frontiers | Research Topics

Small area estimation in forest inventories: New needs, methods, and tools

Edited by
Barry Wilson, John Coulston, Steve Prisley and Philip Radtke

Published in
Frontiers in Forests and Global Change

ISSN 1664-8714
ISBN 978-2-83251-647-8
DOI 10.3389/978-2-83251-647-8

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Small area estimation in forest inventories: New needs, methods, and tools

Topic editors

Barry Wilson — Northern Research Station, Forest Service (USDA), United States
John Coulston — Southern Research Station, Forest Service (USDA), United States
Steve Prisley — National Council for Air and Stream Improvement, Inc (NCASI), United States
Philip Radtke — Virginia Tech, United States

Citation

Wilson, B., Coulston, J., Prisley, S., Radtke, P., eds. (2023). *Small area estimation in forest inventories: New needs, methods, and tools*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83251-647-8

Table of contents

frontiers in Forests and Global Change

SYSTEMATIC REVIEW
published: 15 July 2021
doi: 10.3389/ffgc.2021.695929

A Systematic Review of Small Domain Estimation Research in Forestry During the Twenty-First Century From Outside the United States

*Richard W. Guldin**

Guldin Forestry LLC, Silver Spring, MD, United States

Small domain estimation (SDE) research outside of the United States has been centered in Canada and Europe—both in transnational organizations, such as the European Union, and in the national statistics offices of individual countries. Support for SDE research is driven by government policy-makers responsible for core national statistics across domains. Examples include demographic information about provision of health care or education (a social domain) or business data for a manufacturing sector (economic domain). Small area estimation (SAE) research on forest statistics has typically studied a subset of core environmental statistics for a limited geographic domain. The statistical design and sampling intensity of national forest inventories (NFIs) provide population estimates of acceptable precision at the national level and sometimes for broad sub-national regions. But forest managers responsible for smaller areas—states/provinces, districts, counties—are facing changing market conditions, such as emerging forest carbon markets, and budgetary pressures that limit local forest inventories. They need better estimates of conditions and trends for small sub-sets of a national-scale domain than can be provided at acceptable levels of precision from NFIs. Small area estimation research is how forest biometricians at the science-policy interface build bridges to inform decisions by forest managers, landowners, and investors.

Keywords: estimating forest conditions for small areas, using national forest inventory data at small spatial scales, remotely sensed imagery as auxilliary data for imputation, small area estimation research in Europe, driving forces spurring small area estimation research

OPEN ACCESS

Edited by:
Gretchen Moisen, United States Forest Service (USDA), United States

Reviewed by:
Paula Soares, University of Lisbon, Portugal
Steve Prisley, National Council for Air and Stream Improvement, Inc. (NCASI), United States

***Correspondence:**
Richard W. Guldin
rich@guldinforestry.com

Specialty section:
This article was submitted to Forest Management, a section of the journal Frontiers in Forests and Global Change

Received: *15 April 2021*
Accepted: *21 June 2021*
Published: *15 July 2021*

Citation:
Guldin RW (2021) A Systematic Review of Small Domain Estimation Research in Forestry During the Twenty-First Century From Outside the United States. Front. For. Glob. Change 4:695929. doi: 10.3389/ffgc.2021.695929

INTRODUCTION

Defining Small Domain Estimation and Small Area Estimation

A study domain is a major segment of some population for which separate statistics are needed. A domain can be defined in many ways, including a demographic characteristic (e.g., an age stratum or an ethnic group) or an economic characteristic (e.g., a category of manufacturers) or a geographical area (e.g., a political jurisdiction, the range of a tree species, a hydrologic basin) (Lavrakas, 2008).

According to Brackstone (1987), small domain estimation (SDE) began several centuries ago—in the eleventh century in England and the seventeenth century in Canada. Those were

based on data from complete enumerations of the population's members, commonly called censuses. But today, even what are called mandatory censuses rarely obtain data from all members of the population. Further the costs and administrative burdens of complete enumerations are no longer affordable. Therefore, surveys based on statistical sampling designs applied to a population have largely replaced complete enumerations as data collection tools, and statistical estimation procedures are used to make inferences from the sample data about the characteristics of the entire domain's population. The sampling and statistical plan for the survey normally specifies the level of detail and reliability or precision required in the estimates sought from the sample data. But despite well-constructed sampling and statistical plans that carefully balance users' information needs and available funding, demands from data users continually arise for attribute estimates for sub-populations in sub-domains that were not envisioned in the original plans (U.S. Census Bureau, 2021).

