

Small area estimation of forest biomass via a two-stage model for continuous zero-inflated data

Grayson W. White ^{a,b}, Josh K. Yamamoto^c, Dinan H. Elsyad^d, Julian F. Schmitt^e, Niels H. Korsgaard^d, Jie Kate Hu^f, George C. Gaines, III ^g, Tracey S. Frescino^h, and Kelly S. McConvilleⁱ

^aDepartment of Forestry, Michigan State University, East Lansing, MI, USA; ^bDepartment of Statistics & Probability, Michigan State University, East Lansing, MI, USA; ^cRedcastle Resources, Inc., Salt Lake City, UT, USA; ^dDepartment of Statistics, Harvard University, Cambridge, MA, USA; ^eEnvironmental Science and Engineering, California Institute of Technology, Pasadena, CA, USA; ^fDepartment of Statistics, The Ohio State University, Columbus, OH, USA; ^gRocky Mountain Research Station Forest Inventory & Analysis, USDA Forest Service, Missoula, MT, USA; ^hRocky Mountain Research Station Forest Inventory & Analysis, USDA Forest Service, Riverdale, UT, USA; ⁱDominguez Center for Data Science, Bucknell University, Lewisburg, PA, USA

Corresponding author: Grayson W. White (email: whitegra@msu.edu)

Abstract

Nationwide Forest Inventories (NFIs) collect data on and monitor the trends of forests across the globe. Users of NFI data are increasingly interested in monitoring forest attributes such as biomass at fine geographic and temporal scales, resulting in a need for assessment and development of small area estimation techniques in forest inventory. We implement a small area estimator and parametric bootstrap estimator that account for zero-inflation in biomass data via a two-stage model-based approach and compare the performance to a Horvitz–Thompson estimator, a post-stratified estimator, and to the unit- and area-level empirical best linear unbiased prediction (EBLUP) estimators. We conduct a simulation study in Nevada with data from the United States NFI, the Forest Inventory and Analysis Program, and remote sensing data products. Results show the zero-inflated estimator has the lowest relative bias and the smallest empirical root mean square error. Moreover, the 95% confidence interval coverages of the zero-inflated estimator and the unit-level EBLUP are more accurate than the other two estimators. To further illustrate the practical utility, we employ a data application across the 2019 measurement year in Nevada. We introduce the R package, *saeczi*, which efficiently implements the zero-inflated estimator and its mean squared error estimator.

Key words: small area estimation, forest inventory and analysis, zero-inflated models, biomass estimation, remote sensing, statistical software

Résumé

L'inventaire forestier des États-Unis (*Nationwide Forest Inventories*, NFIs) recueille les données forestières et surveille de près l'état des forêts à travers le globe. L'intérêt croissant des usagers de l'inventaire forestier dans le suivi des caractéristiques des forêts, telles que la biomasse, sur des échelles géographiques et temporelles fines, engendre la nécessité de développer des techniques d'estimation pour de petites régions géographiques adaptées au contexte des inventaires forestiers. Nous mettons en œuvre un estimateur pour petites régions géographiques ainsi qu'un estimateur paramétrique par rééchantillonnage (« Bootstrap ») qui prennent en compte l'excès de zéros dans les données de biomasse grâce à une approche en deux étapes fondée sur des modèles. Nous comparons la performance de cet estimateur à l'estimateur post-stratifié et au meilleur prédicteur linéaire sans biais empirique (MPLSBE) aux niveaux unitaires et régionaux. Nous réalisons une étude de simulation dans le Nevada, aux États-Unis, en se basant sur des données provenant de l'inventaire forestier américain et de produits de données de télédétection. Les résultats montrent que l'estimateur avec excès de zéros présente le biais relatif le plus faible et la plus petite erreur quadratique moyenne empirique. De plus, les couvertures des intervalles de confiance à 95% de l'estimateur avec excès de zéros et du MPLSBE au niveau unitaire sont plus précises que celles des deux autres estimateurs. Afin d'illustrer davantage l'aspect pratique, nous réalisons une application de données à partir des observations du Nevada de l'année 2019. Nous introduisons également le package en R, *saeczi*, qui met en œuvre l'estimateur avec

excès de zéros et son estimateur quadratique de manière efficiente. "[Ceci est une traduction fournie par l'auteur du résumé en anglais.]"

Mots-clés : estimation de petits domaines, analyse et inventaire forestier, modèle avec excès de zéros, estimation de la biomasse, télédétection, logiciel statistique

1. Introduction

The United States (US) Forest Service Forest Inventory and Analysis (FIA) Program is the strategic Nationwide Forest Inventory (NFI) of the US. FIA collects and analyzes data to evaluate and report on the extent and status of, as well as time-varying trends in, the forests of the US. These data consist primarily of measurements of forest vegetation traits made at permanent sample locations distributed across most of the nation and selected according to an equal probability, quasi-systematic sample design. The design prescribes superimposition of a grid of approximately 2400 ha hexagons across the US, and selects one plot location at random within each hexagon to sample. Another important feature of the design is the system of time-interpenetrating panels (10 years in the western US, and 5 or 7 years in the eastern US) that govern inventory cycles.

The FIA sample design permits unbiased direct estimation of forest population parameters, and estimation of associated variances, in a design-based inferential framework. The FIA design produces adequately precise design-based estimates for populations of historic importance to FIA such as, US state-level estimates over an inventory cycle-wide time period. More broadly though, users of FIA data and data from other NFIs have had an increased interest in using NFI data to produce estimates for subpopulations spanning sub-state geographies and/or short-time intervals. In these scenarios, design-based estimates are often too unstable to permit meaningful inferences on forest attributes of interest.

These imprecise design-based estimates for subpopulations are particularly relevant due to increased interest in forest attribute estimates in sub-state regions such as counties, ecological regions, and subdivisions thereof, watersheds, recently burned regions, and more. This increased interest motivated extensive research into NFI applications of small area estimation (SAE) methods (see [Rao and Molina \(2015\)](#) for SAE methods) ([Wiener et al. 2021](#)). The resounding demand for the FIA-based SAE methods and tools was acknowledged by law in recent appropriations bills, wherein US Congress directed FIA "...to further explore the use of available technologies like remote sensing and methodologies such as SAE to further refine county and state biomass estimates as outlined in Sec. 8632 of the Agriculture Improvement Act of 2018" ([U.S. Senate 2023](#)).

FIA's standard estimator is a direct, design-based and model-assisted, post-stratified estimator ([Bechtold and Patterson 2005](#)). This estimator combines plot-level data with a single categorical auxiliary variable and can be represented as a weighted sum of sample means within post-strata.

When these post-strata are homogeneous within post-strata and heterogeneous between post-strata, the post-stratified estimator allows for variance reduction over a Horvitz–Thompson estimator. However, as mentioned above,

this estimator provides insufficient precision for small areas of interest with few numbers of plots. Further explorations of design-based, model-assisted estimators, such as the generalized regression estimator, still exhibit insufficient precision in small sample size scenarios in forest inventory ([Frescino et al. 2022](#)).

The instability of the direct estimator described above highlights the need to consider alternative approaches to estimation in these small areas, and in particular, the viability of *model-based* and *indirect* estimators. For reference, *indirect* estimators consider data outside of the domain of interest to help inform estimates, while *direct* estimators, such as the post-stratified estimator, only consider data in the domain of interest; further the *model-based* inferential paradigm relies on distributional assumptions of the proposed model, while the *design-based* inferential paradigm relies only on the design of the sample. Within the model-based paradigm, we can classify the approaches to fitting models for SAE into two categories: *area-level* and *unit-level*. *Area-level* models are fit to an areal summary of the response and auxiliary data, while *unit-level* models are fit to response and auxiliary data at the granularity of the observational unit. Two common indirect, model-based approaches in small sample size cases are the empirical best linear unbiased prediction (EBLUP) estimators based on the Fay–Herriot and Battese–Harter–Fuller models ([Fay III and Herriot 1979](#); [Battese et al. 1988](#)), which are fit at the area- and unit-level, respectively, but are analogous in the sense that each of them implements an area-specific random intercept, which captures the between-area variation of the data. In forest inventory estimation applications, it is common to use auxiliary information derived from aerial and satellite-borne remote sensing platforms, including radiometric, climatic, topographic, canopy height, and other data products to inform the model. Investigations of these model forms feature prominently in the remote sensing-based forest inventory literature ([Breidenbach and Astrup 2012](#); [Ver Planck et al. 2018](#); [Temesgen et al. 2021](#); [Cao et al. 2022](#); [Frescino et al. 2022](#)); however, despite precision gains shown for small area estimates, they can exhibit significant limitations in forest inventory applications.

Oftentimes, even with the rich auxiliary data available to use in the modern day, the EBLUP estimator based on the Battese–Harter–Fuller model is poorly specified as the relationship between the forest inventory data and these auxiliary data cause modeling assumptions to not be met (see, for example, the relationship between the response and auxiliary variables in [Frescino et al. \(2022\)](#) fig. 9, subfigure A and the resulting disagreement between direct and Battese–Harter–Fuller estimates in subfigure C), in turn causing inaccurate estimates. A modeling assumption that is commonly violated here is the assumption that the error term is normally distributed. Formally, the assumption of normality for the error term of a linear regression model is specified as

$\epsilon \sim N(0, \sigma^2)$, where ϵ is the error term in the model and σ^2 is the variance of the error term, to be estimated from the sample. In order to test this assumption, we resort to examining the observed residuals, most often via normal quantile-quantile plots and residual versus fitted plots. The violation of the error term normality assumption is often caused in part by zero-inflation in forest attribute response variables.

As the name suggests, data are canonically classified as being zero-inflated when they contain a significant proportion of zeroes. While it's hardly ever very productive to spell out a definition for a phrase that is its own definition, we do so here to emphasize the fact that to call data zero-inflated is to only say something very broad about how those data are distributed. There is no commonly accepted cutoff for at what proportion of zeros our data deserve the label zero-inflated, and there is no restriction on the distribution of the nonzero data. While the work done in this paper concerns zero-inflated data with no constraint on the level of "zero-ness", we do require that the nonzero data are positive and continuously distributed. Moreover, as we are working in an estimation setting, when we say our data are zero-inflated we mean that the *response variable* is zero-inflated. In the forest inventory context, forest attributes often exhibit this specific type of zero-inflation due to the moderate frequency at which there is a lack of trees and vegetation at plot locations. Most often this occurs as a result of a plot location falling in a non-forested area, but can also be the result of a disturbance, such as a fire, at a plot site.

The difficulty of specifying a model which meets modeling assumptions, such as the normality of the error term, at the unit-level can sometimes lead researchers to take an area-level approach such as the EBLUP estimator based on the Fay–Herriot model (White et al. 2021; May et al. 2023). In forest inventory applications, the relationship between forest attributes of interest and auxiliary data is often stronger and more linear at the area-level. However, area-level models substantially reduce the number of data points and do not take full advantage of the auxiliary data available at fine resolutions. Further, the EBLUP estimator based on the Fay–Herriot model treats the within-area variation as a fixed and known quantity, while in reality this variation is an estimate derived from the unit-level sample. This assumption is particularly troublesome in small area applications, where the data used to inform the within-area variation is limited.

Despite the difficulties of specifying a useful model for estimating forest biomass, accurate estimates of forest biomass and carbon within small domains are essential for effective forest management, planning, development of sound policy and legislation, and compliance with greenhouse gas emissions reporting requirements. Furthermore, as discussed in section 5.2 of Finley et al. (2011), much of the world's tree cover is no longer in forests and misclassification of forested and nonforested regions through satellite imagery is substantial. These factors, along with the limitations of the estimators described above motivate an approach that achieves the "best of both worlds": a model-based approach that takes advantage of rich unit-level data and meets modeling assumptions while accurately quantify average biomass in regions

with mixed land cover types (e.g., counties or other politically defined regions).

A two-stage modeling approach for SAE that accounts for zero-inflation at the unit-level was first proposed and implemented by Pfeffermann et al. (2008). This method first specifies a linear mixed model fit to the positive response values in the sample data, and then a generalized linear mixed model with binomial response fit to the entire sample data to classify the probability of nonzero for a given unit. By fitting the linear mixed model to only the positive response values in the sample data, the assumption of normality of the error term is much more easily met. These models were fit in a Bayesian framework with Markov chain Monte Carlo simulations. Chandra and Sud (2012) propose a similar modeling approach in a frequentist framework, and estimate the mean squared error of small area predictions through a parametric bootstrap approach. Similar zero-inflated estimators have been applied to forest inventory data as well. In particular, Finley et al. (2011) propose a Bayesian two-stage model-based approach with spatial random effects to account for zero-inflation in forest inventory data for pixel-level prediction of forest attributes of interest; however, they do not implement or assess small area estimates. These approaches all take quite similar form: a two-stage approach with a linear model fit to the positive response data and then a binomial model fit to the entire sample. However, one may consider different approaches to this estimation problem. For example, a forest mask could be used in place of a binomial model. Further, a machine learning model such as a random forest, which does not rely on the same modeling assumptions, could be used for prediction. These approaches, while they have their merits, make it much more difficult to quantify prediction uncertainty accurately and are less interpretable from an inferential standpoint.

