Small area estimation of zero-inflated forest biomass

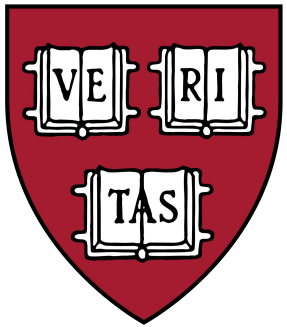
Josh Yamamoto

PSAE Meeting 2024 Tucson, Arizona April 4th 2024

Undergraduate Forestry Data Science Research Group

A USFS funded collaboration with Dr. Kelly McConville:

- 10 week program devoted to working on SAE oriented forestry research projects.
- Direct interaction with FIA scientists.
- Producing technical reports and R packages.



Zero-Inflation

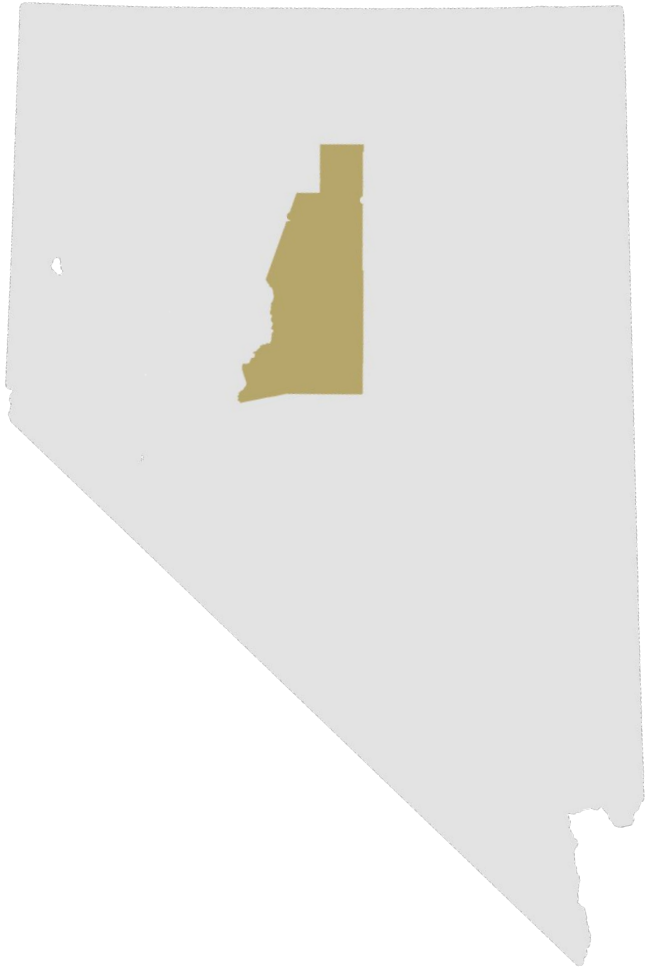
(in forest attributes)

Sometimes key forest attributes are zero on a large number of plots for an area of interest.