A clear definition of SDE and SAE can be obtained by melding points made by three authors. Lavrakas (2008, p. 675) distinguished between large domains or areas and small domains or area, saying, "A domain is considered 'large' or 'major' if the domain sample is sufficiently large so that it can provide *direct estimates* of the domain parameter, for example, the mean, with adequate precision. A domain or area is regarded as "small" if the domain-specific sample is not large enough to produce an estimate with reliable precision. Areas or domains with small samples are called *small areas, small domains, local areas, subdomains,* or *substates.*"

Pfeffermann (2013, p. 41) acknowledged and clarified definitional confusion, "The term 'small area estimation' is somewhat confusing because it's the size of the sample from the area that causes estimation problems, not size of the area."

Ghosh (2020, p. 2) said that what is important in defining *small area* is, "… the 'smallness' of the targeted population within an area that constitutes the basis for SAE." He elaborated further that, "A domain is regarded as 'small' if domain-specific sample size is not large enough to produce estimates of desired precision." and "A domain (area) specific estimator is 'direct' if it is based only on the domain-specific sample data." In comparison, an indirect estimator is one that requires additional data. The additional data may come from the same geographic area though not part of the original domain-specific sample or it may come from related geographic areas and/or time periods. In all cases, the additional data are used, "… to increase the "effective" sample size. This is usually done through the use of models, mostly 'explicit', or at least 'implicit' that links the related areas and/or time periods …" to the original domain-specific sample data to yield estimates of acceptable precision.

The *raison d'etre* for SDE/SAE is that the data in one domain of information is too sparse to yield acceptable precision—root mean square error (RMSE) or other error statistic—for a desired estimate computed from the domain's sample data. To improve the precision/reduce the RMSE of the desired estimate, small area methods "borrow strength" (Ghosh and Rao, 1994) from auxiliary data, as the quotations from Ghosh (2020) in the previous paragraph outlined. Direct estimators, using only plots from within the domain of interest, may also incorporate auxiliary data for improved precision. However, the sample within that specific domain may not be large enough for estimates to be made sufficiently precise, even with the assistance of auxiliary data. This review includes methods of SAE as defined in the Introduction, as well as closely related direct methods that improve estimates of forest conditions.

When the definitional aspects of U.S. Census Bureau (2021), Lavrakas (2008), Pfeffermann (2013) and Ghosh (2020) are melded—all rooted in prior work by Ghosh and Rao (1994), Pfeffermann (2002) and Rao (2003)—definitional clarity emerges on two points. First, "small area" in SAE is synonymous with "small domain," thus SAE and SDE are identical conceptually. Although SDE is the still the proper statistical term, when the domain is defined spatially, it's become common to refer to SDE as SAE. In the January 2019 issue of *The Survey Statistician*—newsletter of the International Association of Survey Statisticians—the editors said, "*Small area estimation is one of the most popular topics in survey statistics of the 21st century.*" (Krapavickaite and Rancourt, 2019, p. 3). Following the custom of the International Association of Survey Statisticians, the term SAE will be used instead of SDE in the remainder of this paper. Second and more important definitionally, "area" and "domain" are surrogates for the fact that the defined geospatial area or subdomain has a dataset too small to yield credible direct estimates. Thus, indirect estimators are required where data are few or even non-existent.

Forest Sector Interest in SAE

In forestry today, the word "area" is commonly considered a geospatial term that describes the space enclosed by a set of arcs (lines or boundaries to a closed polygon, defined by a set of vertices or coordinate points). The space inside the polygon—its "area"—is measured in units such as acres or hectares.

