384 plot-level observations. While the mean of a subsection with 384 observations can likely be well-approximated using direct estimates like PS, a region with only 28 observations is considered a small

area and therefore would benefit from indirect estimators that borrow strength across subsections.

We denote our subsection level estimators as $\hat{\mu}_j$, indexed $j = 1, \cdots, J$ and hope to build a model that has both low bias, variance, and mean squared error (MSE). We write $\text{Var}(\hat{\mu}_j)$ and $\text{MSE}(\hat{\mu}_j)$ for the variance and MSE of our estimated parameters, respectively.

2.2 Current Suite of Estimators

The range of estimators we hope to compare to the zero-inflation model include PS, area-EBLUP, and unit-EBLUP, which were identified as common estimators in communication with the FIA and appear frequently in recent literature [1, 4, 5]. Before fully introducing the estimators we will take a moment to describe the various methodologies of small area estimation.

Recall that a common goal for the FIA is to produce estimates at the subsection level. In other words, we are interested in generating estimates for small areas. For the purposes of this paper we will describe two main "types" of SAE estimators: direct estimators which estimate $\hat{\mu}_j$ using only data within region j and indirect estimators which use data across all subsections to estimate $\hat{\mu}_j$. Broadly:

- **Direct Estimators**: only make use of sampled data from within the small area on which we are trying to generate an estimate.
- **Indirect Estimators**: make use of sampled data from both inside and outside of the small area on which we are trying to generate an estimate. These estimators incorporate data from outside the area of interest using an explicit model. Within this type of SAE estimator there are two other nested categories: unit-level and area-level:
  - *Unit-Level*: Built on data at the unit level at which it was collected. In our case this means building the estimator on the plot-level data
  - *Area-Level*: Built on data that has been aggregated to the level of the small area on which we are generating estimates. In our case this means building the estimator using subsection means of pixel-level auxiliary variable observations.

Direct estimators have the advantage of being more easily interpreted and having low computational cost, however indirect estimators, while more complicated, can often lead to a reduction in variance as a result of making use of more of the available data and models if they are specified correctly. Based on communication with the FIA, the Post-Stratified estimator is currently the most commonly used estimator.

2.2.1 Post-Stratified

The post-stratified (PS) estimator is a weighted average of the post-stratified means, or Horovitz-Thompson (HT) estimators. The area of interest is divided, or stratified, by a single categorical auxiliary variable, in our case the tree/no tree (tnt) boolean, and the sample mean is calculated for each region. Next, the weights are generated by computing the fraction of the area which belongs to each strata which is used to weight our HT estimates. Thus the PS estimator is a weighted mean of means. Formally, let our categorical auxiliary variable have H levels indexed by $h = 1, ..., H$. Next, let $i = 1, ..., n_h$ index the sampled units in a given category h with sample size n_h. Now let N and N_h denote the total number of population units and the total number of population units in category h of our auxiliary variable, respectively. Our PS estimator for a given small area can thus be written as:

$$\hat{\mu}^{PS} = \sum_{h=1}^{H} \frac{N_h}{N} \left[\frac{1}{n_h} \sum_{i=1}^{n_h} y_i \right] = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h \tag{3}$$

Thus, we can see that the post-stratified estimator weights the strata-level means by the proportion of the population units that category h occupies within that small area and then sums them all up. Recall that since this is a *direct estimator*, this calculation is being done only using the data from within a singular small area. PS is an unbiased and consistent estimator for the population mean μ_j.

2.2.2 Unit-Level EBLUP

Now let $i = 1, ..., N_j$ denote the population units in small area j and let $j = 1, ..., J$ denote the small areas in our population. Thus our explanatory variables ($p = 1, ..., P$) for unit i in domain j are denoted by $\mathbf{x}_{ij} = (x_{ij}^1, ..., x_{ij}^P)^T$. We build a linear mixed model with random intercepts as follows

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + u_j + \varepsilon_{ij} \tag{4}$$

Here $\mathbf{x}_{ij}^T$ is a $P \times 1$ vector of covariates, $\boldsymbol{\beta}$ is a $1 \times P$ vector of fixed effects, u_j is the random effect associated with area j, and ε_{ij} is the individual level random effect for observation i in area j. We assume

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

In this case σ_u^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 $\hat{u}_j$ are estimated as laid out in Rao 2015 [6]. Of course, we are interested in getting estimates for our small areas of interest so we must now aggregate to get estimates for an individual area j:

$$\hat{Y}_j = \bar{\mathbf{X}}_j \hat{\boldsymbol{\beta}} + \hat{u}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \left[\mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + \hat{u}_j \right] \tag{5}$$

2.2.3 Area-Level EBLUP

Under the same indexing set up as in 2.2.2 we build a linear mixed model with random intercepts as follows:

$$y_j = \mathbf{x}_j^T \boldsymbol{\beta} + u_j + \varepsilon_j \tag{6}$$

This time $\mathbf{x}_j^T$ is a $P \times 1$ vector of the means for all of our explanatory variables in area j, while ε_j is the individual level random effect for area j. And again,

$$u_j \sim \mathcal{N}(0, \sigma_u^2) \qquad \text{and} \qquad \varepsilon_j \sim \mathcal{N}(0, \sigma_e^2)$$

the parameters σ_u^2 and σ_e^2 are again estimated using either method of moments or REML, and $\hat{\boldsymbol{\beta}}$ and $\hat{u}_j$ are estimated as is laid our in Rao (2015) [6]. Note that since this is an *area-level* estimator, it is built on data aggregated to the small area level. In the end we have the model

$$\hat{Y}_j = \mathbf{x}_j^T \hat{\boldsymbol{\beta}} + \hat{u}_j$$

2.3 The Unit-Level Zero-Inflation Estimator

We now turn to a description and derivation of the zero-inflation model. Following [7], where y represents the response and R represents the covariates and random effects, we have the following underlying model structure:

$$\mathbb{E}[y|R = r] = \underbrace{\mathbb{E}[y \mid R = r, y = 0]}_{= 0} \mathbb{P}(y = 0 \mid R = r) + \mathbb{E}[y \mid R = r, y > 0]\mathbb{P}(y > 0 \mid R = r) \tag{7}$$

$$= \mathbb{E}[y \mid R = r, y > 0]\mathbb{P}(y > 0 \mid R = r) \tag{8}$$

Out of equation 7 comes a wonderful intuition for what our zero-inflation models will look like. By expanding the LHS of equation 7 by conditioning on whether y is positive or not, we end up with an estimator in which we fit a model to the nonzero portion of the data and then weigh those model outputs by the probability that that point is zero. The allure of this structure is that the model for $\mathbb{E}[y \mid R = r, y > 0]$ is no longer model-misspecified when fit only to the non-zero data points when compared to the unit-EBLUP. What's more, if the model for $\mathbb{P}(y > 0 \mid R = r)$ is accurate then data points that truly have zero response would get sent to zero with a high probability by our estimator and non-zero data points would be accurately modeled

by a less biased model.