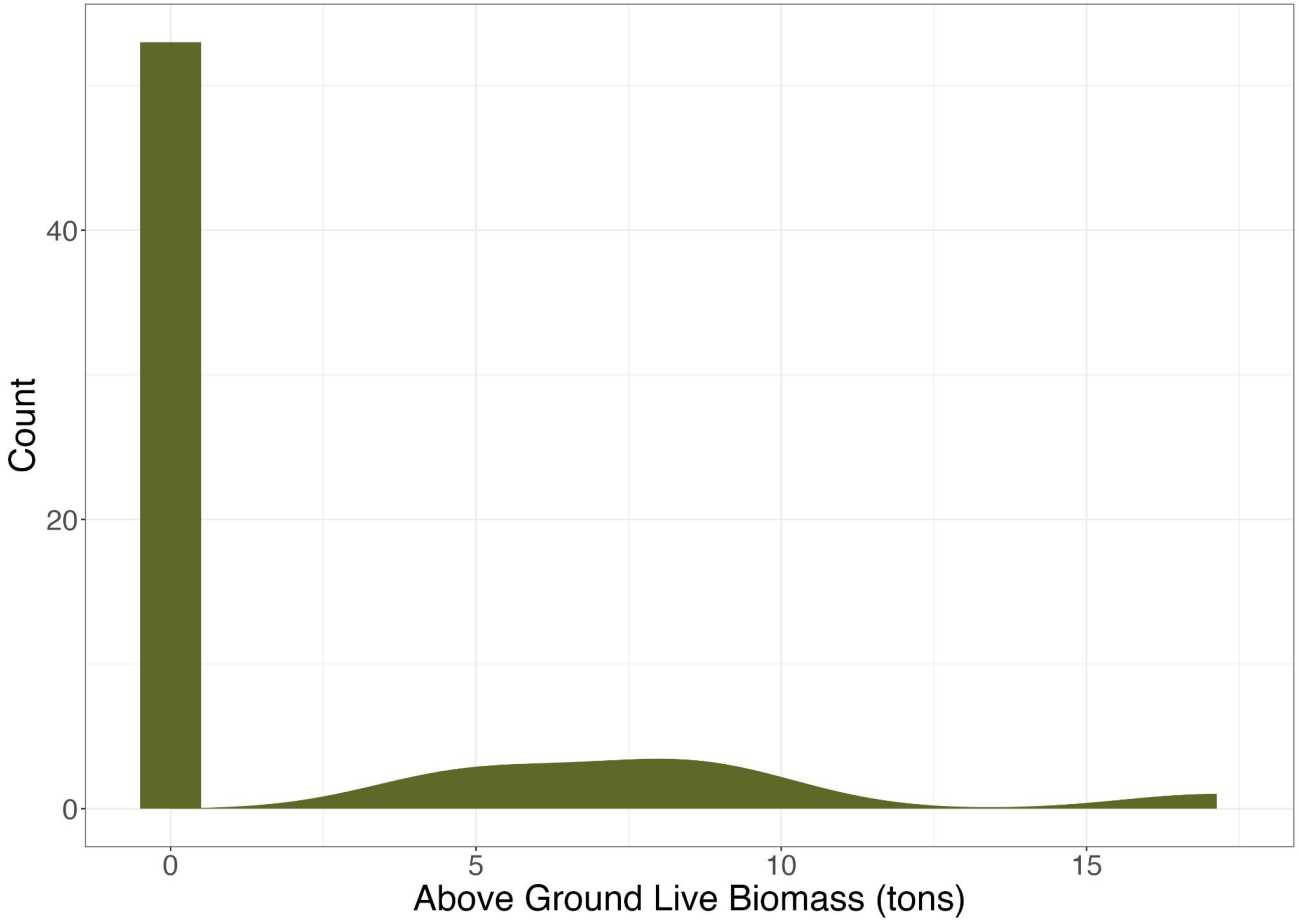
Usually when an area of interest has little forest area.

Nevada sky islands

Example:
Lander County, NV



Distribution of **biomass** from 2019 Lander County FIA plots



Common Estimators

Post-Stratified

- Uses a single categorical auxiliary variable
- Weighted average of post-stratified means

Area-level linear mixed model

- Linear model fit to domain-level means with domain-level random effects

Unit-level linear mixed model

- Linear model fit to plot-level data with domain-level random effects
- Aggregated to get domain-level estimates

Issues with existing estimators

Post-Stratified

- Already in a setting where we worry about the variance of direct design based estimators.

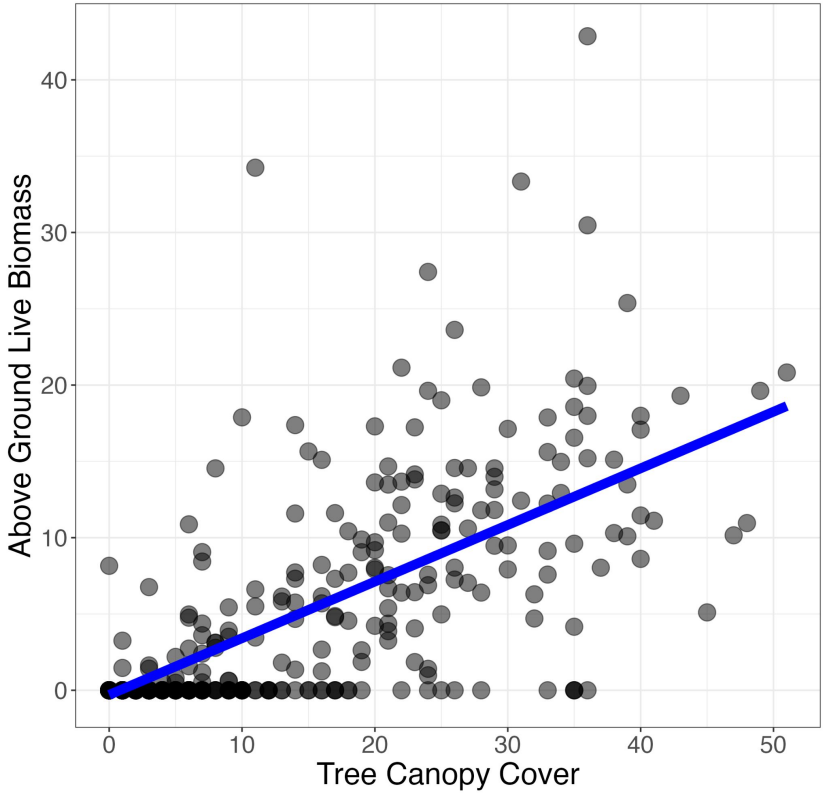
Area-Level Linear Mixed Model

- Usually good model fit, but doesn’t leverage all of the rich auxiliary data.