Here, we implement the two-stage frequentist small area estimator (henceforth zero-inflated estimator) and parametric bootstrap algorithm as described in Chandra and Sud (2012) to predict average live above-ground square root biomass (henceforth average square root biomass) in small areas of interest. In Section 2, we introduce the response and auxiliary data, define the estimators used for this analysis, and discuss our simulation framework. In Section 3.1, we conduct a simulation study in the US state of Nevada (henceforth Nevada), using FIA plot data and a variety of remotely sensed data products with Nevada's counties as our small areas of interest. The counties have a large proportion of plots with no biomass, ranging from 58% to 98% zero. Next, in Section 3.2, we apply these estimators to predict average square root biomass in one measurement year in Nevada, where each county defines a small area of interest. In both Sections 3.1 and 3.2, we compare the zero-inflated estimator to the EBLUP estimators based on the Battese–Harter–Fuller and Fay–Herriot models, and to a post-stratified estimator and a Horvitz–Thompson estimator. In Section 3.3, we describe the technical implementation of the zero-inflated estimator through the introduction of our R package, *saeczi*. Finally, in Section 4, we discuss model diagnostics and the broader impacts and implications of our work.

Table 1. Auxiliary variables used throughout the analyses.

Variable	Units	Description
ndvi	Index	LANDFIRE 2010 Landsat Normalized Difference Vegetation Index (Nelson and Steinwand 2015; LANDFIRE 2016; LANDFIRE Remap 2020)
evi	Index	LANDFIRE 2010 Landsat Enhanced Vegetation Index (Nelson and Steinwand 2015; LANDFIRE 2016; LANDFIRE Remap 2020)
tcc	%	National Land Cover Dataset (NLCD) Analytical Tree Canopy Cover (Yang et al. 2018)
elev	Meters	LANDFIRE 2010 digital elevation model (DEM) (U.S. Geological Survey 2019)
eastness	(−100 to 100)	Transformed aspect degrees to eastness (U.S. Geological Survey 2019)
northness	(−100 to 100)	Transformed aspect degrees to northness (U.S. Geological Survey 2019)
rough	Meters	Degree of irregularity of the surface (U.S. Geological Survey 2019)
tri	Index	Terrain Ruggedness Index (U.S. Geological Survey 2019)
tpi	Index	Topographic Position Index (U.S. Geological Survey 2019)
ppt	mm × 100	Parameter-elevation Regressions on Independent Slopes Model (PRISM) mean annual precipitation—30 year normals (1991–2020) (Daly et al. 2002)
tmean	°C × 100	PRISM mean annual temperature—30 year normals (1991–2020) (Daly et al. 2002)
tmax	°C × 100	PRISM mean annual maximum temperature—30 year normals (1991–2020) (Daly et al. 2002)
tmin	°C × 100	PRISM mean annual minimum temperature—30 year normals (Daly et al. 2002)
tmin01	°C × 100	PRISM mean minimum temperature (Jan)—30 year normals (Daly et al. 2002)
tdmean	°C × 100	PRISM mean annual dewpoint temperature—30 year normals (1991–2020) (Daly et al. 2002)
vpdmax	Pa × 100	PRISM max annual vapor pressure deficit—30 year normals (1991–2020) (Daly et al. 2002)
def	mm	TOPOFIRE mean annual climatic water deficit—30 year normals (1981–2010) (Holden et al. 2019)
wc2cl	Class	European Space Agency (ESA) 2020 WorldCover global land cover product (Zanaga et al. 2021)
tnt	Class	LANDFIRE 2014 tree/nontree lifeform mask (Rollins 2009; Picotte et al. 2019)

2. Methods

2.1. Data

Throughout this article, our goal is to estimate average square root biomass in a variety of small areas of interest, and to assess the quality of the estimators used. As a case study, we focus on average square root biomass estimation in the counties of Nevada as our response variable. The data for our response variable come from ground plots collected by the US Forest Service FIA Program, with sampling design as described in Section 1. Note that the choice to model and predict average square root biomass is discussed in Section 4.1. Further, we have wall-to-wall auxiliary data from a variety of

remotely sensed and climatic raster files, which have all been resampled to a 90 m × 90 m pixel size. For our analyses, we focus on 19 auxiliary variables as listed and described in Table 1. The FIA ground plots and auxiliary data were extracted and matched with the FIESTA R package (Frescino et al. 2023).

The dataset used in Section 3.1 contains 11 848 FIA plot observations across the 17 counties in Nevada, linked with 19 auxiliary variables listed in Table 1. We used the most current sampled measurement for each plot in the FIA database downloaded on 8 February 2023 (Burrill et al. 2023). For the analyses in Section 3.1, we do not need pixel-level data in Nevada, because in the simulation study we treat the entire set of plots as our target population and draw Monte Carlo

samples from this set. This approach is described in detail in [Section 2.3](#).

For the data application in [Section 3.2](#), we used a subset of the data described above to demonstrate estimator efficacy at constrained temporal scales. In particular, we used only the FIA plot observations from measurement year 2019, resulting in a sample dataset of just 1155 observations. Considering only FIA measurements made in 2019, Nevada county sample sizes range from 3 to 190 observations with a median of 60 observations. Further, for our analysis we have access to each of the auxiliary variables described in [Table 1](#) across the 33.7 million 90 m pixels comprising Nevada, our population of interest.

2.2. Estimators

2.2.1. Notation

Before describing the estimators, we introduce our notation. Let the indexed set $U = \{1, 2, \dots, N\}$ denote a finite population with N units. U can be partitioned into J small areas indexed by j where each small area population, U_j , contains N_j units. Let $s \subset U$ denote our sample of n units and the subset of sampled units s_j represent the sampled units in small area j . Finally, the individual units are indexed by i .

We use y to denote the response variable of interest, square root biomass, where y_{ij} represents the observed value for the i th unit in small area j . We are interested in estimating $\mu_j = \frac{1}{N_j} \sum_{i \in U_j} y_{ij}$, the population mean of square root biomass in small area j . Since we are interested in zero-inflation in y , we also introduce z_{ij} , a binary variable indicating whether y_{ij} is positive, formally

$$z_{ij} = \begin{cases} 1 & \text{if } y_{ij} > 0, \\ 0 & \text{if } y_{ij} = 0. \end{cases}$$

Further, for modeling we denote the unit-level vector of predictors by

$$\mathbf{x}_{ij} = (1, x_{1ij}, x_{2ij}, \dots, x_{p-1ij})^T,$$

containing the intercept and $p - 1$ predictor values for the j th county's i th plot. At the area-level, we denote the vector of predictors by

$$\bar{\mathbf{X}}_j = (1, \bar{X}_{1j}, \bar{X}_{2j}, \dots, \bar{X}_{p-1j})^T,$$

where each predictor in $\bar{\mathbf{X}}_j$ is the population mean for that predictor in a given small area. We use this quantity for both model fitting and prediction. For prediction with the zero-inflated estimator and unit-level EBLUP estimator, we use $\mathbf{X}_{ij}$ to denote the vector analogous to $\mathbf{x}_{ij}$ with the subtle difference that this is a vector of population units.

2.2.2. Horvitz–Thompson

We first introduce the Horvitz–Thompson estimator of μ_j ([Horvitz and Thompson 1952](#)), which is given by

$$\hat{\mu}_j^{\text{HT}} = \frac{1}{n_j} \sum_{i \in s_j} y_i.$$

The Horvitz–Thompson estimator does not take into account any auxiliary data and simplifies to the sample mean here because of our equal-probability sampling design. The mean square error estimator is given by

$$\widehat{\text{MSE}}(\hat{\mu}_j^{\text{HT}}) = \frac{1}{n_j(n_j - 1)} \sum_{i \in s_j} (y_i - \hat{\mu}_j^{\text{HT}})^2,$$

as described in [Särndal et al. \(2003\)](#) when we ignore the finite population correction term. Because the Horvitz–Thompson estimator is approximately unbiased under equal probability designs, its variance and mean squared error are roughly equivalent ([Särndal et al. 2003](#)). Therefore, we present the uncertainty error in terms of an MSE estimator instead of a variance estimator to be consistent with the model-based estimators presented in this article.

2.2.3. Post-stratified

As mentioned in [Section 1](#), FIA still often relies on the direct, model-assisted, post-stratified estimator for production processing of its estimates. The post-stratified estimator is the result of taking a weighted sum of Horvitz–Thompson estimates that are built over the H post-strata of a single categorical auxiliary variable within a small area, with individual post-strata indexed by h . Each weight is simply the proportion of population units in that small area that belong to the corresponding post-stratum. In our case, we use the binary categorical auxiliary variable that classifies each unit in the population into either a forested or nonforested post-stratum, so $H = 2$. The post-stratified estimator ([Holt and Smith 1979](#)) is as follows:

$$\hat{\mu}_j^{\text{PS}} = \frac{1}{N_j} \sum_{h=1}^2 \frac{N_{jh}}{n_{jh}} \sum_{i \in s_{jh}} y_i = \sum_{h=1}^2 \frac{N_{jh}}{N_j} \hat{\mu}_{jh}^{\text{HT}}.$$

As mentioned above, $\hat{\mu}_{jh}^{\text{HT}}$ represents a Horvitz–Thompson estimate for post-stratum h in small area j , and it is weighted by N_{jh}/N_j , which is the proportion of population units in small area j that belong to post-stratum h . Moreover, the MSE estimator for the post-stratified estimator ([Cochran 1977](#)) in small area j is as follows:

$$\widehat{\text{MSE}}(\hat{\mu}_j^{\text{PS}}) = \frac{1}{n_j} \left(\sum_{h=1}^2 \frac{N_{jh}}{N_j} n_{jh} \widehat{\text{MSE}}(\hat{\mu}_{jh}^{\text{HT}}) + \sum_{h=1}^2 \left(1 - \frac{N_{jh}}{N_j} \right) \frac{n_{jh}}{n_j} \widehat{\text{MSE}}(\hat{\mu}_{jh}^{\text{HT}}) \right).$$

2.2.4. Unit-level EBLUP

Thus far, we have considered direct, design-based estimators. We now turn to the unit-level EBLUP estimator, which moves beyond the direct, design-based estimation framework into an indirect, model-based estimation framework, which allows us to leverage data from outside of the small area that we are generating an estimate for, often referred to as “borrowing strength”. Borrowing strength occurs through the explicit modeling of between-area variation. For the unit-level EBLUP estimator, the underlying model is the Battese–Harter–Fuller model, a unit-level, linear mixed model with a random intercept term for small area

$$(1) \quad y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + v_j + \varepsilon_{ij}$$

where $\boldsymbol{\beta}$ is the vector of fixed effects regression coefficients, v_j is the small area random effect and ε_{ij} is random noise, distributed as follows:

$$(2) \quad v_j \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_v^2), \quad \varepsilon_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2), \quad \text{and} \quad v_j \perp \varepsilon_{ij}$$

where $\perp$ represents independence. Notice that the random intercepts pertain to the small areas in our population and its these terms that allow us to account for between-area variation. When the model captures the grouped structure of the data through the random effect and through incorporation of valuable auxiliary variables by way of the fixed effects, the unit-level EBLUP estimator can produce more precise estimates than the direct estimators described above. The parameters of this model can be estimated with a restricted maximum likelihood (REML) approach.

We produce areal estimates by aggregating the unit-level predictions to the small area level. For small area j , the unit-level EBLUP estimator of μ_j is given by

$$\hat{\mu}_j^{\text{UEBLUP}} = \frac{1}{N_j} \sum_{i \in U_j} (\mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + \hat{v}_j).$$

Furthermore, the associated MSE estimator of $\hat{\mu}_j^{\text{UEBLUP}}$ can be expressed as

$$\widehat{\text{MSE}}(\hat{\mu}_j^{\text{UEBLUP}}) = g_{1j}(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2) + g_{2j}(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2) + 2g_{3j}(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2)$$

where σ_v^2 and σ_ε^2 are estimated via REML, and g_{1j} , g_{2j} , and g_{3j} capture the within-area variation, variance in estimating regression parameters, and model-variance estimation, respectively. The formal derivations and definitions of g_{1j} , g_{2j} , and g_{3j} come from section 7.2.2 of Rao and Molina (2015) as well as Breidenbach and Astrup (2012) and are included in Appendix A.

2.2.5. Area-level EBLUP

The area-level EBLUP takes a similar linear mixed model approach as the unit-level EBLUP but is fit at the area-level. It relies on the Fay–Herriot model, which is a linear mixed model with random intercepts on area-level data. Specifically,

the following linking model is assumed:

$$\mu_j = \tilde{\mathbf{X}}_j^T \boldsymbol{\beta} + v_j, \quad v_j \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_v^2).$$

Further, we assume the following data generation model:

$$\mu_j = \hat{\mu}_j^{\text{HT}} + \varepsilon_j, \quad \varepsilon_j \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2).$$

By combining the data generation model and linking model, we obtain the Fay–Herriot model

$$(3) \quad \hat{\mu}_j^{\text{HT}} = \tilde{\mathbf{X}}_j^T \boldsymbol{\beta} + v_j + \varepsilon_j$$

with the following assumptions

$$v_j \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_v^2), \quad \varepsilon_j \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2), \quad \text{and} \quad v_j \perp \varepsilon_j.$$

We obtain area-level estimates directly from eq. (3). For small area j , the area-level EBLUP estimator of μ_j is given by

$$\hat{\mu}_j^{\text{AEBLUP}} = \tilde{\mathbf{X}}_j^T \hat{\boldsymbol{\beta}} + \hat{v}_j.$$

Further, our MSE estimator of $\hat{\mu}_j^{\text{AEBLUP}}$ is calculated with

$$\widehat{\text{MSE}}(\hat{\mu}_j^{\text{AEBLUP}}) = f_{1j}(\hat{\sigma}_v^2) + f_{2j}(\hat{\sigma}_v^2) + 2f_{3j}(\hat{\sigma}_v^2),$$

where σ_v^2 is estimated via REML, and the intuition for each of f_{1j} , f_{2j} , and f_{3j} is the same as the intuition for g_{1j} , g_{2j} , and g_{3j} defined in Section 2.2.4, respectively. Their derivations and definitions are described in Rao and Molina (2015) section 6.2.1. and are also included in Appendix A.