The model operates in a nested two-level form where the first level is the area and the second level is the individual plot. The estimator is built for plot i in area j:

$$\mathbb{E}[y_{ij} \mid R_{ij} = r] = \mathbb{E}[y_{ij} \mid R_{ij} = r, y_{ij} > 0]\mathbb{P}(y_{ij} > 0 \mid R_{ij} = r) \tag{9}$$

We model the two parts of the RHS of 8 using two separate mixed models.

2.3.1 Linear Mixed Model (LMM)

We model $\mathbb{E}[y_{ij} \mid R_{ij} = r, y_{ij} > 0]$ as the following linear mixed model with random intercepts fit to the nonzero portion of the sample data. The subscript nz simply specifies that we are referring to the non-zero portion of the sample data. We superscript our response with an asterisk (*) to denote the prediction is from our linear model and is not our final estimate for y_{ij}.

$$y^*_{ij,nz} = \mathbf{x}^T_{ij,nz}\boldsymbol{\beta}_{nz} + u_{j,nz} + \varepsilon_{ij,nz} \qquad \text{where} \qquad u_{j,nz} \sim \mathcal{N}(0, \sigma^2_{u,nz}) \ , \ \varepsilon_{ij,nz} \sim \mathcal{N}(0, \sigma^2_{e,nz}) \tag{10}$$

In this case $\mathbf{x}^T_{ij,nz} = (x^1_{ij,nz}, ..., x^P_{ij,nz})$ is a $P \times 1$ vector of covariates and $\boldsymbol{\beta}_{nz}$ is a $1 \times P$ vector of fixed effects. Furthermore $\sigma^2_{u,nz}$ is the between area variance parameter and $\sigma^2_{e,nz}$ is the within area variance parameter.

2.3.2 Logistic Mixed Model (GLM)

We model $\mathbb{P}(y_{ij} > 0 \mid R_{ij} = r)$ as a logistic mixed model. The asterisk (*) on u^*_j is to differentiate between the random effect in our linear mixed model. This will become important when we combine the two models in 2.3.3.

$$p_{ij} = \frac{\exp\left(\mathbf{x}^T_{ij}\boldsymbol{\gamma} + u^*_j\right)}{1 + \exp\left(\mathbf{x}^T_{ij}\boldsymbol{\gamma} + u^*_j\right)} \qquad \text{where} \qquad u^*_j \sim \mathcal{N}(0, \sigma^2_{u^*}) \tag{11}$$

where again, $\mathbf{x}^T_{ij}$ is a $P \times 1$ vector of covariates, $\boldsymbol{\gamma}$ is a $1 \times P$ vector of fixed effects, and $\sigma^2_{u^*}$ is the value of the between area variance parameter. Note that the lack of the subscript nz tells us that this model is built on the entire sample data set.

2.3.3 Combining the Models

We then combine these two estimators to create the final model estimate

$$y_{ij} = y^*_{ij}p_{ij} = \left[\mathbf{x}^T_{ij,nz}\boldsymbol{\beta}_{nz} + u_{j,nz} + \varepsilon_{ij,nz}\right] \cdot \left[\frac{\exp\left(\mathbf{x}^T_{ij}\boldsymbol{\gamma} + u^*_j\right)}{1 + \exp\left(\mathbf{x}^T_{ij}\boldsymbol{\gamma} + u^*_j\right)}\right] \tag{12}$$

For simplicity we assume that the random effects between the two parts of the model are uncorrelated, i.e

$$\text{Corr}\left(u_{j,nz}, u^*_j\right) = 0 \qquad \forall j \in 1, ...J$$

While this is very likely not a correct assumption, [7] showed that taking the correlations into account only slightly improved the accuracy of their estimates while introducing much more complexity to the model itself. For this reason we choose a simpler model at the expense of slightly more accuracy.

We employ a frequentist approach by which we estimate models (2) and (3) to get the estimates $\hat{\boldsymbol{\beta}}_{nz}, \hat{\boldsymbol{\gamma}}, \hat{u}_{j,nz}$, and $\hat{u}^*_j$. Following the structure of the R package `lmer` these parameters are estimated using restricted maximum likelihood (REML).

With these we generate $\hat{y}^*_{ij}$ and $\hat{p}_{ij}$:

$$\hat{y}^*_{ij} = \mathbf{x}^T_{ij,nz}\hat{\boldsymbol{\beta}}_{nz} + \hat{u}_{j,nz}$$
$$\hat{p}_{ij} = \frac{\exp\left(\mathbf{x}^T_{ij}\hat{\boldsymbol{\gamma}} + \hat{u}^*_j\right)}{1 + \exp\left(\mathbf{x}^T_{ij}\hat{\boldsymbol{\gamma}} + \hat{u}^*_j\right)}$$

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 $\hat{y}_{ij} = \hat{y}^*_{ij}\hat{p}_{ij}$. Then, since we are interested in area level estimates, we aggregate to the area level:

$$\hat{Y}_j = \frac{1}{N_j}\sum_{i=1}^{N_j}\hat{y}^*_{ij}\hat{p}_{ij} \tag{13}$$

where N_j is the population size of area j. Notice that since we are summing over the total number of pixel-level data points in area j, we must predict both of our models in our two part estimator on the entire pixel-level data set.