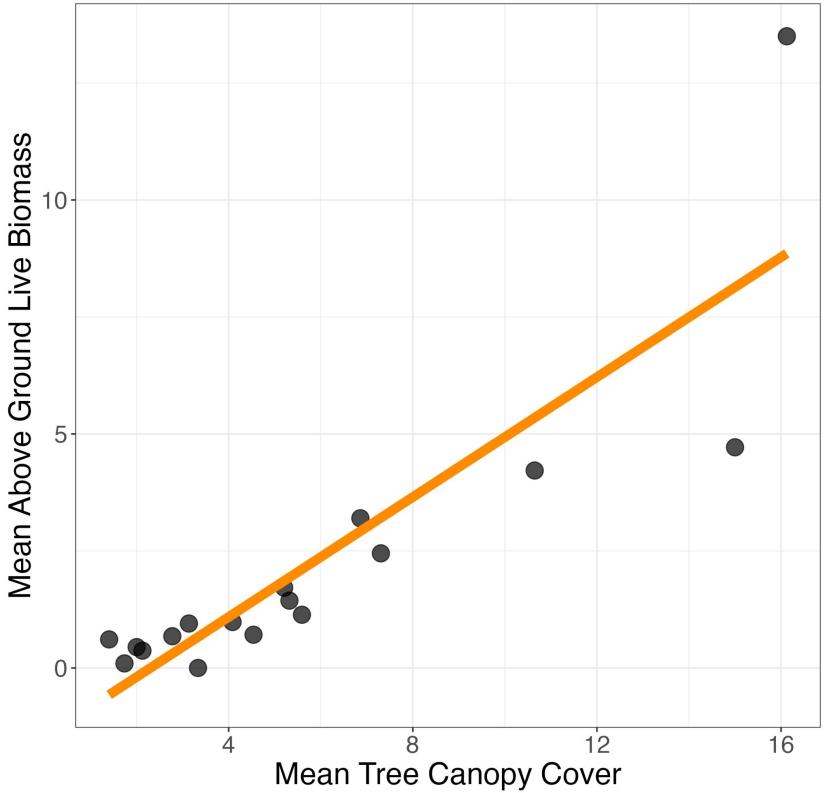
Unit-Level Linear Mixed Model

- Model misspecification

Unit-level linear model



Area-level linear model



Zero-Inflation Estimator

Two-Part Model (intuition)

1. Linear mixed model fit to the *non-zero* plots. Predicting biomass.
2. Generalized linear mixed model with binomial response fit to all of the plots. Predicting P(biomass != 0)

Each model is applied to a new individual unit and the results are multiplied to get a prediction.

The prediction from the first model is “weighted” by how likely the second model thinks that the unit is non-zero.

Notation

U represents our finite population with **N** pixels

i: indexes an individual pixel

j: indexes a domain

x_ij: is our design matrix at the plot level

X_ij: is our design matrix at the pixel level

Individual Models

Linear mixed model:

$$y_{ij}^{*} = \mathbf{x}_{ij}^{*}\boldsymbol{\gamma} + u_j + \varepsilon_{ij}$$

Logistic mixed model:

$$p_{ij} = \frac{1}{1 + \exp\Big(- \big(\mathbf{x}_{ij}\boldsymbol{\delta} + w_j\big)\Big)}$$

u_j and w_j are domain level random effects

Prediction

Models are fit at the plot-level but prediction happens on each pixel in the population.

$$\hat{y}_{ij}^{ZI} = \left[\mathbf{X}_{ij}\hat{\boldsymbol{\gamma}} + \hat{u}_j\right] \cdot \left[\frac{1}{1 + \exp\left(-\left(\mathbf{X}_{ij}\hat{\boldsymbol{\delta}} + \hat{w}_j\right)\right)}\right].$$

Pixel-level predictions are then aggregated to the domain level to get a small area estimate.

$$\hat{\mu}_j^{ZI} = \frac{1}{N_j}\sum_{i \in U_j} \hat{y}_{ij}.$$

Uncertainty estimation

To estimate the MSE of our estimator we implement the parametric bootstrap first introduced by Chandra and Sud (2012)

- Parametrically generates a bootstrap pixel-level population.
- Generates B-many bootstrap samples from it.
- Fits the zero-inflation estimator to each sample.
- Attains MSE estimates by comparing results to bootstrap population domain means.