Whether an area with forests is considered "large" or "small" is typically a function of the jurisdiction of the public forester or private forest manager, landowner, or investor. To a forest manager or landowner responsible for hundreds to several thousand acres, "small area" might mean a specific stand, compartment, or management unit—a small subset of a property composed of tens to a hundred stands or compartments. To a state/provincial forester responsible for a million hectares or several million acres of forests, "small area" might mean a county, municipality, or group of counties and municipalities within their state or province. The point is the total forest area within a jurisdiction or ownership creates meaning and context for those responsible for the jurisdiction or ownership about what they consider a "large area" vs. a "small area." This not only means that different foresters have different notions about relative sizes regarded as "small," but their notions differ conceptually and fundamentally from what a "small area" means to a forest biometrician or statistician in the context of SAE. Consequently, there are currently misconceptions and misunderstandings within the forestry community over what "small area" in SAE really means. One of the purposes of this special issue of "Frontiers in Forests and Global Change" is to clarify conceptions and improve understanding at the science-policy interface between forest biometricians and statisticians

on the one hand who have made much scientific progress the past two decades in SAE, and on the other hand, foresters, landowners, and investors who need reliable, credible estimates of forest conditions and trends to make well-informed resource management and policy decisions for lands within their purview.

In the forest sector, the domain/population for a national forest inventory (NFI)[1] is typically the entire country. While the NFI sampling design and intensity may yield credible estimates[2] at the national level, often the low spatial sampling intensities for NFI purposes may yield datasets too sparse for making estimates with acceptable precision at state/provincial, or smaller spatial scales. Forest managers and policy makers most commonly turn to digital aerial photography or data from other passive or active sensors[3] as the auxiliary datasets to "borrow strength" (Ghosh and Rao, 1994) for improving the RMSEs of desired NFI data attributes. Through imputation models, including both parametric (e.g., multiple regression) and non-parametric models (e.g., k-nearest neighbor), image pixels are assigned to classes, such as species or cover types or stand heights, or given values, such as growing stock volume or biomass volume. Then, after geospatial boundary files are defined for the small area, the imputed pixel values are summarized using some algorithm to make an SAE with acceptable RMSE.

Organizers of an SAE-focused technical session at the 2019 Forest Inventory and Analysis (FIA) Stakeholders Science meeting[4] invited the author to present a review of ***recent*** SAE research ***pertaining to forests*** by researchers ***outside the United States*** to provide context—both historic and current—for SAE presentations by U.S. researchers using NFI data from the FIA program. The invitation set the sideboards and shaped the survey design for this systematic review of SAE research.

[1]National forest inventory means an inventory of all the forests in a nation, not an inventory of just federally owned forests.

[2]The Forest Inventory and Analysis (FIA) program—the U.S.A.'s NFI—is based on a three-phase sampling design, the second phase of which is an array of roughly 127,000 hexagons, each 6,000 acres in size with one permanent plot. Precision standards for phase 2 population estimates are plus/minus 3% per million acres of Timberland and plus/minus 5% per billion cubic feet of growing-stock volume in the Eastern United States. See Bechtold and Patterson (2005) for more details.

[3]Passive sensors measure natural radiation (e.g., reflected sunlight) while active sensors use their own energy source to emit radiation and record what's reflected to the sensor. The term "imagery" usually refers to data collected only by passive sensors, often called optical sensors because they "see" reflected sunlight across visible spectral bands or re-emitted sunlight energy across near infrared, thermal infrared, and/or short-wave infrared bands. An example is the Thematic Mapper (TM) sensor on LANDSAT that collects data across seven spectral bands. Passive sensor data are often characterized on two principal ways—by spectral resolution (the number of bands of reflected/reemitted radiation recorded) or by spatial resolution of the recorded information [hundreds of meters (e.g., MODIS and AVHRR), tens of meters (e.g., TM or Sentinel), or meters (e.g., IKONOS)]. In contrast, data collected by active sensors, such as LiDAR or synthetic aperture radar, is not usually referred to as "imagery." Active sensor resolution is usually characterized as a function of the radiation emitter. Some sensors are satellite-borne, like GEDI on the International Space Station, TM on LANDSAT or MODIS on the EOS-AM and EOS-PM satellites. Other sensors are carried by piloted aircraft or unmanned aerial vehicles (UAVs) or terrestrial-based. See Gutman (2010) or Canadian Centre for Remote Sensing (CCRS) (2019). See https://www.gedi.umd.edu/mission/mission-overview for details.

[4]Held November 19–21, 2019, in Knoxville, Tennessee, USA.

METHODS

Defining the Survey Criteria

Recent

Small area estimation research accelerated in the mid-1990s. There was an upsurge across many disciplines from basic statistics to diverse applied statistics disciplines. Starting in 2000, articles about developing SAE from NFI data began appearing in the forestry and remote sensing literature. This review assumed that papers published in the first decade of this century are already well-known. Thus, 2010 was the threshold chosen for defining "recent." However, information is included from the previous decade (2000–2009) to provide context for the more recent decade.