2.4 More intuition for the Zero-Inflation Estimator

For these examples and images we will use downsampled data from all across Eco-Province M333. Suppose we are building a model where we are trying to predict Basal Area using one of our remote sensed auxiliary variables: Tree Canopy Cover. If we plot these variables against each other we see the following pattern: While there is certainly a moderately strong positive linear relationship for a portion of the data, we do see

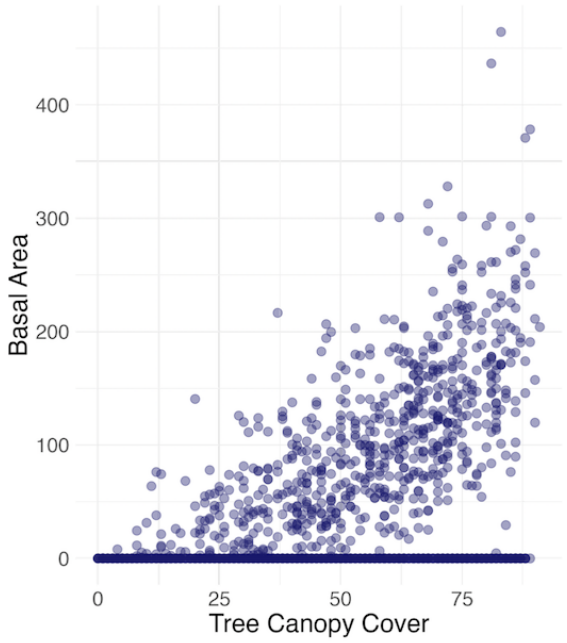

Figure 5: The relationship between our key FIA forest attribute of interest, Basal Area, and Tree Canopy Cover, a predictor known in forestry to have strong correlation.

a large portion of data points that lie directly on the x-axis due to the large proportion of zero-inflation in Basal Area. We'll now walk through how some of the estimators described in the previous section fare with this unique data structure.

2.4.1 Unit-Level EBLUP

While this is an oversimplification of what the unit-level EBLUP actually does, it does highlight the impracticality of using a unit-level that relies solely on an underlying linear model.

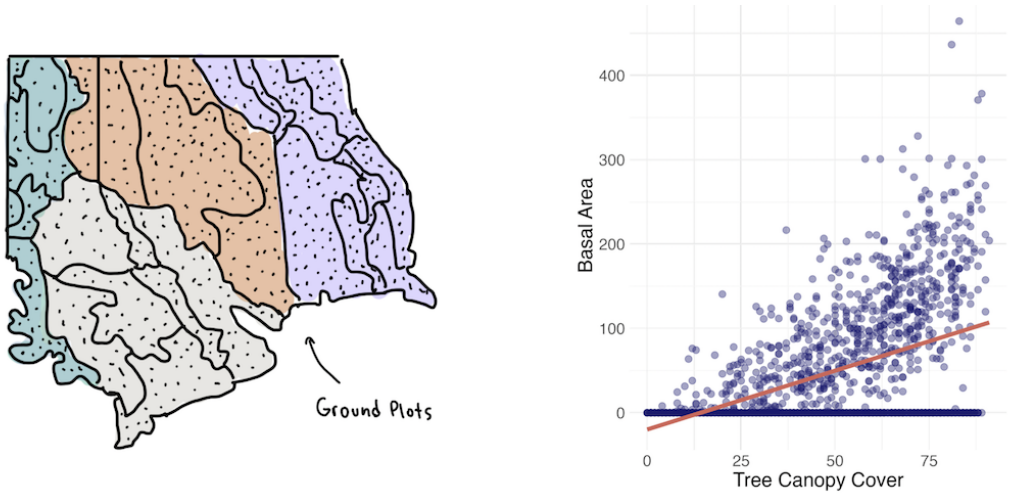

Figure 6: On the left we show an image of M333 with each dot representing an example of a ground-plot. On the right we show those ground plots plotted in the same way as in Figure 5, but with a OLS model fit to the data.

The underlying linear model of a unit-level EBLUP is clearly model misspecified in Figure 6. What's more, as Basal Area becomes more and more zero-inflated we can imagine the linear model becoming less and less appropriate for the data. One major way to get around this model misspecification is to use an area-level model.

2.4.2 Area-Level EBLUP

Again, the area-level EBLUP is more complex than simply fitting a linear model to the area level data, but this example helps give intuition for why the area-level EBLUP might be used in zero-inflation situations. When the data are aggregated to the area-level and we plot mean basal area against mean tree canopy cover we get a better model fit: While an area-level estimator with an underlying linear model does fix the

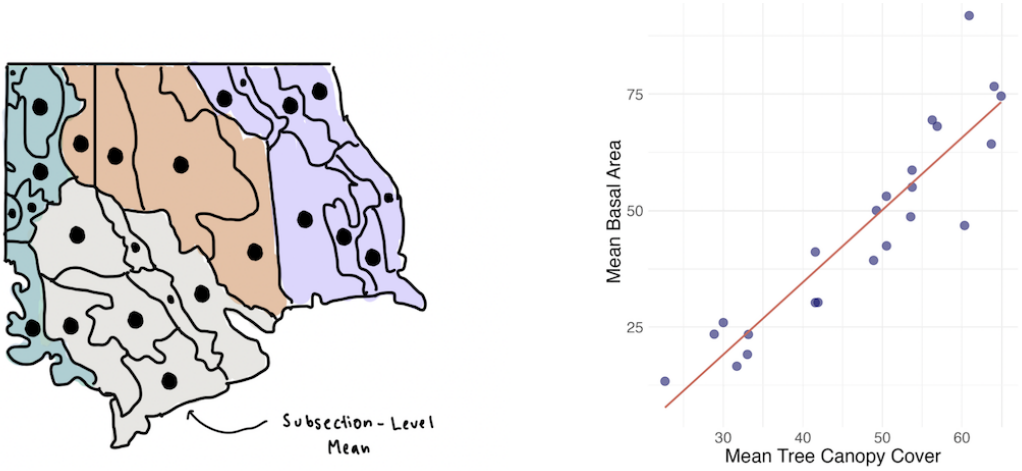

Figure 7: On the left we show an image of M333 with each dot representing an area level mean. On the right we show those means plotted, with an OLS model fit to the data.

model-misspecification problem, we are making use of so much less data in this approach. It's this loss of data that often makes area-level estimators unmanageably variable.