Simulation Study

Treated *all* Nevada FIA plots as our population (intensified plots removed).

Created 1000 samples from this population

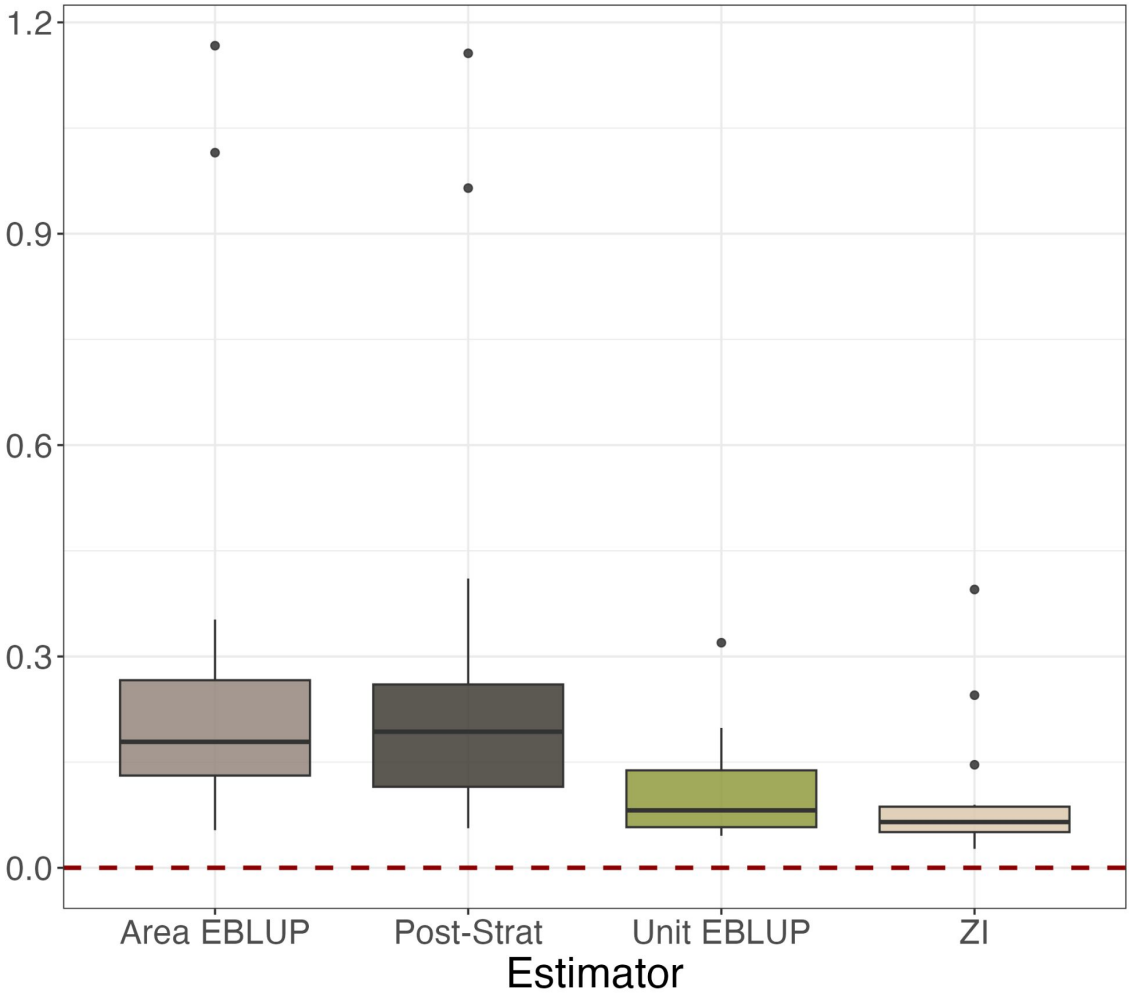
- Sampled 3% of FIA plots in each county to create a single sample
- This percent was empirically chosen as it produces roughly 20 plots per county for each sample.