2.2.6. Zero-inflated

We now formally introduce the zero-inflated small area estimator. In order to specify the estimator, we initially specify two models. First, a linear mixed model is fit to the nonzero sample data. Let s_j^* represent the subset of s_j where for $i \in s_j$, it holds that $y_{ij} > 0$ and let $s^* = \bigcup_{j=1}^J s_j^*$. Then we define y_{ij}^* and $\mathbf{x}_{ij}^*$ to be the values of the response and auxiliary variables for the i th unit in s_j^* , respectively. For the sample data in s^* , we fit the following linear mixed model:

$$(4) \quad y_{ij}^* = \mathbf{x}_{ij}^{*T} \boldsymbol{\gamma} + u_j + \varepsilon_{ij}$$

where $\boldsymbol{\gamma}$ is the vector of fixed effects regression coefficients, u_j is the small area random effect, and ε_{ij} is random noise, distributed as follows:

$$(5) \quad u_j \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_u^2), \quad \varepsilon_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2), \quad \text{and} \quad u_j \perp \varepsilon_{ij}.$$

In fact, eq. (4) and eq. (5) are analogous to eq. (1) and eq. (2) but instead of being fit to the sample data in s , they are fit to the sample data in s^* . Now, assume that for unit i in small area j , z_{ij} follows a Bernoulli distribution where $p_{ij} = \mathbb{P}(z_{ij} = 1)$. Using a generalized linear mixed model with a logit link func-

tion for p_{ij} , we obtain

$$(6) \quad p_{ij} = \frac{1}{1 + \exp\left(-\left(\mathbf{x}_{ij}^T \boldsymbol{\delta} + w_j\right)\right)}$$

where $\boldsymbol{\delta}$ is the vector of fixed effects regression coefficients, and the small area random effect w_j is distributed as follows:

$$w_j \sim \mathcal{N}(0, \sigma_w^2).$$

It is important to note that while eq. (4) is fit on only the sample data where the response variable is nonzero, s^* , eq. (6) is fit to the entire sample, s . The parameters in eq. (4) and eq. (6) can be estimated using REML.

Then for $j = 1, \dots, J$ and each $i \in U_j$, we can predict y_{ij} by taking the product of the two fitted models:

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

and we can estimate μ_j by aggregating to the small area level:

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

We now turn to estimation of the MSE for the zero-inflated estimator. The MSE estimation method that we employ is a parametric bootstrap technique first introduced in Chandra and Sud (2012), which consists of the following steps:

1. Generate the bootstrap population.

- Fit the zero-inflated estimator to the original sample data and extract the estimated parameters $\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}}, \hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2$, and the realizations of the random effect for the binomial model ($\hat{w}_j$).
- Use $\hat{\sigma}_u^2$ and $\hat{\sigma}_\varepsilon^2$ to generate synthetic area-level random effects and individual random errors by drawing randomly from the following distributions:

$$(7) \quad \tilde{u}_j \sim \mathcal{N}(0, \hat{\sigma}_u^2); \quad (j = 1, \dots, J),$$

and

$$(8) \quad \tilde{\varepsilon}_{ij} \sim \mathcal{N}(0, \hat{\sigma}_\varepsilon^2); \quad (i = 1, \dots, N_j), \quad (j = 1, \dots, J).$$

Here, we use the tilde ($\sim$) over the quantities generated in eq. (7) and eq. (8), and throughout the bootstrap process to indicate that these are generated quantities associated with the bootstrap process, not estimated from the sample data. Note that by randomly drawing these generated quantities using the variance parameters estimated by the linear model component of the zero-inflated estimator, we introduce the uncertainty associated with that model into the generation of the synthetic bootstrap population.

- Using the estimated parameters obtained from fitting the zero-inflated estimator to the original sample data,

calculate the probability of positive value at each population unit,

$$(9) \quad \hat{p}_{ij} = \frac{1}{1 + \exp\left(-\left(\mathbf{x}_{ij}^T \hat{\boldsymbol{\delta}} + \hat{w}_j\right)\right)}; \quad (i = 1, \dots, N_j), \quad (j = 1, \dots, J).$$

- Generate an indicator variable by drawing from a Bernoulli distribution using the probabilities calculated in eq. (9).

$$(10) \quad \tilde{z}_{ij} \sim \text{Bernoulli}(\hat{p}_{ij}); \quad (i = 1, \dots, N_j) \\ (j = 1, \dots, J).$$

While the uncertainty associated with the generalized linear mixed model component of the original zero-inflated estimator fit is not incorporated into the bootstrap process in the same way as it is for the linear model, we still account for uncertainty in the binary outcomes by drawing the indicator variables in this way.

- From the quantities generated in eq. (7), eq. (8), and eq. (10), and $\hat{\boldsymbol{\gamma}}$ estimated from the original sample data, generate the response variable for the bootstrap population data,

$$(11) \quad \tilde{y}_{ij} = \left(\mathbf{x}_{ij}^T \hat{\boldsymbol{\gamma}} + \tilde{u}_j + \tilde{\varepsilon}_{ij} \right) \cdot \tilde{z}_{ij}; \quad (i = 1, \dots, N_j) \\ (j = 1, \dots, J).$$

The full bootstrap population data consist of matching each $\tilde{y}_{ij}$ with the $\mathbf{X}_{ij}$ that was used to generate it,

$$\left\{ \mathbf{X}_{ij}, \tilde{y}_{ij} \right\}; \quad (j = 1, \dots, J), \quad (i = 1, \dots, N_j).$$

2. Calculate synthetic area-level means from the bootstrap population data generated in eq. (11),

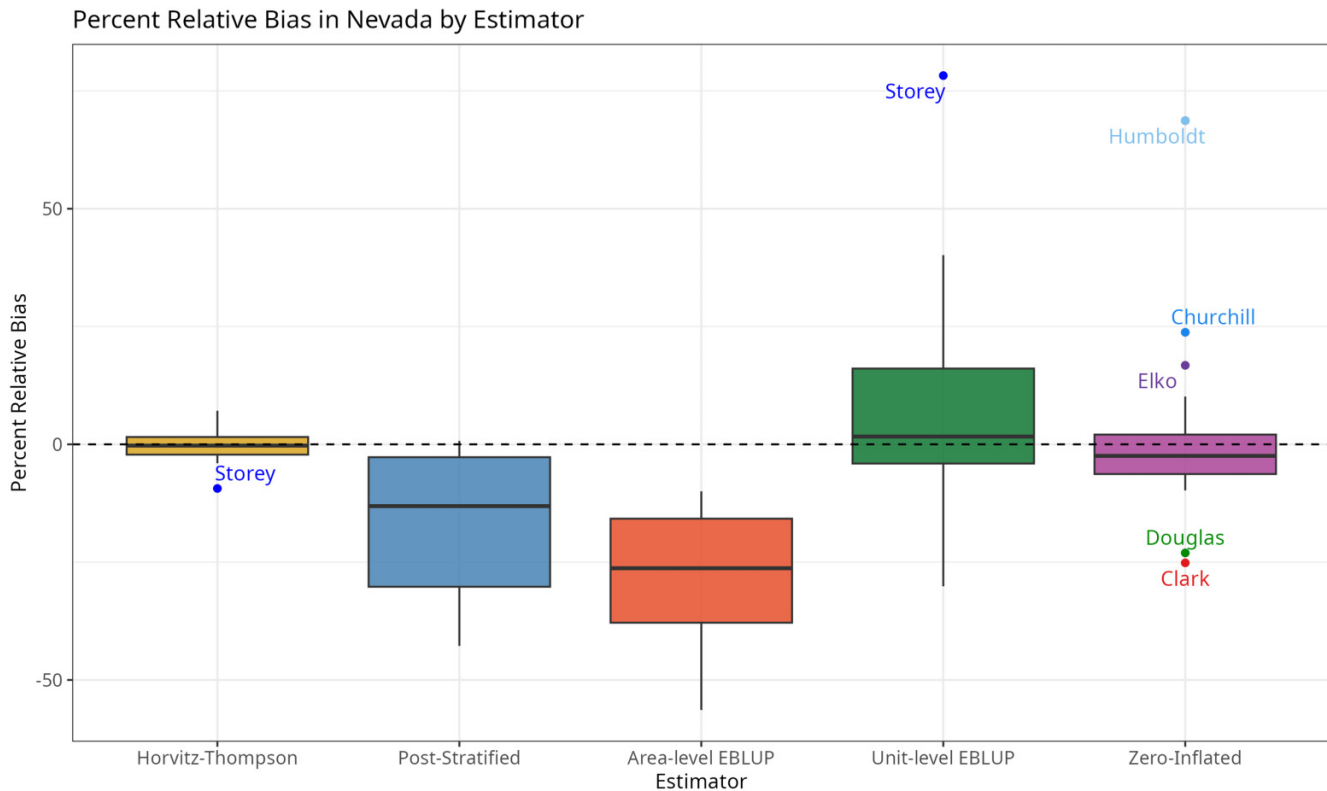
$$\tilde{\mu}_j = \frac{1}{N_j} \sum_{i \in U_j} \tilde{y}_{ij}; \quad (j = 1, \dots, J).$$

- The synthetic area-level means are the true population means of the small areas for the bootstrap population generated in Step 1.
3. Create B bootstrap samples from the bootstrap population data.
 - To do so, sample n_j units with replacement from each small area in the bootstrap population data B times. It is important to emphasize that the sampling is stratified by small area in order to preserve the sample size in each small area.
 4. Fit the zero-inflated estimator on each bootstrap sample to obtain a bootstrap distribution of estimates for the parameter of interest in each small area

$$\hat{\mu}_j^{(1)}, \dots, \hat{\mu}_j^{(B)}; \quad (j = 1, \dots, J).$$

- By fitting the zero-inflated estimator to each bootstrap sample, we implicitly introduce the uncertainty of the

Fig. 1. Percent relative bias of average square root biomass estimates in each county. The x-axis and the boxplot's fill color indicate the estimator used, while the y-axis indicates the percent relative bias value. This plot shows the percent relative bias across samples for each county. Outliers are labeled by county name and are colored consistently between estimators. EBLUP, empirical best linear unbiased prediction.



fixed and random effect parameters from both models into the bootstrap process.

- Obtain MSE estimates using the bootstrap estimates and the synthetic area-level means

$$\widehat{\text{MSE}}(\hat{\mu}_j) = \frac{1}{B} \sum_{b=1}^B \left(\hat{\mu}_j^{(b)} - \tilde{\mu}_j \right)^2; \quad (j = 1, \dots, J).$$

2.3. Simulation study

We conducted a simulation study set in Nevada to assess the quality of the zero-inflated estimator and its parametric bootstrap MSE estimator. A large majority of the land area of Nevada is classified as desert or other nonforest vegetation cover types. However, isolated mountain ranges known as “sky islands” are sparsely distributed throughout the state, and are often substantially forested. Nevada is also a large state at over 110 000 square miles and is divided into just 17 counties. These features make Nevada a great candidate for our simulation study, since (1) we would like to test the zero-inflated estimator in a location where we believe it will be useful, that is, where a significant proportion of the FIA plot data are zero; and (2) the large county size allows us to treat the set of all plots in a given county as the population for that county and then draw a small proportion of those plots for

the sample. With this in mind, we now turn to the formal framework of our simulation study.

2.3.1. Framework

We let the collection of all FIA plots in Nevada be the population, U , of size $N = 11\,848$. Further, U is partitioned by $J = 17$ counties of size N_j , where j indexes over county. We will denote the set containing all FIA plots in a given county by U_j . For our simulation study, we create $K = 1000$ Monte Carlo samples from the population U , and denote the k th sample as s_k . In order to generate a given sample, s_k , we implement Algorithm 1.