2.4.3 Unit-Level Zero-Inflation Estimator

Finally, we'll show visually how the zero-inflation estimator tackles modeling this data structure. As a reminder, a linear model is fit to the non-zero portion of the sample data. Next a logistic regression model is fit to the full sample data. We get our final predictions for a given plot in a given small area by multiplying the output of the linear model by the output of the logistic regression model. By only fitting a linear model to the non-zero portion of the data we end up with a model that is more specified to the data structure. We can gain intuition for the combination of the two models by imagining that we are weighting our linear regression outputs by how likely our logistic regression model thinks that that plot has zero basal area.

2.5 Bootstrapping MSE

In order to generate mean squared error estimates we follow the bootstrapping technique laid out in Chandra and Sud (2012) to estimate the MSE for the zero-inflated SAE estimator [8]. We run the following bootstrapping technique on the full suite of sample data sets:

1. Fit the zero-inflation model to the original data set and extract the parameters from the two part model. This provides estimates of the true population parameters:
   (a) $\hat{\boldsymbol{\beta}}$, the fixed effects from the linear mixed model.
   (b) $\hat{\sigma}_u^2$, the variance of the area-specific random errors, and $\hat{\sigma}_e^2$, the variance of the individual-level random errors.

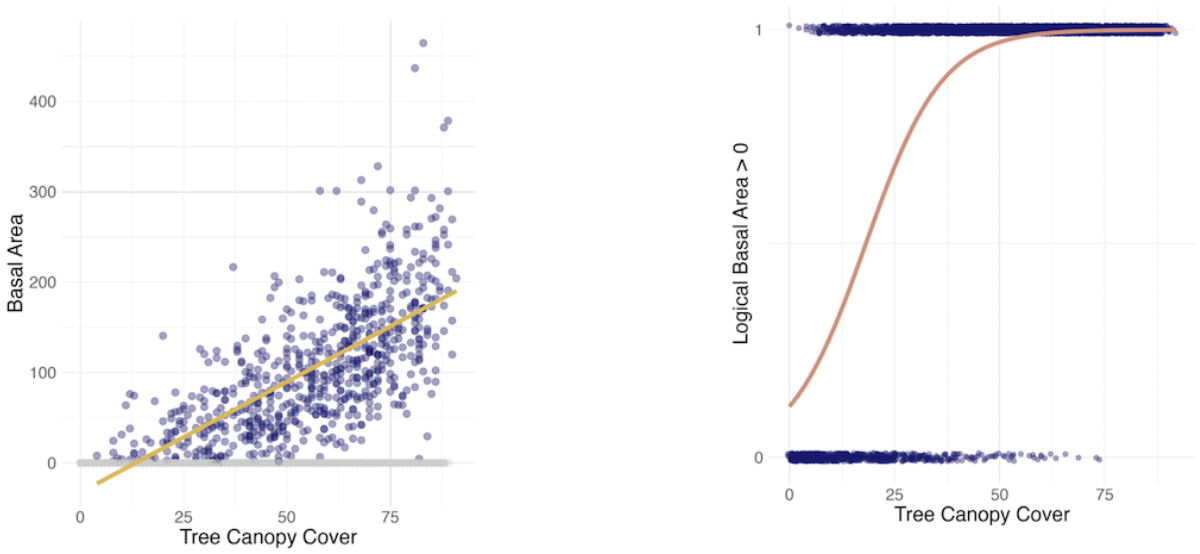

Figure 8: On the left we show a linear model fit to the non-zero portion of the sample data. On the right we show a logistic regression model fit to the full sample data set.

(c) $\hat{\boldsymbol{\gamma}}$, the fixed effects from the logistic mixed model.

(d) $\hat{u}_j^*$, the random effects for each eco-section in the logistic model.

Using these parameters, we generate the population bootstrap data set by first computing the probability that each point is non-zero using,

$$\hat{\boldsymbol{p}}^*_{\boldsymbol{ij}} = \frac{\exp(\boldsymbol{x_{ij}}\hat{\boldsymbol{\gamma}} + \hat{u}_j^*)}{1 + \exp(\boldsymbol{x_{ij}}\hat{\boldsymbol{\gamma}})} \tag{14}$$

then, treating these as bernoulli trials, we generate indicator random variables δ_{ij}^* which will generate zeros in the bootstrap data. Now, we generate area-specific random errors for each domain U_j for $j = 1, 2, \cdots, d$ and individual random errors for each pixel in U. Intuitively, these are fit from a normal distribution as we assumed they were drawn from one. Thus our random errors become:

$$u_i^* \sim \mathcal{N}(0, \hat{\sigma}_u^2) \qquad\qquad e_{ij}^* \sim \mathcal{N}(0, \hat{\sigma}_e^2) \tag{15}$$

We generate our bootstrap y_{ij} data from the pixel-level x_{ij} data, as follows:

$$y_{ij}^* = \left(\boldsymbol{x_{ij}}\boldsymbol{\beta} + u_i^* + e_{ij}^*\right)\delta_{ij}^* \;\; j = 1, 2, \cdots, d, \;\; i = 1, 2 \cdots N_j \tag{16}$$

Now that we have generated y_{ij}^* for each pixel in the dataset we can now compute the population parameter of interest. For our study the population parameter of interest, $\boldsymbol{\theta}$ the domain mean for each subsection which we can now compute via the sample mean:

$$\boldsymbol{\theta} = \begin{bmatrix}\theta_1 & \theta_2 & \cdots & \theta_d\end{bmatrix} \tag{17}$$

$$\theta_j = \frac{1}{N_j}\sum_{i=1}^{N_j} y_{ij}^* \text{ for } j = 1, \cdots, d \tag{18}$$

$\boldsymbol{\theta}$ becomes our true population parameter, even though it has been generated artificially. We now can calculate bias and empirical MSE using this as our true value.

2. Generate B bootstrap samples from the bootstrap population data set of equal size to the original sample and fit the zero-inflation model to the each one, extracting the model estimates for each one $\hat{\boldsymbol{\theta}}^{(b)}$.

3. Compute an estimate for the MSE by taking $\frac{1}{B}\sum_{b=1}^{B}(\hat{\boldsymbol{\theta}}^{(b)} - \theta)^2$.

2.6 R Function Code

The first working version of an R function that will fit this a zero-inflation model to a given dataset has been written. It uses `lme4:lmer()` and `lme4::glmer()` to fit the linear mixed model and the logistic mixed model respectively. Both of these R functions default to using restricted maximum likelihood to find the parameter estimates, and we used this default throughout this project. Additionally if the user wants MSE estimates the function will perform the bootstrap MSE estimation procedure as described in Section 2.5.