Pertaining to Forests

Many articles have been published about SAE research across many different disciplines. They range from pure statistical theory to applied statistical research. Regarding applied research, results of case studies span all domains, from social (e.g., health care, education) and economic (e.g., poverty, marketing) to ecological (e.g., farm crop production, meteorology, and forests). Again, this paper ignores all the excellent work in other domains and sub-domains to focus tightly on applied statistical research related to making estimates of forest conditions from small geographic areas.

Outside the United States

Early SAE research was accomplished within the pure statistics community by statisticians outside the United States. Leading statisticians who summarized the state-of-science at various times were from Canada [J.N.K. Rao (Ghosh and Rao, 1994; Rao, 2003; Rao and Molina, 2015)], the United Kingdom [Danny Pfeffermann (Pfeffermann, 2002, 2013), Ayoub Saei and Ray Chambers (Saei and Chambers, 2003)], and Australia [Azizur Rahman (Rahman, 2008)]. Although their reviews included citations of work by statisticians in the United States, many—if not most—of the articles they reviewed were from researchers in Europe, India, and elsewhere.

A challenge to focusing on SAE research results from individuals and teams from outside the United States is that Ronald McRoberts,[5] was deeply involved with colleagues from other countries in seminal SAE research pertaining to forests. The early international collaboration emerged from his activities within the International Union of Forest Research Organizations (IUFRO). Beginning in the late 1990s, McRoberts published many SAE articles with international coauthors. Therefore, it is impossible to tease apart completely the international SAE research progress pertaining to forests from domestic SAE research progress because so much of the early progress here and abroad was led or influenced by McRoberts.

With these three survey design criteria in place, the rest of the paper presents an overview of applied research since

[5]Adjunct Professor, Department of Forest Resources, University of Minnesota, and Principal, Raspberry Ridge Analytics LLC. Formerly, USDA Forest Service, Northern Research Station FIA program until June 2019.

2010 about connecting national- and regional-scale forest inventories to smaller geographic areas from researchers outside the United States.

Designing the Review Survey

There is little evidence that auxiliary data other than spatial information from passive or active sensors have been used in any of the forestry disciplines in the United States. For example, census data haven't been used to improve SAEs for woodland owners' characteristics or attributes. Nor have economic data from the Commerce Department's Bureau of Economic Analysis been used to improve SAEs for forests. Therefore, the search for international SAE activities will focus mainly on applications using remotely sensed data as auxiliary data, and secondarily on socio-economic census data.

The fact that remotely sensed spatial data have been the prime source of auxiliary data guided the design of the search for international examples down three pathways. First, a broad-based search for SAE research and applications by scientists from other countries was conducted using Google Scholar and ResearchGate. Two mandatory search terms were used ("small area estim" and "forest") to identify journals that had the large numbers of SAE articles related to forestry. Second, detailed searches of the archives of the leading journals publications were conducted. Third, additional searches were conducted, based on literature cited in articles from the leading journals, and international authors who had published in the leading journals and may have published in journals less-frequently identified in the initial searches. Finally, based on these results, personal contacts were made with the leading experts working in NFIs in Europe and Canada to understand their current research programs underway and recent progress that perhaps had not yet been published. The experts were identified through the author's IUFRO network connections.

RESULTS

Google Scholar and ResearchGate Search Engines

These two search engines take different approaches to identifying relevant content. Google Scholar is a web crawler that provides citations of articles that have the named search terms in their titles, abstracts, and keywords. But for full-text articles, the user must go to the publication's website. ResearchGate is a membership application whose members can upload citations and full-text articles that are then available to other members for downloading. Guldin (2018) contrasted these two applications and the relative difficulties they provide to practicing foresters searching for scientific information to use in their daily work. ResearchGate provides greater likelihood for free access to full-text articles.

Analysis of the initial search results showed that two journals dominated the forest-related applied SAE niche: *Remote Sensing of Environment*[6] and *Remote Sensing*.[7] Therefore, the review dove deeply into their article archives.

TABLE 1 | Authors and co-authors of articles on small area estimation published in *Remote Sensing of Environment* since 2000.