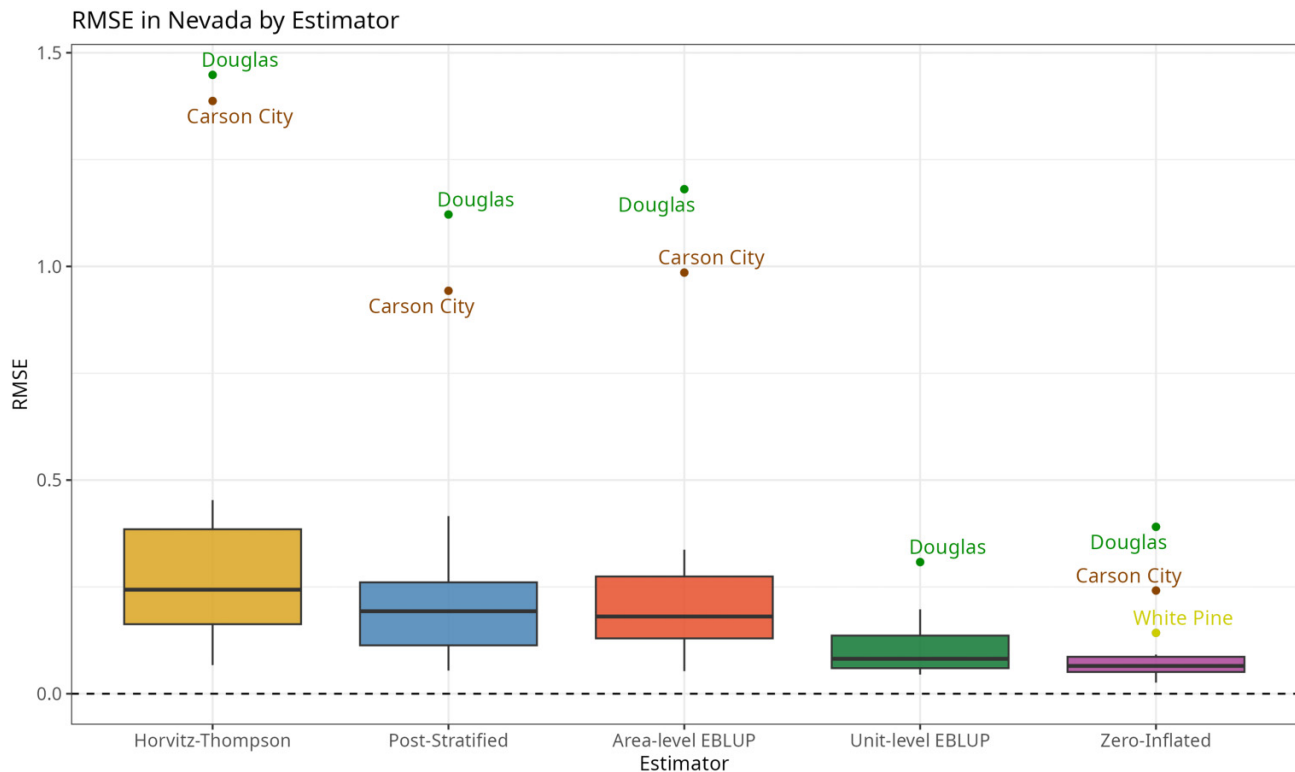
Algorithm 1. Creating one sample dataset, s_k .

```

 $s_k \leftarrow \emptyset$ 
for  $j \in \{1: J\}$  do
   $s_j \leftarrow \text{sample}(U_j, \text{proportion} = 0.03, \text{replacement} = \text{false})$ 
  if  $|s_j| < 2$  then
     $s_j \leftarrow \text{sample}(U_j, n = 2, \text{replacement} = \text{false})$ 
  end if
 $s_k \leftarrow s_j \cup s_k$ 
end for

```

Fig. 2. Empirical root mean squared error (RMSE) values in each county. The x-axis and the boxplot's fill color indicate the estimator used, while the y-axis indicates the empirical RMSE value. This plot shows the empirical RMSE values across samples for each county. Outliers are labeled by county name and are colored consistently between estimators. EBLUP, empirical best linear unbiased prediction.



Notably, we sample 3% of FIA plots in each county to give us an average number of FIA plots per county of approximately 20. Further, if sampling 3% of the county's FIA plots results in a sample of less than two, we sample two plots rather than one or zero. This allows for the within-area variance to be computed for the area-level EBLUP estimator. Repeating Algorithm 1 $K = 1000$ times gives us the 1000 sample datasets, we will assess our estimators on. The minimum number of FIA plots per county across our simulated samples was 2 and the maximum was 57.

2.3.2. Aims

Our simulation study aims to assess the zero-inflated estimator in comparison to the EBLUP estimators based on the Battese–Harter–Fuller and Fay–Herriot models, the post-stratified estimator, and the Horvitz–Thompson estimator. In order to make these comparisons, we first introduce our model fitting process, parameters of interest, and metrics for assessing the statistical performance of the estimators.

The estimators are fit in the R language using a variety of packages (R Core Team 2024). The post-stratified and Horvitz–Thompson estimators are fit using the `mase` package (McConville et al. 2018), the area-level EBLUP is fit using the `sae` package (Molina and Marhuenda 2015), the unit-level EBLUP is fit using the `hbsae` package (Boonstra 2022), and the zero-inflated estimator is fit with our package, `saeczi`,

introduced in Section 3.3 (Yamamoto et al. 2024). In order to properly and fairly assess the estimators in question, we perform model selection for each estimator across each simulation rep and consider the set of predictors from Table 1, not including any interaction terms. We use the `glmnet` package to perform model selection for each estimator on each sample created from the simulation study (Tay et al. 2023). In particular, we use the least absolute shrinkage and selection operator technique for variable selection and select the largest regularization parameter λ such that error is within one standard error of the minimum cross validation error. This choice of λ puts a greater penalty on the magnitude of the values of the estimated coefficients, leading to a more parsimonious model and a decreased chance of overfitting to the sample data.

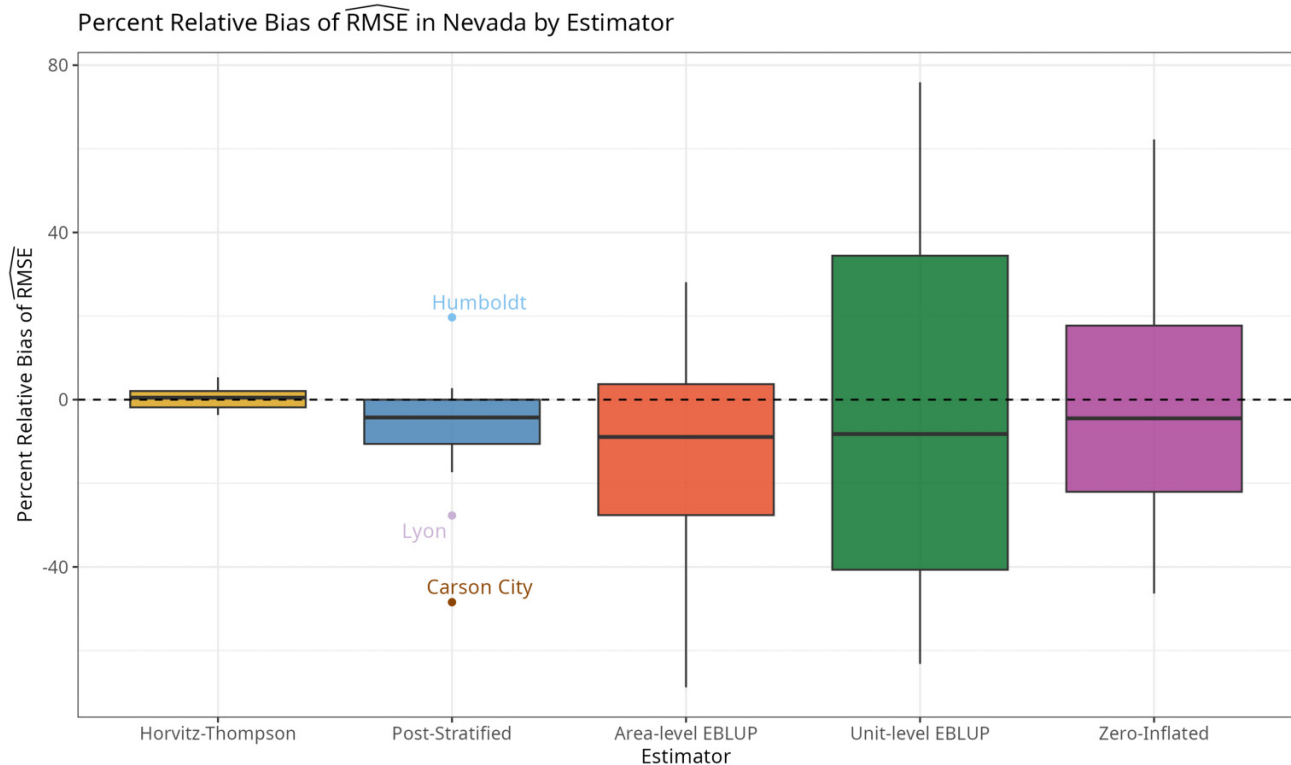
We now turn to discussing the parameters of interest in the simulation study. For a given county j , we have two parameters of interest. First

$$\mu_j = \frac{1}{N_j} \sum_{i \in U_j} y_{ij},$$

the average square root biomass in county j and second

$$(12) \quad \text{RMSE}(\hat{\mu}_{j\ell}) = \sqrt{\frac{1}{K} \sum_{k=1}^K (\hat{\mu}_{jk\ell} - \mu_j)^2},$$

Fig. 3. Percent relative bias of root mean squared error (RMSE) estimators in each county. The x-axis and the boxplot's fill color indicate the estimator used, while the y-axis indicates the percent relative bias value. This plot shows the percent relative bias of the RMSE estimators across samples for each county. Outliers are labeled by county name and are colored consistently between estimators. EBLUP, empirical best linear unbiased prediction.



the empirical root mean square error (RMSE) of the average square root biomass in county j , and $\hat{\mu}_{jkl}$ is the estimated mean square root biomass in the j th county, k th Monte Carlo sample, and ℓ th estimator. Recall that each estimator we are evaluating produces an estimate of the average square root biomass and an estimate of the RMSE of the estimator for a given county, sample, and estimator. These estimators and their corresponding MSE estimators are defined in Section 2.2, and for this analysis we consider the square root of their MSE estimator, the RMSE estimator.

We now move to discussing the metrics to assess estimator performance. To assess the bias of the estimator for county j and estimator form ℓ , denoted $\hat{\mu}_{j\ell}$, we compute the percent relative bias (PRB) as follows:

$$\text{PRB}(\hat{\mu}_{j\ell}) = \frac{\mathbb{E}[\hat{\mu}_{j\ell}] - \mu_j}{\mu_j} \cdot 100\%$$

where $\mathbb{E}[\hat{\mu}_{j\ell}]$ is the empirical expected value for estimator ℓ of mean square root biomass in county j across all Monte Carlo samples,

$$(13) \quad \mathbb{E}[\hat{\mu}_{j\ell}] = \frac{1}{K} \sum_{k=1}^K \hat{\mu}_{jkl}.$$

Beyond bias, we explore the efficiency of the estimators through the county level empirical RMSEs, defined in eq. (12).

To determine if our uncertainty estimators are underestimating or overestimating the true uncertainty, we also compute the PRB of the RMSE:

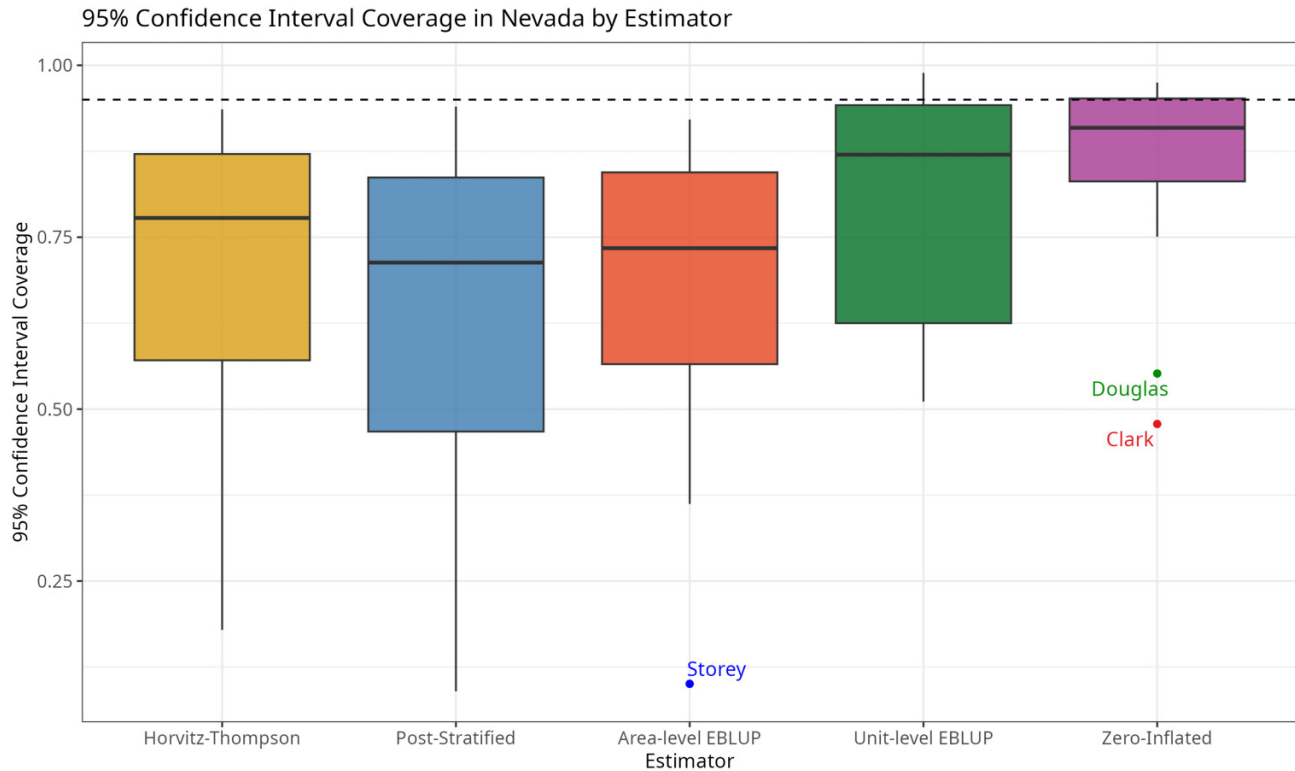
$$\text{PRB}(\widehat{\text{RMSE}}(\hat{\mu}_{j\ell})) = \frac{\mathbb{E}[\widehat{\text{RMSE}}(\hat{\mu}_{j\ell})] - \text{RMSE}(\hat{\mu}_{j\ell})}{\text{RMSE}(\hat{\mu}_{j\ell})} \cdot 100\%$$