In order to evaluate this zero-inflation model we want to be able to meaningfully compare it to the other SAE models that the FIA employs. We do this through the use of a simulation study using our function code and all of the estimators described in the methods.

2.7 Simulation Study

To effectively compare the performance of the zero-inflation estimator relative to the estimators that the FIA most commonly uses, we designed a simulation study to compare the Post-Stratified Estimator, the Unit-Level EBLUP, the Area-Level EBULP, and the Zero-Inflation Estimator. Following the work of Morris, White, and Crowther (2019) we designed our simulation study as follows [9].

2.7.1 Aims

- To understand under what conditions the zero-inflation estimator outperforms other estimators currently in use by the FIA for small area estimation.
- Assess the performance, as measured by empirical percent relative bias and empirical mean squared error, of zero-inflation small area estimation models (ZI SAE) in comparison to the unit-level EBLUP, area-level EBLUP, and Post-Stratified estimators.
- Generate heuristics for when the FIA should be concerned about model mis-specificity caused by zero-inflated data, when performing small area estimation.

2.7.2 Data Generating Mechanisms

Pixel-Level Data Generation
We chose to use basal area as our response variable because it had sufficient and variable levels of zero-inflation. However, as basal area is collected at the plot-level, pixel-level data for them simply does not exist. As pixel-level observations are required to assess estimator bias in the simulation study, we used a K-Nearest Neighbors (KNN) imputation process to generate our full pixel level data. We chose a non-parametric model to do the imputation as assuming additional underlying structure would be likely to advantage certain models. For example, if we used linear regression on another auxiliary variable to impute basal area, we would favor models that also carried linear assumptions, including area and unit EBLUP. As we decided to use Basal Area as the forest attribute of interest for all of our analyses and so we only imputed this variable.

Using the R function `knn()` from the package `class` we trained our knn model setting $k = 2$ on the plot-level data set using the following predictors: Enhanced Vegetation Index (evi), Tree Canopy Cover (tcc), and Mean Temperature (tmean) where the classification for each point was the value of Basal Area. This model was then applied to the pixel-level data set that we had already been given to give us pixel-level data for Basal Area. This unconventional use of a KNN model resulted in imputed data that was quite lumped together because we used $1,204$ data points to generate data for over 3.7 million pixels. To fix this issue, we added random noise to the Basal Area value associated with each pixel from a uniform distribution over the interval $(-3, 3)$ to only pixels that were non-zero.

Sample Data Generation
In order to compare our estimators across a large number of trials we had to come up with a process for generating a large number of sample data sets on which we would test the models.

We followed a semi-systematic random sampling design that mimicks how the FIA samples ground plots. Recall that the FIA overlays a hexagonal grid over the U.S. and then randomly samples locations from within those hexagons to determine where actual field measurements will be taken. Thus every plot-level data point is associated with a unique hexagon. Because of this, our each of our imputed pixel-level data points are associated with a unique hexagon as well. The sampling design was thus as follows:

1. Sample k different hexagons from the total number of unique hexagons.
2. Sample 1 pixel from each of the k chosen hexagons,– this results in k total data points that we will treat as a plot-level data point for that simulation run.
3. Rinse and repeat.

We ran through this process 2000 times for each value of k that we used. This resulted in 2000 different sample "plot-level" data sets for each run the simulation. As the hexagons can contains portions of multiple eco-subsections this process does not results in exactly $k/8$ (there are 8 eco-subsections in M333A), sampled plots per eco-subsections but it was generally close, especially for larger values of k.

2.7.3 Dials

To get a more robust sense of when the zero-inflation estimator outperforms the area EBLUP, unit EBLUP, and PS estimator, we tune the total number of samples used to fit the model.

- Sample Size (30, 50, 100): The main dial that we wanted to turn was the sample size of each eco-subsections. As mentioned previously the FIA is especially interested in developing estimators that can perform well on very small sample sizes and so we wanted to compare the performance of our zero-inflation estimator relative to the other estimators across different approximate sample sizes.
  - We used $k = 240, 400,$ and 800 in order to get subsections with sample sizes that were approximately $30, 50,$ and 100 respectively.
  - This resulted in 3 different simulation datasets of 2000 samples where the difference between the 3 sets was the approximate sample size in each subsection.

2.7.4 Estimands

- μ_j^{BA}: The mean of our response variable (Basal Area) in eco-subsection j for a given estimator.

2.7.5 Performance Metrics

We use the following metrics to evaluate our models over the $S = 2000$ samples for each ecosubsection, $j = 1, \cdots, d$, and for each of the 3 simulation runs, $k = 30, 50, 100$:

- Empirical Percent Relative Bias: $\left[\frac{\hat{\mu}_j^k - \mu_j}{\mu_j}\right] \cdot 100, \quad \text{where } \hat{\mu}_j^k = \left(S^{-1} \sum_{s \in S} \hat{\mu}_j^{(s)}\right)$
- Empirical Variance: $\hat{\text{Var}}(\mu_j^{(BA)k}) = \frac{1}{S-1}\left[\sum_{s \in S} (\hat{\mu^k}_j^{(s)} - \mu_j])^2\right]$
- Empirical Mean Squared Error: $\text{MSE}(\hat{\mu}_j^{(BA)k}) = \hat{\text{Var}}(\hat{\mu^k}_j) + [\hat{\mu^k}_j]^2$

2.7.6 Computation

As the zero-inflation model required that our estimators be predicted on the entire pixel-level data set, we utilized the Harvard Cluster to help with the huge amount of computation. Due to time constraints we were unable to run the MSE estimation procedure as described in [8].