Used these 1000 sample data sets to evaluate the following estimators:

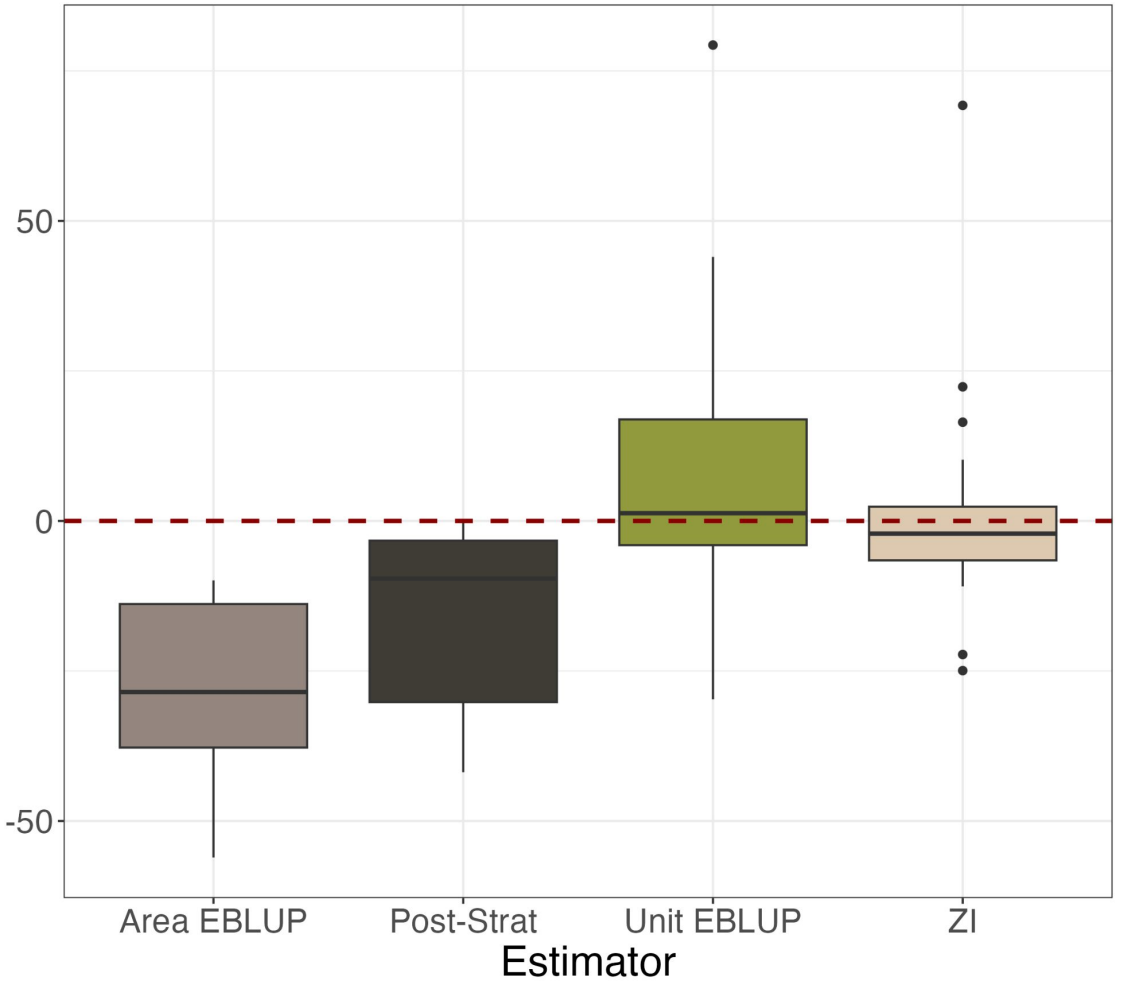
- Post-Stratified, Area-level EBLUP, Unit-level ELUP, zero-inflation

Results

Root Mean Squared Error



Percent Relative Bias



Takeaways

- Zero-Inflation can accurately quantify average biomass in regions with mixed land cover types.

- Simulation study was run in a setting favorable to the zero-inflation estimator.

    - We don’t expect it to perform well in every situation (i.e entirely forested areas).

R Package: `saeczi`

Features

- Efficient under-the-hood MSE estimation
- Syntax similar to that of existing SAE R packages (`hbsae`, `sae`)
- Written in a combination of R and C++
- Parallelized capability

Version 0.1.2 out on CRAN

Source code on GitHub at **harvard-ufds/saeczi**

Usage

```
library(saeczi)
data(pop)
data(samp)

result <- saeczi(samp_dat = samp,
                 pop_dat = pop,
                 lin_formula =  DRYBIO_AG_TPA_live_ADJ ~ tcc16 + elev,
                 log_formula = DRYBIO_AG_TPA_live_ADJ ~ tcc16,
                 domain_level = "COUNTYFIPS",
                 mse_est = TRUE,
                 B = 1000L)
```

Thank you!