where $\mathbb{E}[\widehat{\text{RMSE}}(\hat{\mu}_{j\ell})]$ is the empirical expected value for estimator ℓ of the RMSE for mean square root biomass in county j across all samples defined similarly to eq. (13). Lastly, we consider the confidence interval coverage probability. We define the 95% confidence interval coverage for a given county j and estimator ℓ as the proportion of samples out of the all K samples where the true parameter, μ_j , falls within the 95% confidence interval. Formally, we define the estimated confidence interval for the j th county, k th sample, and ℓ th estimator form, C_{jkl} ,

$$C_{jkl} = \left[\hat{\mu}_{jkl} - \Phi^{-1}(0.975) \cdot \widehat{\text{RMSE}}(\hat{\mu}_{jkl}), \hat{\mu}_{jkl} + \Phi^{-1}(0.975) \cdot \widehat{\text{RMSE}}(\hat{\mu}_{jkl}) \right]$$

where Φ is the cumulative density function of the standard normal distribution. Further, we define the indicator

Fig. 4. Empirical 95% confidence interval coverage in each county. The x-axis and the boxplot's fill color indicate the estimator used, while the y-axis indicates the empirical 95% confidence interval coverage. This plot shows the empirical 95% confidence interval coverage across samples for each county. Outliers are labeled by county name and are colored consistently between estimators. EBLUP, empirical best linear unbiased prediction.



function $\mathbb{1}_{C_{j\ell}}$,

$$\mathbb{1}_{C_{j\ell}}(\mu_j) = \begin{cases} 1, & \text{if } \mu_j \in C_{j\ell}, \\ 0, & \text{if } \mu_j \notin C_{j\ell}. \end{cases}$$

The 95% confidence interval coverage (CICOVG) for the ℓ th estimator form and j th county is thus

$$\text{CICOVG}(\hat{\mu}_{j\ell}, \widehat{\text{RMSE}}(\hat{\mu}_{j\ell})) = \frac{1}{K} \sum_{k=1}^K \mathbb{1}_{C_{j\ell}}.$$

3. Results

3.1. Simulation study

We now turn to discussing the results of the simulation study. Here, we examine side-by-side boxplots of each metric described in Section 2.3.2 in order to assess the performance of the estimators. Because each metric was computed by county, there are 17 values for a given metric and a given estimator.

First, we examine the PRB of the estimator for mean square root biomass in Fig. 1. Here, we see that the both the zero-inflated estimator's and Horvitz-Thompson estimator's PRB estimates are distributed closely around 0, while the other estimators have more variation in their PRB estimates. Fur-

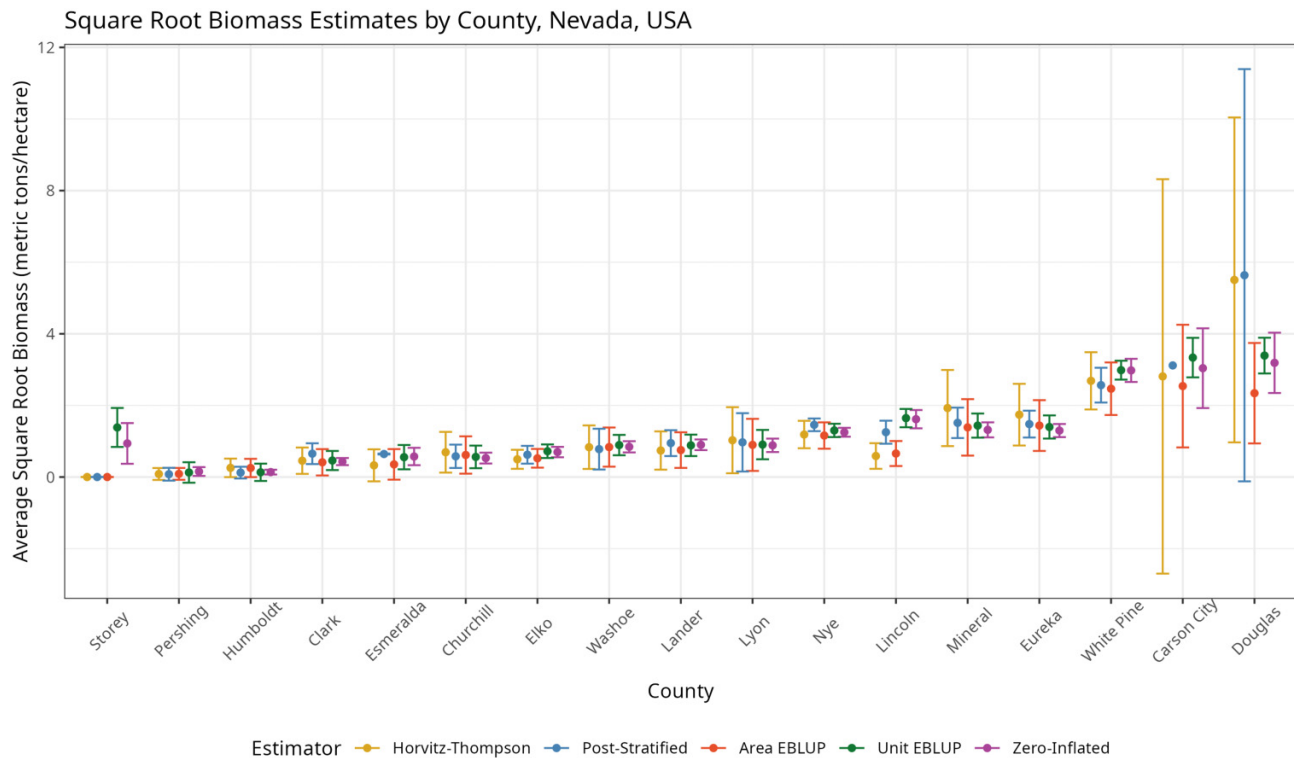
ther, the area-level EBLUP estimator is consistently underestimating the average square root biomass in the 17 counties.

The precision of the estimators is summarized in Fig. 2. We see that the zero-inflated estimator tends to have a smaller true RMSE for most counties and that unit-level estimation leads to greater precision than area-level estimation. To assess how well the RMSE estimators are quantifying the true uncertainty, we calculate the PRB of the RMSE estimator, shown in Fig. 3. All estimators (sans the Horvitz-Thompson estimator) tend to underestimate the true uncertainty and the zero-inflated estimator is competitive with the others. The unit-level EBLUP estimator has much larger variation in its bias, but on median its RMSE estimator is still close to 0.

Given the underestimation of the uncertainty in our estimates, it is not surprising that there is undercoverage by the 95% confidence intervals, as shown in Fig. 4. Here, we again see comparable performance between the unit-level EBLUP estimator and the zero-inflated estimator, with the three other estimators having much lower confidence interval coverage.

Overall, the zero-inflated estimator tends to have less bias and greater precision than the other estimators and its uncertainty estimator is competitive with the other uncertainty estimators. This indicates that the zero-inflated estimator has much promise and potential to accurately estimate forest

Fig. 5. Estimates and 95% confidence intervals of average square root biomass in each county in Nevada for measurement year 2019. Colored dots indicate the estimate and the error bars depict the 95% confidence interval. Colors distinguish between estimators, with Horvitz–Thompson estimates in yellow, post-stratified estimates in blue, area-level EBLUP estimates in red, unit-level EBLUP estimates in green, and zero-inflated estimates in purple. Counties are shown on the x-axis and are ordered by average estimate value between estimators. EBLUP, empirical best linear unbiased prediction.



inventory metrics such as biomass. In the next section we present the results of the application of these estimators to spatiotemporal domains whose sample sizes are constrained both by geography and by the limited availability of measurements made in a particular year of interest.

3.2. Data application

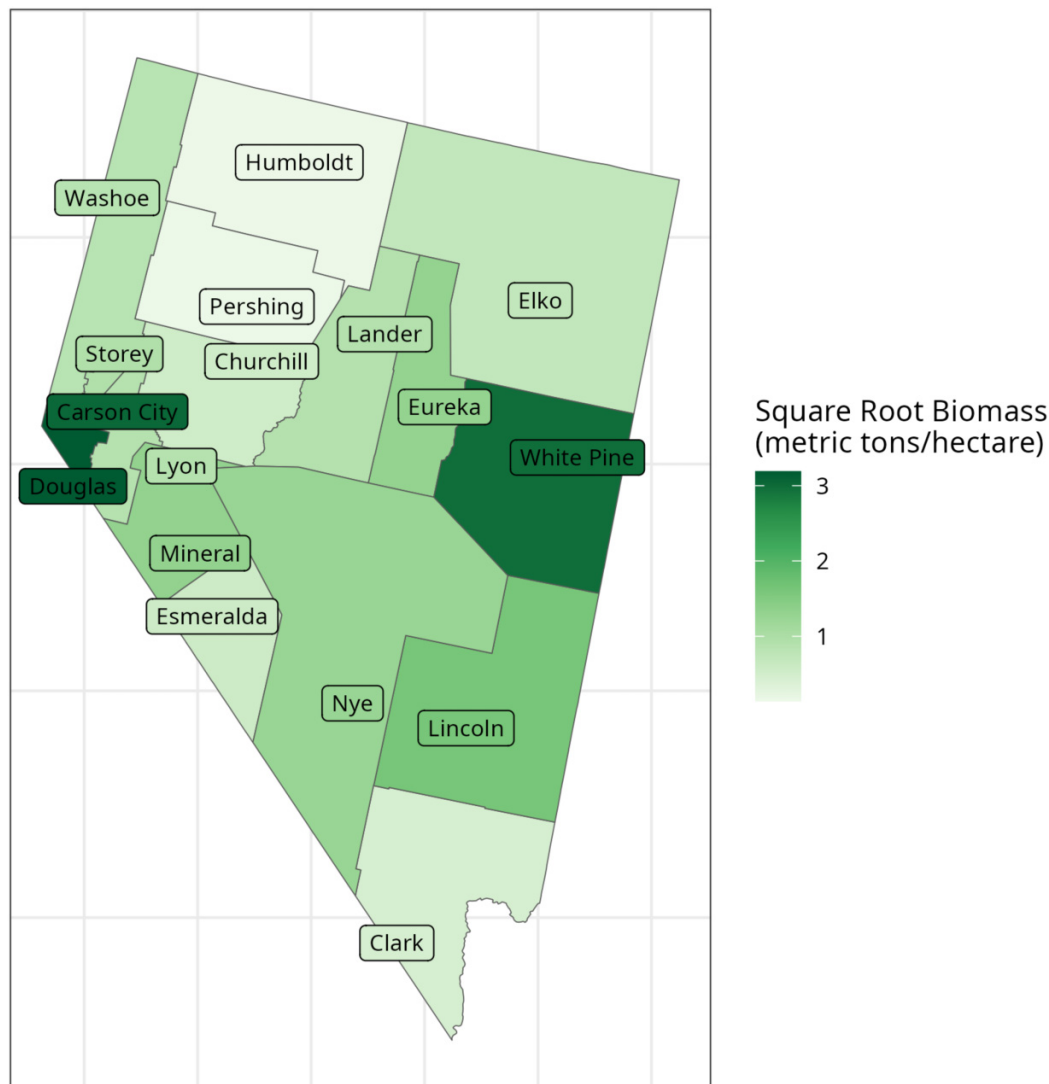
Rao and Molina (2015) define a “small area” as any domain for which direct estimates with adequate precision cannot be produced. In Section 3.1, the entire land area, forested and nonforested alike, contained by counties in Nevada, considered over a 10-year time period, served as the domains of interest. We subsampled plot measurements made in any year during the associated 10-year FIA inventory cycle using a procedure intentionally devised to yield domain sample sizes small enough to present challenges to both the design- and model-based estimators under investigation. This was particularly convenient as we were able to use the direct estimates for each county across the 10-year FIA inventory cycle as a proxy for truth.

Now, we turn to applying estimators to an authentic small area problem. In particular, we still consider counties in Nevada as our areas of interest, except now we treat only plots measured in the most recent measurement year (2019) of data as our sample. This approximately reduces our sample size by a factor of 10 and focuses the estimators on es-

timating average square root biomass in 2019 rather than implicitly assuming domain stasis over the 10-year FIA inventory cycle by incorporating all measurements, irrespective of their measurement year. For this data application, only tree canopy cover and elevation were used as auxiliary variables in the models for the model-based estimators, and the binary tree/nontree variable for the post-stratified estimator.