3 Results

3.1 Simulation Data Generation

,

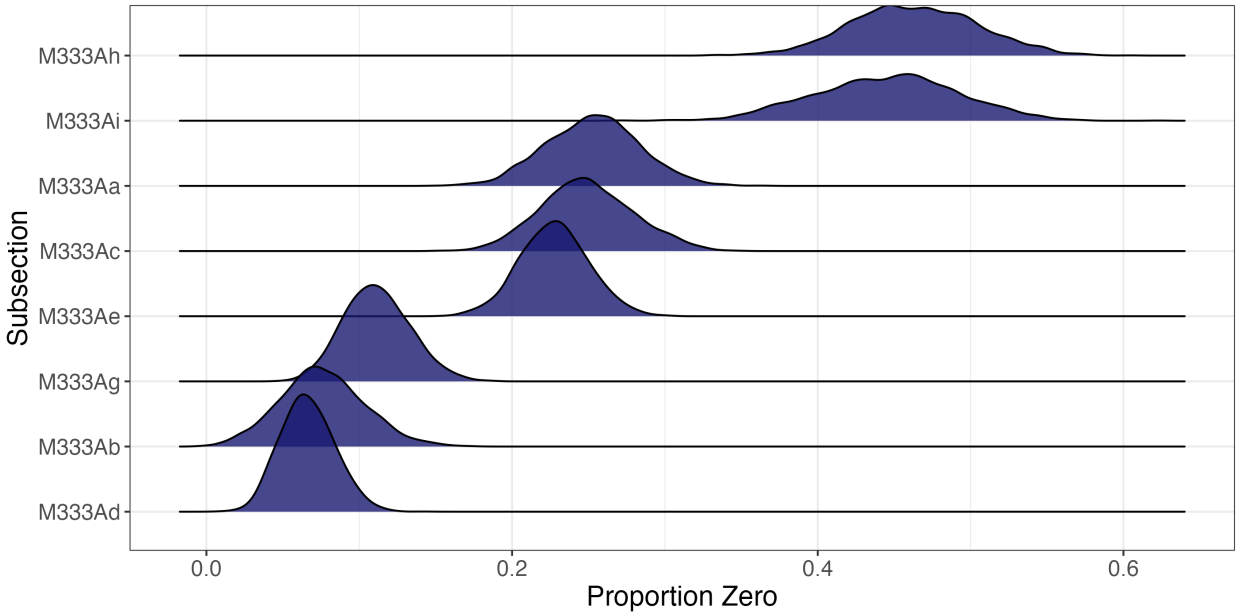

Figure 9: Distribution of zero-inflation of the generated simulation study samples for each Eco-Subsection ordered by zero proportions.

We begin by showing the distribution of zero-inflated in eco-subsection. For each sample that we generated for our simulation study, we computed the proportion of Basal Area that was zero for each subsection and plotted in Figure 9. Understanding how the proportion of zero inflation relates to estimator performance is essential for our simulation study aim of determining when to recommend the zero-inflated model. The main results plots will be ordered from least zero-inflated to most zero-inflated which was determined by looking at the means of the distributions in Figure 9.

3.2 Evaluating Model Bias

Figure 10 shows the Empirical Percent Relative Bias across all 3 simulation studies and 8 eco-subsections for each model. Across the 24 simulation size and eco-subsection model comparisons we found that the PS estimator performed best, having the lowest bias 15 times, the zero-inflation model had the lowest bias 8 times, the area-EBLUP once, and the unit-EBLUP never performed the best. We expected the PS estimator to perform particularly well because it is asymptotically unbiased and our sample sizes were relatively large. Comfortingly, we see that amongst the other models, the zero-inflation model is a clear second, outperforming both the area and unit EBLUP models. While the PS estimator had the least bias the most times, when computing the average bias across all 24 trials we found that the PS and zero-inflation models tie at an average absolute bias of 1.82%, see Table 3.2. This indicates the PS estimator often has significant bias, see Figure 10, making the zero-inflation model seem like a safer choice for the FIA. For example, in subsection M333Ah, the unit zero-inflation model has roughly half the bias of the PS estimator, which is nearly 10% off on average across the 2000 trials.

Model	Lowest Abs. Bias Count	Avg. Abs. Bias	Lowest MSE Count	Avg. MSE
Area EBLUP	1	5.32 %	0	66.1
PS	15	1.82 %	0	101.0
Unit EBLUP	0	4.09 %	0	25.8
ZI-SAE	8	1.82 %	24	21.4

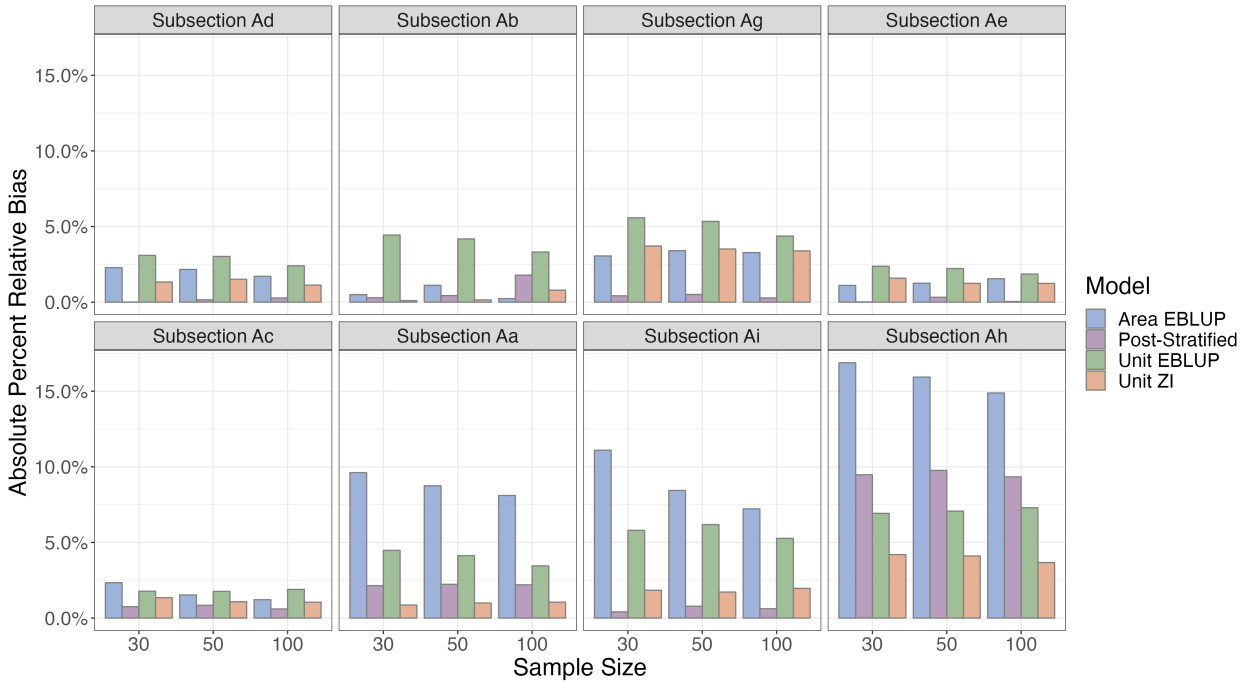

Figure 10: Bias results for the Simulation Study across 3 sample sizes, and 4 estimators: Post-Stratified, Area-EBLUP, Unit-EBLUP, and Zero-Inflated Model. Each facet represents a different eco-subsection in M333A, with the x-axis of each facet differentiating between the sample sizes. We plot the absolute value of the biases. Plots are ordered left to right top to bottom with increasing zero-inflation as shown in Figure 9