Estimates and 95% confidence intervals for each of the estimators fit are shown in Fig. 5. Notably, the zero-inflated estimator oftentimes has narrower confidence intervals than the other estimators, and sometimes “splits the difference” between the area- and unit-level EBLUP estimators (e.g., Storey and Douglas counties). Figure 6 displays a choropleth map of average square root biomass estimates produced by the zero-inflated estimator, which provides a sense of the large-scale spatial variation in biomass across the state of Nevada. Further, Fig. 7 displays the estimated relative efficiency of each model-based estimator compared to the Horvitz–Thompson estimator. Notably, Storey County has no estimate for relative efficiency due to the Horvitz–Thompson estimator having an estimated RMSE of zero in that county. Otherwise, the area-level EBLUP estimator has similar estimated efficiency to the Horvitz–Thompson estimator, likely due to high variability of direct estimates that are used in the model fitting process for the area-level EBLUP estimator. The zero-inflated estimator,

Fig. 6. Estimates of average square root biomass produced by the zero-inflated estimator for each county in Nevada mapped to the polygons of each county. Darker green indicates higher values, while light green indicates lower values. Labels indicate county name. Figure was created using R version 4.3.2 with package ggplot2. The shapefile, county boundaries, and county names courtesy of the FIESTA R package (Frescino et al. 2023). Base map from the ggplot2R package (Wickham 2016).



on average, shows the most estimated precision compared to the Horvitz–Thompson estimator. Further, in contrast to the zero-inflated estimator, which displays relatively consistent variation across counties, the unit-level EBLUP estimator shows a wider variety of variability and relative efficiency across counties, which can be seen in Figs. 5 and 7. This variation in the unit-level EBLUP estimator's uncertainty estimates is consistent to what we saw in Fig. 3, where the RMSE estimator for the unit-level EBLUP estimator is nearly unbiased on median, but exhibited a huge variety of bias across counties.

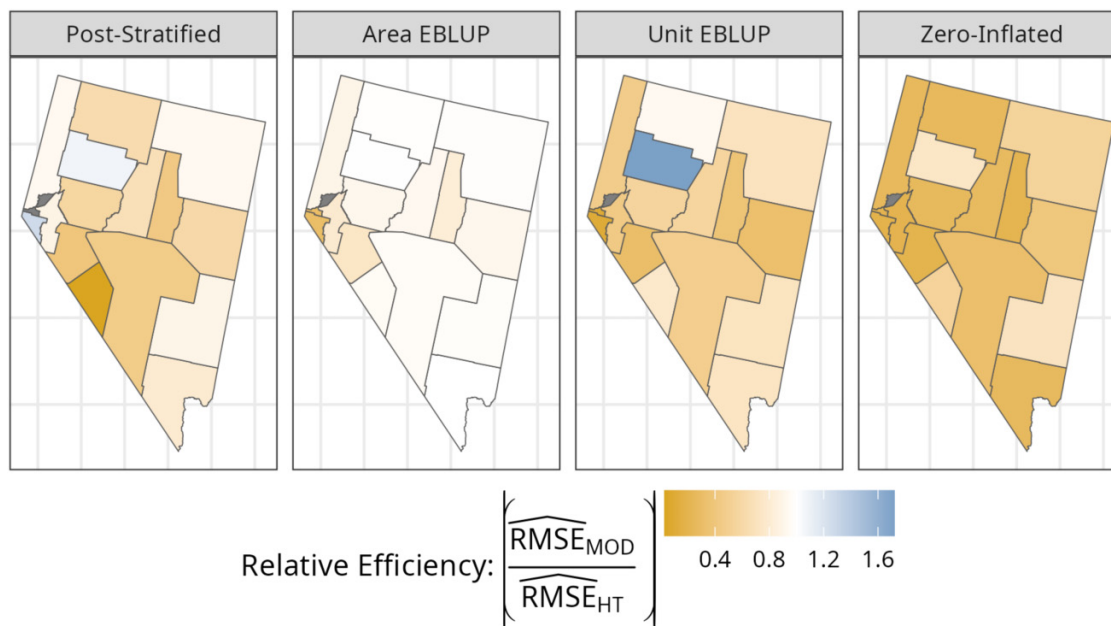
3.3. R package

In order to calculate the results for the simulation study and data application discussed in Sections 3.1 and 3.2, respectively, and to increase the accessibility of using the zero-inflated estimator, we wrote the R package *saeczi*. This R package implements both the zero-inflated estimator and its MSE estimator through efficient and parallelizable R code

and integration with C++. While Chandra and Sud (2012) introduce the zero-inflated estimator and its MSE estimator, they do not provide software to implement the estimator in practice. *saeczi* implements these estimators through an intuitive function with similar syntax to popular SAE packages in R, informative messaging and progress output as the function runs, and clear and concise examples.

The implementation of the zero-inflated estimator is a manageable programming task, however, efficiently implementing the parametric bootstrap algorithm required for MSE estimation takes more care. Not only is the bootstrap a fairly involved process, but it also is computationally intensive as each iteration requires fitting a zero-inflated estimator on a bootstrap sample and using it to predict on the bootstrap population. Thus, it does not take an inordinately large population dataset for the bootstrap to become a very time-intensive process under a standard implementation of the algorithm. *saeczi* allows a user to obtain MSE estimates

Fig. 7. Relative efficiency of average square root biomass estimates for each model-based and -assisted (MOD) estimator compared to the Horvitz–Thompson (HT) estimator. Yellow values indicate counties that have a lower estimated relative efficiency than the Horvitz–Thompson estimator and blue values indicate counties with a higher estimated relative efficiency. From left to right, the plots show values for the post-stratified estimator, area-level EBLUP estimator, the unit-level EBLUP estimator, and the zero-inflated estimator. Figure was created using R version 4.3.2 with package ggplot2. The shapefile and county boundaries courtesy of the FIESTA R package (Frescino et al. 2023). Base map from the ggplot2R package (Wickham 2016). EBLUP, empirical best linear unbiased prediction; RMSE, root mean square error.



for a zero-inflated estimator without having to spend a significant amount of time programming the algorithm in an efficient manner. Further, *saeczi* allows for a user with adequate computational resources to run the estimation process in parallel across threads.

In addition to the computational efficiency of *saeczi*'s implementation, we also focused on the user experience. We wrote *saeczi* to follow design choices of other existing small-area-estimation R packages, such as *sae* and *hbsae* (Molina and Marhuenda 2015; Boonstra 2022). While the user must still understand the zero-inflated estimator, the alignment of *saeczi*'s form with existing SAE R packages removes as much friction as possible associated with its use. We hope that using *saeczi* should feel familiar, and thus integrable, into existing SAE workflows.

4. Discussion

We implement the zero-inflated estimator described in Chandra and Sud (2012) within a forest inventory estimation context and compare the efficacy of this estimator to other small area estimators. We now turn to formally discussing the modeling approaches and diagnostics in Section 4.1, and the implications of our work in Section 4.2.

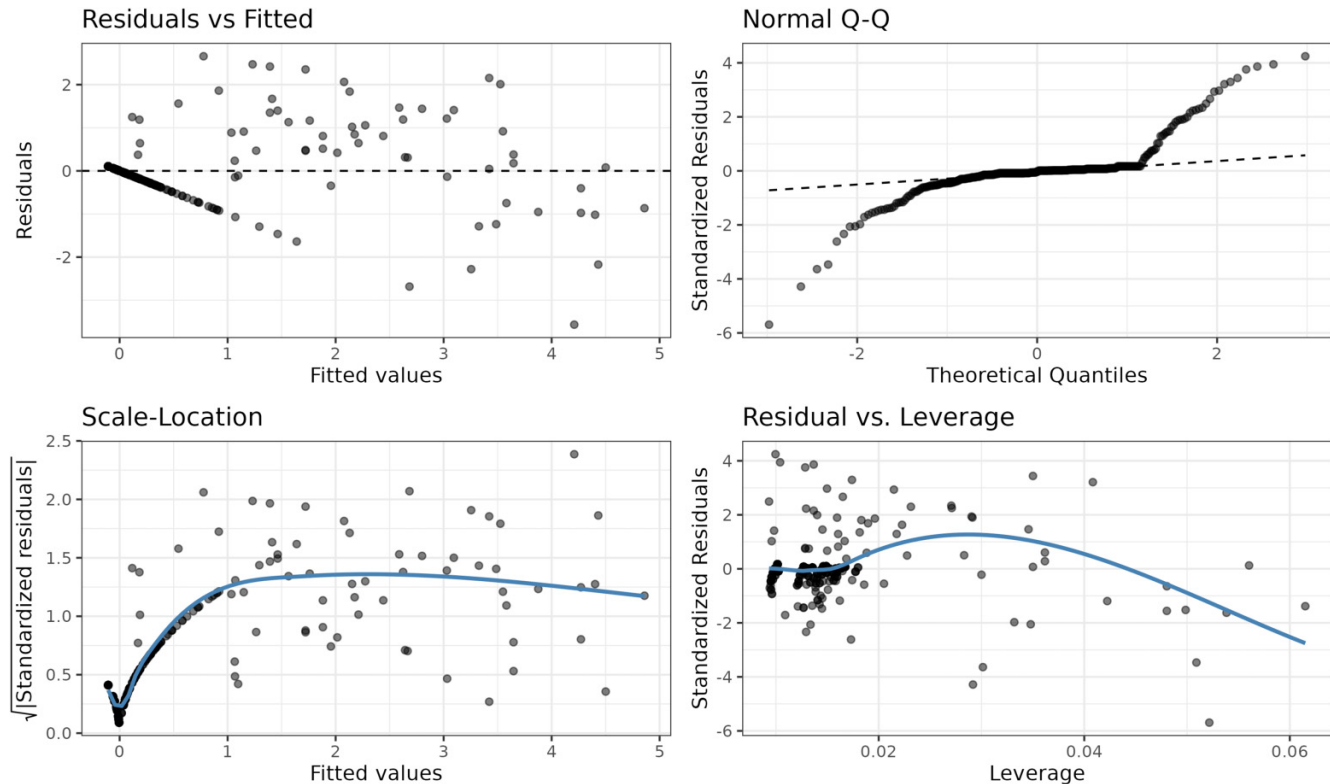
4.1. Diagnostics

For this article, we have focused on estimating average square root biomass, due to large amounts of heteroskedasticity when estimating average biomass. In general, we

would back-transform to biomass, however, the area-level EBLUP estimator produces estimates based on means of pixel-level population data, rather than the pixel-level population data itself, making it impossible to perform the back-transformation. Further, the software used to fit the unit-level EBLUP estimator also uses population means for estimation as a major computational shortcut. Thus, as the goal of this article is primarily to compare estimators under valid modeling assumptions (outside of the zero-inflation in the data, of course), we chose to model the square root biomass. It is important to note that while we chose to transform the response variable in order to handle the heteroskedasticity in the data, other approaches to handling this problem may be preferred. For example, Breidenbach et al. (2018) discuss versions of the unit- and area-level EBLUP estimators that account for heteroskedasticity via constants specified in the models and is more generalizable to differing degrees of heteroskedasticity.

When examining the validity of the modeling assumptions through formal diagnostic procedures, the area-level EBLUP estimator generally appeared to meet assumptions. We believe this to be due to the relationship between our response variable and auxiliary data to be much stronger and more linear at the area-level. Further, by aggregating to the area-level, we have entirely removed the issue with zero-inflated plots, as they are absorbed by the average. While choosing an area-level model can be convenient and easy to implement, it also comes with significant losses in performance compared to unit-level alternatives as seen in

Fig. 8. Quartet of diagnostic plots produced from one sample in the simulation study for the linear mixed model specified for the unit-level empirical best linear unbiased prediction estimator. The top left plot shows a residual versus fitted value dot plot with a dotted horizontal line $y = 0$. The top right plot shows a normal QQ plot with a dotted line indicating the $y = x$ line. The bottom left plot shows a scale-location plot with fitted values on the x-axis and the square root of absolute standardized residuals on the y-axis. The blue line is a trend line produced with local polynomial regression fitting. The bottom right plot shows a residual versus leverage plot with leverage values on the x-axis and standardized residuals on the y-axis. The blue line is a trend line produced with local polynomial regression fitting.



Figs. 1, 2, and 4. Especially in small sample size cases, the area-level EBLUP estimator can perform badly due to unstable direct estimates and direct variance estimates (which are assumed to be a fixed and known quantity in the Fay–Herriot model).

We now turn to discuss modeling assumptions between the unit-level model-based estimators. Recall that the unit-level EBLUP estimator fits a linear mixed model to the sample data, but the zero-inflated estimator fits a linear mixed model on the *positive* sample data. This distinction is a large motivation for the zero-inflated estimator, and by examining common diagnostic plots we can see why the model fit on the positive sample data is much more sensible than the one fit on the entire sample data. For one sample from the simulation study using just tree canopy cover and elevation as linear predictors, **Fig. 8** displays the quartet of modeling diagnostic plots for the linear mixed model fit on the entire sample data (i.e., the Battese–Harter–Fuller model), and **Fig. 9** displays the quartet of modeling diagnostic plots for the linear mixed model fit on the positive sample data (i.e., the linear model fit by the zero-inflated estimator).