3.2.1 Estimating Model affect on Bias by Percent Zero-Inflation

To provide an estimate for when the FIA should consider the zero-inflation model as a function of bias, a natural starting point would be to study how the winning model changes as a function of the percent zero-inflation. Combining the average percent zero-inflation results from Figure 9 for each subsection with the absolute percent relative bias estimates from the simulation, we display the bias as a function of percent zeros in the data stratified by model type, see Figure 11. We summarize the regression results in Table 3.2.1. Based on bias alone we recommend that the FIA should consider using zero-inflation model when the proportion of zero inflation exceeds 25%, as this is the threshold above which the zero inflation model has the lowest bias.

Model	Intercept	Slope
Area EBLUP	-1.93	32.77
Post-Stratified	-1.41	13.04
Unit EBLUP	2.99	5.53
ZI-SAE	0.955	3.86

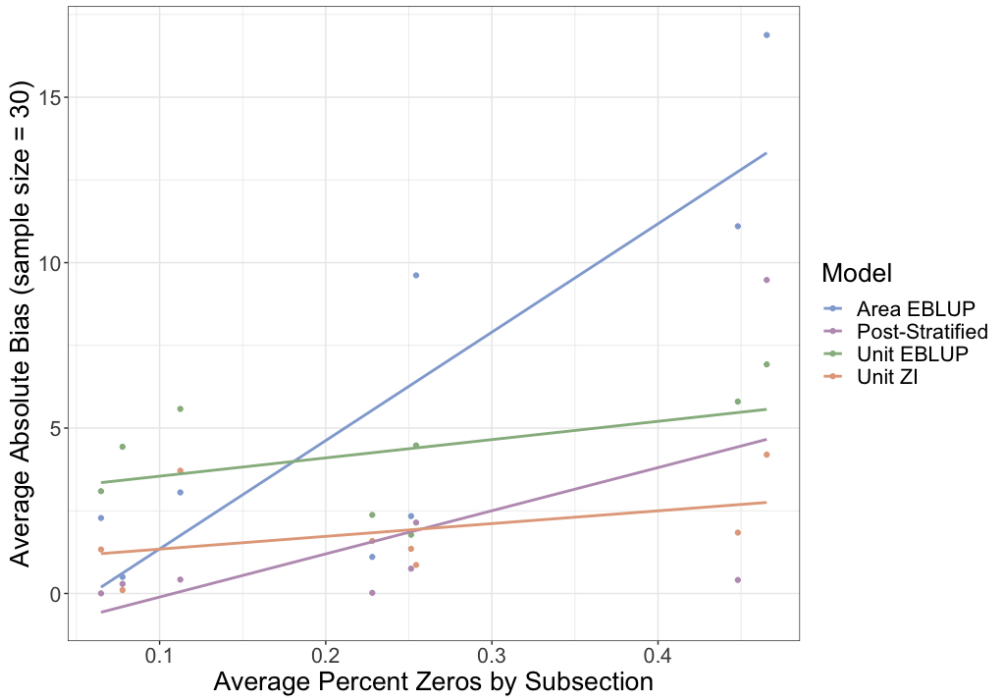

Figure 11: Average Absolute Bias as a function of subsection percent zero-inflation. We stratify by model and fit a regression to the corresponding points, shaded in the same color. Changing the sample size to 50 or 100 observations per subsection had little effect on regression coefficients and did not affect conclusions.

3.3 Empirical MSE Results

Figure 12 shows the empirical MSE results across all 3 simulation studies and 8 eco-subsections for each model. Again comparing across the 24 simulation and eco-subsection comparisons, we find that the zero-inflation model has the lowest MSE across all comparisons and has the lowest average empirical MSE, see Table 3.2. When looking solely at the bias results in Figure 10, we could conclude that the PS estimator could be a strong candidate, however, we find that the MSE is on average 101 compared to the zero-inflation model with 21.4. The empirical MSE results indicate a strong preference for unit zero-inflation models at this level of zero-inflation in the data.

4 Discussion

In this study we found that the zero-inflation model has potential to improve forestry estimates by significantly reducing the empirical MSE without increasing model bias. The analysis of empirical bias as a function of percent zero inflation revealed that at 25 % zero-inflation and above, the zero-inflation model had less bias than the unit and area EBLUP. While subsection Ai demonstrated that the PS estimator could still have low bias at high levels of zero-inflation, however the corresponding large empirical MSE values are prohibitive. We conclude that the FIA should consider employing the zero-inflation model over post-stratification when the percent of zeros in the data exceeds this 25% threshold. However, this threshold should be lowered when we begin to factor in empirical MSE results as the PS estimator has a standard error over 2 times larger than the zero-inflation model, which could revise our estimate to data having 10% zeros, per the intersection of the area-EBLUP and unit ZI models. Considering more models, additional eco-Provinces and subsections, and simulations will help tune this threshold.