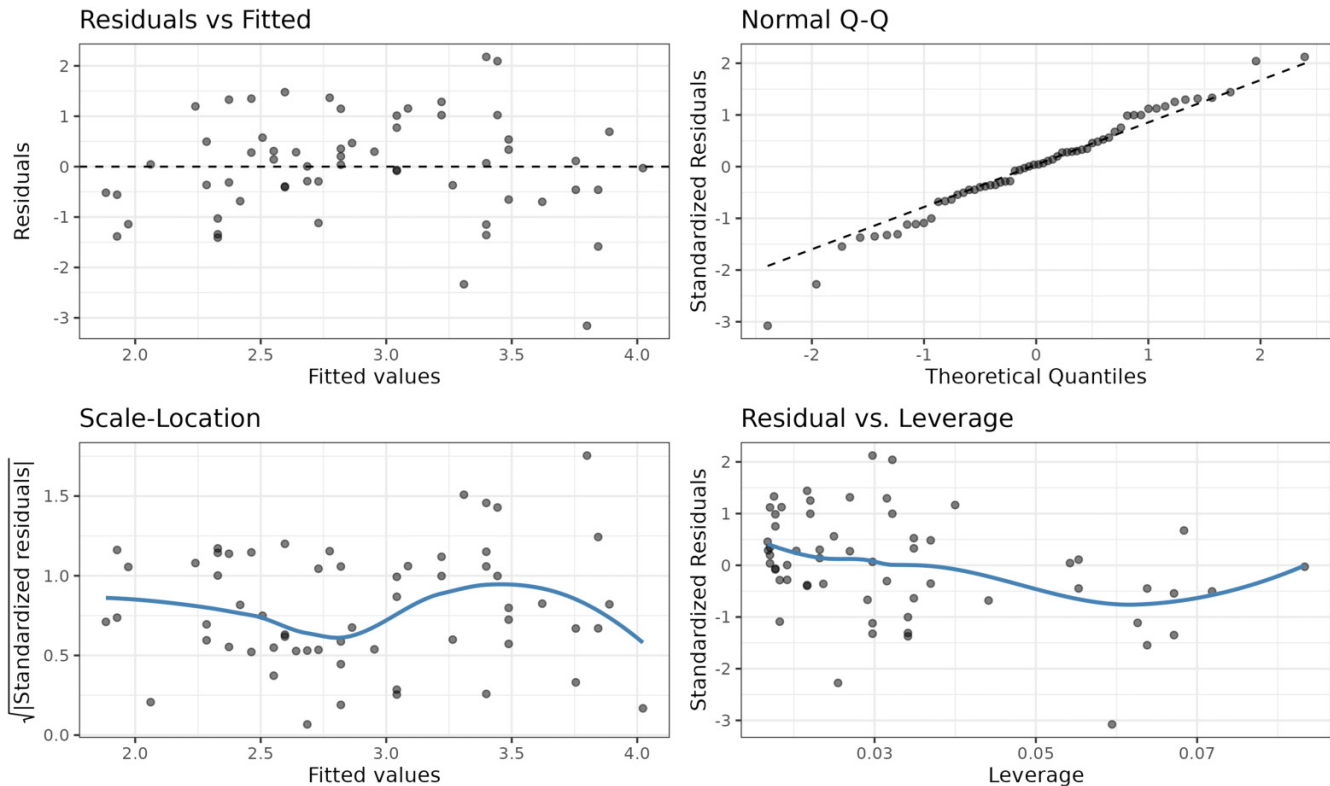
The modeling assumptions are clearly violated in **Fig. 8**. Most notably, the residual versus fitted plot trends downwards with a heavy diagonal line caused by the zero-valued

plots, the normal-QQ plot has extremely heavy tails, and the scale-location plot suffers from a distinctive “check mark” of points caused by the zero-values plots. However, these issues are largely corrected in **Fig. 9**. Here, we see a much more sensible residuals versus fitted plot where the points uncorrelated with each other and normally distributed around zero, the points on the normal-QQ plot generally fall on the theoretical line, and there are no patterns in the scale-location plot. This significant improvement of modeling diagnostics strongly motivates the zero-inflated estimator, and showcases the pitfalls of using a unit-level EBLUP estimator when the sample data contain zero-valued cases, but is otherwise positive and continuous.

4.2. Implications

Our work compares the zero-inflated estimator to two common small area estimators, the unit- and area-level EBLUP estimators, and to two direct, design-based estimators, the post-stratified estimator and Horvitz–Thompson estimator. We achieve these comparisons and assessments through a simulation study and data application, both set in Nevada. The simulation study shows that the zero-inflated estimator has much promise for estimating forest inventory attributes of interest, such as biomass, in areas where levels of zero-

Fig. 9. Quartet of diagnostic plots produced from one sample in the simulation study for the linear mixed model specified for the zero-inflated estimator. The top left plot shows a residual versus fitted value dot plot with a dotted horizontal line $y = 0$. The top right plot shows a normal QQ plot with a dotted line indicating the $y = x$ line. The bottom left plot shows a scale-location plot with fitted values on the x -axis and the square root of absolute standardized residuals on the y -axis. The blue line is a trend line produced with local polynomial regression fitting. The bottom right plot shows a residual versus leverage plot with leverage values on the x -axis and standardized residuals on the y -axis. The blue line is a trend line produced with local polynomial regression fitting.



inflation are substantial (in this case, we studied small areas where the proportion of zero-valued sample data ranged from 58% to 98%). Our work is limited in the sense that we only implement estimators in Nevada, however, we believe the zero-inflated estimator's strong performance will generalize to other similar regions where there are large amounts of land without biomass. In entirely forested regions, we do not expect the zero-inflated estimator outperform simpler estimators such as the unit- and area-level EBLUP estimators. Specifically, we expect the zero-inflated estimator to consistently outperform the simpler unit-level EBLUP estimator in scenarios where the unit-level EBLUP estimator is substantially model-misspecified, and we expect performance gains to be positively correlated with the amount of model-misspecification. The expectation is under the assumption that auxiliary data are of similar quality as to what we have used here in Nevada. Further work must be carried out to verify this expectation.

It's worth noting that the area-level EBLUP estimator is quite different from the other estimation approaches taken in our analyses. We included the area-level EBLUP estimator not only because it is a common indirect model-based approach to SAE, but also because it is often used in forest inventory settings where we see model-misspecification in unit-

level estimators (White et al. 2021; May et al. 2023). Area-level models exhibit further preferable properties including low computational cost and only requiring access to area-level summaries of inventory data. Area-level models in forest inventory often capture moderately strong linear relationships between the forest attribute of interest and auxiliary data, however, this comes at the cost of only being able to leverage a fraction of the available data. Through our simulation study and data application, we show that this trade-off is unfavorable, as the area-level EBLUP estimator consistently produces lower precision (Fig. 2) and higher bias (Fig. 1) compared to the zero-inflated estimator. In Section 1, the zero-inflated estimator was presented as a "best of both worlds" approach due to the fact that it takes advantage of the rich unit-level data while simultaneously meeting modeling assumptions. Here we have shown that the additional complexity required to achieve this is indeed worthwhile as the zero-inflated estimator outperforms both the unit- and area-level EBLUP estimators. One caveat though, in regards to the zero-inflated estimator and unit-level EBLUP estimator, is the assumption that the auxiliary data have been extracted at the precise plot location. FIA only provides perturbed plot locations publicly, and without auxiliary data extracted at the exact plot locations, even with "moderate" spatial resolution auxiliary data

such as the 90 m resolution we use here, the unit-level approaches may perform poorly. This caveat is worth considering when deciding between estimation techniques.

One might posit that because the zero-inflated estimator is more complex than the methods used for comparison, i.e., it has more parameters, is a two-stage model, etc., its estimates will improve simply due to the added complexity. However, recent work by [Schmitt \(2023\)](#) assessed the same set of estimators alongside a random forest and mixed-effects random forest, in a similar forest inventory setting, and found that the added complexity did not lead to any performance gains over the zero-inflated estimator. Machine learning methods such as random forests can be useful in forest inventory settings, however, they often produce unreliable estimates on their own. For instance, [Emick et al. \(2023\)](#) found random forest-produced forest biomass maps to be a useful auxiliary variable in parametric models for forest biomass, but largely unreliable for county-level biomass estimates when used on their own. This is all to say that, while increasing model complexity is a common approach to tackling settings where traditional estimators struggle, it does not always lead to improvements in performance. The zero-inflated estimator does increase model complexity, but it does so in a way that is entirely motivated by the structure of the data and stays within the class of parametric estimators. It is this approach to adding complexity that allows the zero-inflated estimator to increase precision and accuracy while remaining interpretable.

In further work, we hope to apply this estimator to even more diverse landscapes and assess its applicability across the US to understand the sorts of areas, this estimator is most useful. In particular, we would like to assess this estimator in areas of differing levels of zero-inflation and in areas where zero-inflation occurs in ways that are systematically different than how it occurs in Nevada. Moreover, we hope to assess the uncertainty quantification of the zero-inflated estimators in comparison to other zero-inflated estimators, especially those rooted in the Bayesian paradigm.

Acknowledgements

The authors thank the Forest Inventory and Analysis Program for the data and Romain Boutelet for the French translation of the abstract.

Article information

History dates

Received: 6 June 2024

Accepted: 13 January 2025

Accepted manuscript online: 22 January 2025

Version of record online: 7 April 2025

Copyright

© 2025 This work is free of all copyright and may be freely built upon, enhanced, and reused for any lawful purpose without restriction under copyright or database law in the United States. This work is licensed under a [Creative Commons Attribution 4.0 International License](#) (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in

any medium, provided the original author(s) and source are credited.

Data availability

Data supporting this research are available for download at <https://apps.fs.usda.gov/fia/datamart/datamart.html> with the constraint that confidential data items used in this study are not available to the public due to U.S. Federal Law. Contact USFS FIA for information regarding FIA sample location coordinates.

Author information

Author ORCIDs

Grayson W. White <https://orcid.org/0000-0003-4993-2792>

George C. Gaines, III <https://orcid.org/0000-0002-4341-6802>

Author contributions

Conceptualization: GW, JH, GG, TF, KM

Data curation: GW, TF

Formal analysis: GW, JY, DE, JS, NK, JH, GG, KM

Funding acquisition: GW, JY, GG, TF, KM

Investigation: GW, JY, DE, JS, NK, JH, GG, TF, KM

Methodology: GW, JY, DE, JS, NK, JH, GG, KM

Project administration: GW, KM

Resources: KM

Software: GW, JY, DE, JS, NK, JH, KM

Supervision: GW, GG, TF, KM

Validation: GW, GG, TF

Visualization: GW, JY, DE, JS, NK

Writing – original draft: GW, JY, JS, TF, KM

Writing – review & editing: GW, JY, JS, JH, GG, KM

Competing interests

JKY is employed by RedCastle Resources, Inc. GWW is part-time employed by RedCastle Resources, Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding information

This work was supported by the US Forest Service, Forest Inventory and Analysis Program, Rocky Mountain Research Station (via agreement 22-JV-11221638-199), and Northern Research Station (via agreement 22-CA-11221638-201); by the US Forest Service, National Forest System Program, Region 9 (via agreement 22-CA-11221638-201); by the US Forest Service, Forest Health Protection Program, State, Private, and Tribal Forestry (via agreement 22-CA-11221638-201); by Reed College; by Harvard University; and by Michigan State University AgBioResearch.