Previous literature and R-package development has focused on encouraging FIA to move away from using only the PS estimator by demonstrating the robustness of area and unit EBLUP models among others [1,

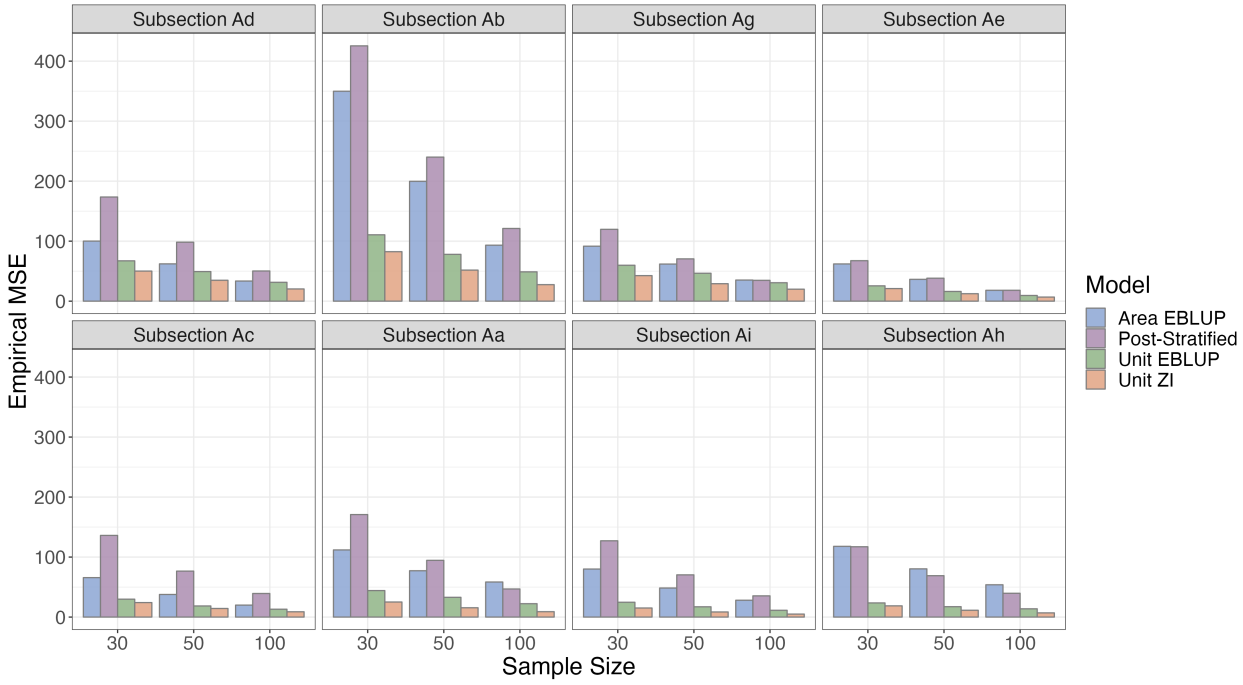

Figure 12: Empirical MSE Results for the Simulation Study across 3 sample sizes, and 4 estimators: Post-Stratified, Area-EBLUP, Unit-EBLUP, and Zero-Inflation Estimator. Each facet shows the results for one of the 8 eco-subsections in M333A. Along the x-axis we have the three approximate sample sizes for the three different simulation run.

5, 10]. In this study, we showed that there was a significant improvement to using the ZI model over these other 3 models in a region where there was a significant amount of zeros present in the data. Our findings exemplify how sophisticated problem-specific models can yield significant performance improvements. While our study was limited in scope to M333A and relied on imputed data for the simulation, the limited tuning we performed suggests that the strength of the zero-inflation model likely persists across a much wider range of scenarios. Furthermore, while the regression was performed on limited data, these results were consistent across sample sizes and are consistent with our theoretical understanding of the models. We hope that future work will verify these findings, particularly across different levels of zero inflation and US regions. We also hope to continue to add expressivitity in our models, perhaps random forests, to fit the complex relationships between forest growth and predictors. Adding a spatial correlation component that can capture regional variability could further improve these estimates would also be a valuable next step. While model performance is important, models must be straightforward to implement, tune, and understand in order to be adopted by the broader community.

Ultimately, the ZI model outperformed the PS, area EBLUP, and unit EBLUP models in our simulation study. Using models which improve empirical bias and MSE in forestry applications can have wide ranging implications for land management and forestry operations, and are especially important at a time when the signal-to-noise ratio of climate change impacts are particularly small [11]. Narrower confidence intervals may allow the FIA to disentangle the climate change signal from background variability, promoting early action which will better preserve US forests for generations to come.

References

1. 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).
2. *The Enhanced Forest Inventory and Analysis Program–national Sampling Design and Estimation Procedures* en (eds Bechtold, W. & Patterson, P.) Google-Books-ID: tryLXDR2c6cC (USDA Forest Service, Southern Research Station, 2005).
3. McNab, W. *et al. Description of ecological subregions: sections of the conterminous United States* en. Tech. rep. WO-GTR-76B (U.S. Department of Agriculture, Forest Service, Washington, DC, 2007), WO–GTR–76B. https://www.fs.usda.gov/treesearch/pubs/48669 (2022).
4. Goerndt, M. E., Monleon, V. J. & Temesgen, H. A comparison of small-area estimation techniques to estimate selected stand attributes using LiDAR-derived auxiliary variables. *Canadian Journal of Forest Research* **41,** 1189–1201. https://doi.org/10.1139%5C%2Fx11-033 (June 2011).
5. Frescino, T. S., Patterson, P. L., Moisen, G. G., Toney, C. & Freeman, E. A. *FIESTA: A Forest Inventory Estimation and Analysis R Package* tracey.frescino@usda.gov. USDA Forest Service, Rocky Mountain Research Station (507 25th street, Ogden, UT, USA, 2009).
6. Rao, J. & Molina, I. *Small Area Estimation* 2nd. ISBN: 978-1-118-73578-7 (Wiley, 2015).
7. 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).
8. Chandra, H. & Sud, U. C. Small Area Estimation for Zero-Inflated Data. *Communications in Statistics - Simulation and Computation* **41,** 632–643 (2012).
9. Morris, T., White, I. & Crowther, M. Using Simulation Studies to Evaluate Statistical Methods. *Statistics in Medicine* **38** (2019).
10. Stanke, H., Finley, A. O., Weed, A. S., Walters, B. F. & Domke, G. M. rFIA: An R package for estimation of forest attributes with the US Forest Inventory and Analysis database. *Environmental Modelling and Software* **127,** 104664. https://www.sciencedirect.com/science/article/abs/pii/S1364815219311089?via%5C%3Dihub (2020).
11. Pörtner, H.-O. *et al.* IPCC, 2022: Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. *Cambridge University Press* (2022).