References

Battese, G.E., Harter, R.M., and Fuller, W.A. 1988. An error-components model for prediction of county crop areas using survey and satellite

- data. *J. Am. Stat. Assoc.* **83**(401): 28–36. doi:[10.1080/01621459.1988.10478561](https://doi.org/10.1080/01621459.1988.10478561).
- Bechtold, W., and Patterson, P.(Editors). 2005. The enhanced Forest Inventory And Analysis Program–national sampling design and estimation procedures. USDA Forest Service, Southern Research Station. Google Books-ID: tryLXDR2c6cC.
- Boonstra, H.J. 2022. hbsae: Hierarchical Bayesian small area estimation. R package version 1.2. Available from <https://CRAN.R-project.org/package=hbsae> [accessed 15 June 2023].
- Breidenbach, J., and Astrup, R. 2012. Small area estimation of forest attributes in the Norwegian National Forest Inventory. *Eur. J. For. Res.* **131**: 1255–1267. doi:[10.1007/s10342-012-0596-7](https://doi.org/10.1007/s10342-012-0596-7).
- Breidenbach, J., Magnussen, S., Rahlf, J., and Astrup, R. 2018. Unit-level and area-level small area estimation under heteroscedasticity using digital aerial photogrammetry data. *Remote Sens. Environ.* **212**: 199–211. doi:[10.1016/j.rse.2018.04.028](https://doi.org/10.1016/j.rse.2018.04.028).
- Burrill, E., Christensen, G., Conkling, B., DiTommaso, A., Lepine, L., Perry, C., et al. 2023. Forest inventory and analysis database: database description and user guide for phase 2 (version: 9.1). Available from <https://www.fs.usda.gov/research/understory/forest-inventory-and-analysis-database-user-guide-nfi> [accessed 15 Jan 2024].
- Cao, Q., Dettmann, G.T., Radtke, P.J., Coulston, J.W., Derwin, J., Thomas, V.A., et al. 2022. Increased precision in county-level volume estimates in the United States National Forest Inventory with area-level small area estimation. *Front. For. Global Change*, **5**: 769917. doi:[10.3389/ffgc.2022.769917](https://doi.org/10.3389/ffgc.2022.769917).
- Chandra, H., and Sud, U.C. 2012. Small area estimation for zero-inflated data. *Commun. Stat. Simul. Comput.* **41**: 632–643. doi:[10.1080/03610918.2011.598991](https://doi.org/10.1080/03610918.2011.598991).
- Cochran, W.G. 1977. Sampling techniques. Wiley Series in Probability and Mathematical Statistics. 3rd ed. John Wiley & Sons, New York-London-Sydney.
- Daly, C., Gibson, W.P., Taylor, G.H., Johnson, G.L., and Pasteris, P. 2002. A knowledge-based approach to the statistical mapping of climate. *Clim. Res.* **22**(2): 99–113. doi:[10.3354/cr022099](https://doi.org/10.3354/cr022099).
- Emick, E., Babcock, C., White, G.W., Hudak, A.T., Domke, G.M., and Finley, A.O. 2023. An approach to estimating forest biomass while quantifying estimate uncertainty and correcting bias in machine learning maps. *Remote Sens. Environ.* **295**: 113678. doi:[10.1016/j.rse.2023.113678](https://doi.org/10.1016/j.rse.2023.113678).
- Fay, R.E., III, and Herriot, R.A. 1979. Estimates of income for small places: an application of James-Stein procedures to census data. *J. Am. Stat. Assoc.* **74**(366a): 269–277. doi:[10.1080/01621459.1979.10482505](https://doi.org/10.1080/01621459.1979.10482505).
- Finley, A.O., Banerjee, S., and MacFarlane, D.W. 2011. A hierarchical model for quantifying forest variables over large heterogeneous landscapes with uncertain forest areas. *J. Am. Stat. Assoc.* **106**(493): 31–48. doi:[10.1198/jasa.2011.ap09653](https://doi.org/10.1198/jasa.2011.ap09653).
- Frescino, T.S., McConville, K.S., White, G.W., Toney, J.C., and Moisen, G.G. 2022. Small area estimates for national applications: a database to dashboard strategy using fiesta. *Front. For. Global Change*, **5**: 779446. doi:[10.3389/ffgc.2022.779446](https://doi.org/10.3389/ffgc.2022.779446).
- Frescino, T.S., Moisen, G.G., Patterson, P.L., Toney, C., and White, G.W. 2023. FIESTA: a forest inventory estimation and analysis R package. *Ecography*, e06428. doi:[10.1111/ecog.06428](https://doi.org/10.1111/ecog.06428).
- Holden, Z.A., Jolly, W.M., Swanson, A., Warren, D.A., Jencso, K., Maneta, M., et al. 2019. TOPOFIRE: a topographically resolved wildfire danger and drought monitoring system for the conterminous United States. *Bull. Am. Meteorol. Soc.* **100**(9): 1607–1613. doi:[10.1175/BAMS-D-18-0178.1](https://doi.org/10.1175/BAMS-D-18-0178.1).
- Holt, D., and Smith, T.F. 1979. Post stratification. *J. R. Stat. Soc. Ser. A*, **142**(1): 33–46. doi:[10.2307/2344652](https://doi.org/10.2307/2344652).
- Horvitz, D.G., and Thompson, D.J. 1952. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* **47**(260): 663–685. doi:[10.1080/01621459.1952.10483446](https://doi.org/10.1080/01621459.1952.10483446).
- LANDFIRE. 2016. Internal communication. U.S. Department of Agriculture and U.S. Department of the Interior.
- LANDFIRE Remap. 2020. U.S. Department of Agriculture Forest Service. Available from https://www.landfire.gov/documents/LFRemap_ImageProcessing.pdf [accessed 15 June 2023].
- May, P., McConville, K.S., Moisen, G.G., Bruening, J., and Dubayah, R. 2023. A spatially varying model for small area estimates of biomass density across the contiguous United States. *Remote Sens. Environ.* **286**: 113420. doi:[10.1016/j.rse.2022.113420](https://doi.org/10.1016/j.rse.2022.113420).
- McConville, K.S., Yamamoto, J., Tang, B., Zhu, G., Cheung, S., and Li, S. 2018. mase: model-assisted survey estimation. Available from <https://cran.r-project.org/package=mase> [accessed 15 June 2023].
- Molina, I., and Marhuenda, Y. 2015. sae: an R package for small area estimation. *R J.* **7**(1): 81–98. doi:[10.32614/RJ-2015-007](https://doi.org/10.32614/RJ-2015-007).
- Nelson, K.J., and Steinwand, D. 2015. A Landsat data tiling and compositing approach optimized for change detection in the conterminous United States. *Photogramm. Eng. Remote Sens.* **81**(7): 573–586. doi:[10.14358/PERS.81.7.573](https://doi.org/10.14358/PERS.81.7.573).
- Pfeffermann, D., Terry, B., and Moura, F. 2008. Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries. *Surv. Methodol.* **34**.
- Picotte, J.J., Dockter, D., Long, J., Tolk, B., Davidson, A., and Peterson, B. 2019. LANDFIRE remap prototype mapping effort: developing a new framework for mapping vegetation classification, change, and structure. *Fire*, **2**(2): 35. doi:[10.3390/fire2020035](https://doi.org/10.3390/fire2020035).
- R Core Team. 2024. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from <https://www.R-project.org/> [accessed 10 March 2024].
- Rao, J., and Molina, I. 2015. Small area estimation. 2nd ed. Wiley.
- Rollins, M.G. 2009. LANDFIRE: a nationally consistent vegetation, wildland fire, and fuel assessment. *Int. J. Wildland Fire*, **18**(3): 235–249. doi:[10.1071/WF08088](https://doi.org/10.1071/WF08088).
- Särndal, C.E., Swensson, B., and Wretman, J. 2003. Model assisted survey sampling. Springer Science & Business Media.
- Schmitt, J.F. 2023. A forest for the trees: using random forests for small area estimation on us forest inventory data. Bachelors thesis, Harvard University Engineering and Applied Sciences. Available from <https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37378277> [accessed 4 Nov 2023].
- Tay, J.K., Narasimhan, B., and Hastie, T. 2023. Elastic net regularization paths for all generalized linear models. *J. Stat. Softw.* **106**(1): 1–31. doi:[10.18637/jss.v106.i01](https://doi.org/10.18637/jss.v106.i01).
- Temesgen, H., Mauro, F., Hudak, A.T., Frank, B., Monleon, V., Fekety, P., et al. 2021. Using Fay–Herriot models and variable radius plot data to develop a stand-level inventory and update a prior inventory in the western cascades, or, United States. *Front. For. Global Change*, **4**: 745916. doi:[10.3389/ffgc.2021.745916](https://doi.org/10.3389/ffgc.2021.745916).
- U.S. Geological Survey. 2019. LANDFIRE elevation. USGS EROS, Sioux Falls, South Dakota.
- U.S. Senate. 2023. S. Rep. 118-83 - Department of the Interior, Environment, and Related Agencies Appropriations Bill. Available from <https://www.govinfo.gov/app/details/CRPT-118srpt83/context> [accessed 10 March 2024].
- Ver Planck, N.R., Finley, A.O., Kershaw, J.A., Jr, Weiskittel, A.R., and Kress, M.C. 2018. Hierarchical Bayesian models for small area estimation of forest variables using lidar. *Remote Sens. Environ.* **204**: 287–295. doi:[10.1016/j.rse.2017.10.024](https://doi.org/10.1016/j.rse.2017.10.024).
- White, G.W., McConville, K.S., Moisen, G.G., and Frescino, T.S. 2021. Hierarchical Bayesian small area estimation using weakly informative priors in ecologically homogeneous areas of the interior western forests. *Front. For. Global Change*, **4**. doi:[10.3389/ffgc.2021.752911](https://doi.org/10.3389/ffgc.2021.752911).
- Wickham, H. 2016. ggplot2: elegant graphics for data analysis. Springer-Verlag, New York. Available from <https://ggplot2.tidyverse.org> [accessed 10 March 2024].
- Wiener, S.S., Bush, R., Nathanson, A., Pelz, K., Palmer, M., Alexander, M.L., et al. 2021. United States forest service use of forest inventory data: examples and needs for small area estimation. *Front. For. Global Change*, **4**: 763487. doi:[10.3389/ffgc.2021.763487](https://doi.org/10.3389/ffgc.2021.763487).
- Yamamoto, J., Elsyad, D., White, G., Schmitt, J., Korsgaard, N., McConville, K.S., and Hu, K. 2024. saeczi: small area estimation for continuous zero inflated data. R package version 0.1.0. Available from <https://CRAN.R-project.org/package=saeczi>. doi:[10.32614/CRAN.package.saeczi](https://doi.org/10.32614/CRAN.package.saeczi).
- Yang, L., Jin, S., Danielson, P., Homer, C., Gass, L., Bender, S.M., et al. 2018. A new generation of the United States national land cover database: requirements, research priorities, design, and implementation strategies. *ISPRS J. Photogramm. Remote Sens.* **146**: 108–123. doi:[10.1016/j.isprsjprs.2018.09.006](https://doi.org/10.1016/j.isprsjprs.2018.09.006).
- Zanaga, D., van De Kerchove, R., De Keersmaecker, W., Souverijns, N., Brockmann, C., Quast, R., et al. 2021. ESA WorldCover 10 m 2020 v100.

Appendix A: EBLUP-estimated MSE equations

A.1. Area-level

The following equations represent the terms for the estimated MSE for the area-level EBLUP estimator (Rao and Molina 2015).

$$\begin{aligned} f_{1j}(\hat{\sigma}_v^2) &= \hat{\gamma}_j \widehat{\text{MSE}}(\hat{\mu}_j^{\text{HT}}) \\ f_{2j}(\hat{\sigma}_v^2) &= (1 - \hat{\gamma}_j)^2 \bar{\mathbf{X}}_j^T \left[\sum_{j=1}^J \bar{\mathbf{X}}_j \bar{\mathbf{X}}_j^T / \left(\widehat{\text{MSE}}(\hat{\mu}_j^{\text{HT}}) + \hat{\sigma}_v^2 \right) \right]^{-1} \bar{\mathbf{X}}_j \\ f_{3j}(\hat{\sigma}_v^2) &= \widehat{\text{MSE}}(\hat{\mu}_j^{\text{HT}})^2 \left(\widehat{\text{MSE}}(\hat{\mu}_j^{\text{HT}}) + \hat{\sigma}_v^2 \right)^{-3} \\ &\quad \times \left[2 \sum_{j=1}^J 1 / \left(\hat{\sigma}_v^2 + \widehat{\text{MSE}}(\hat{\mu}_j^{\text{HT}}) \right)^2 \right]^{-1} \end{aligned}$$

where

$$\hat{\gamma}_j = \hat{\sigma}_v^2 / \left(\widehat{\text{MSE}}(\hat{\mu}_j^{\text{HT}}) + \hat{\sigma}_v^2 \right)$$

A.2. Unit-level

The following equations represent the terms for the estimated MSE for the unit-level EBLUP estimator

(Rao and Molina 2015; Breidenbach and Astrup 2012).

$$\begin{aligned} g_{1j}(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2) &= \hat{\gamma}_j (\hat{\sigma}_v^2 / n_j) \\ g_{2j}(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2) &= (\bar{\mathbf{X}}_j - \hat{\gamma}_j \bar{\mathbf{X}}_j)^T \left[\sum_{j=1}^J \hat{\sigma}_\varepsilon^{-2} \left(\sum_{i=1}^{n_j} \mathbf{x}_{ij} y_{ij} - \hat{\gamma}_j n_j \bar{\mathbf{X}}_j \bar{y}_j \right) \right]^{-1} \\ &\quad \times (\bar{\mathbf{X}}_j - \hat{\gamma}_j \bar{\mathbf{X}}_j) \\ g_{3j}(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2) &= n_j^{-2} (\hat{\sigma}_v^2 + \hat{\sigma}_\varepsilon^2 / n_j)^{-3} h(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2) \end{aligned}$$

where

$$\hat{\gamma}_j = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \hat{\sigma}_\varepsilon^2 / n_j)$$

and

$$\begin{aligned} h(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2) &= \hat{\sigma}_\varepsilon^4 \left[1/2 \sum_{j=1}^J n_j^2 (\hat{\sigma}_\varepsilon^2 + n_j \hat{\sigma}_v^2)^{-2} \right]^{-1} \\ &\quad + \hat{\sigma}_v^4 \left[1/2 \sum_{j=1}^J \left((n_j - 1) \hat{\sigma}_\varepsilon^{-4} + (\hat{\sigma}_\varepsilon^2 + n_j \hat{\sigma}_v^2)^{-2} \right) \right]^{-1} \\ &\quad - 2 \hat{\sigma}_\varepsilon^2 \hat{\sigma}_v^2 \left[1/2 \sum_{j=1}^J n_j (\hat{\sigma}_\varepsilon^2 + n_j \hat{\sigma}_v^2)^{-2} \right] \end{aligned}$$