Author/Co-author	Number of Publications	
	2010–2021	2000–2009
McRoberts (USA)	8	5
Tomppo (Finland)	3	6
Magnussen (Canada)	2	2
Astrup (Norway)	3	0
Breidenbach (Norway)	3	0
Finley (USA)	2	1
Katila (Finland)	0	3
Chirici (Italy)	2	0
Næsset (Norway)	2	0
Rahlf (Norway)	2	0
Ståhl (Sweden)	1	1
Stehman (USA)	0	2
Waser (Switzerland)	2	0
Authors Mentioned Once	28	13
Total authors/Co-authors	**57**	**33**

Remote Sensing of Environment

Leading Authors

The query of this journal's database yielded 12 articles from 2010 to May 2021 that had a total of 57 coauthors and an additional 13 articles from 2000 to 2009 that had 335 coauthors. Several researchers were authors or coauthors on multiple publications, **Table 1**.

Looking at all 25 articles since 2000, McRoberts was an author or coauthor on half of them (sole author on 5, lead author on 2, and coauthor on 6). Leading authors from other countries on three or more articles included Erkki Tomppo (Finland, 9), Steen Magnussen (Canada, 4), Rasmus Astrup (Norway 3), Johannes Breidenbach (Norway, 3), Andrew Finley (USA 3), and Matti Katila (Finland 3).

The data illustrate that although the number of publications in the two time periods were roughly equivalent (12 from 2000 to 2009 vs. 13 from 2010 to 2021), many more researchers have been involved as coauthors in the latter period. This highlights the recent growth in interest and a broadening of the talent studying this issue.

From a networking perspective, Tomppo and McRoberts were central figures. Tomppo's work began earlier, and his seminal contributions were recognized with the Marcus Wallenberg Prize in 1997.[8] Tomppo and McRoberts coauthored six articles

[6]Published by Elsevier (https://www.journals.elsevier.com/remote-sensing-of-environment).

[7]Published by MDPI (https://www.mdpi.com/journal/remotesensing).

[8]His citation read, in part: "*… his unique method of integrating available information sources into one system that is reliable and also allows accurate estimates*

together (2007, twice in 2009, 2011, and twice in 2016). McRoberts coauthored two articles with Magnussen, two with Gherardo Chirici, and single articles with Breidenbach, Astrup, and Finley. Beyond publishing with McRoberts, Tomppo published three times in this journal with Katila, twice in this journal with Magnussen, Chirici, and Waser, and once with 12 others (all but one from Europe).

Leading Topics

The primary foci of these articles were to combine NFI field plot data with remotely sensed data from satellite-borne sensors [e.g., LANDSAT's Thematic Mapper (TM)] and aircraft-borne passive sensors (both panchromatic and infrared digital cameras) or active sensors (e.g., LiDAR). Through non-parametric (primarily) or parametric (secondarily) methods, detailed forest attributes in the field plot data were imputed to pixels in the remotely sensed data. Then the imputed pixel attributes were aggregated to estimate forest conditions (e.g., tree cover area, forest type) for small areas with too few field plots to make estimates with acceptable RMSEs without the sensor-based imputed information. Satellite-borne sensor data were used in 10 of the papers; aerial photography and LiDAR once each.

The ***k*-Nearest-Neighbor** (*k*-NN) algorithm was used in 6 of the 24 papers. The *k*-NN approach is based on similarity in the space of the selected auxiliary variables to impute a value to a pixel when a value is missing. The plot observations and the image pixels' spectral values for the field inventory plots are the "training data set." An image pixel that isn't associated with a field inventory plot is assigned a value based on how closely its pixel spectrum resembles the spectrum of pixels in the training data set for plot locations in in the space of the auxiliary variables. In summary, every pixel in an image with missing values—pixels not associated with known forest inventory plots—can be assigned a value by finding its closest neighbors whose spectra closely resemble it and imputing a weighted mean of its nearest neighbors to it. Most of the early publications about *k*-NN were from the 2006–2010 era. Tomppo, McRoberts and/or Magnussen were the lead author or coauthors in all six papers. The most recent paper in the journal on *k*-NN was a review paper by Chirici et al. (2016).

A key factor in using the *k*-NN approach is the geospatial accuracy of the field plot centers/perimeter coordinates *vis a vis* the geospatial coordinates recorded by the sensor. This aspect was examined in several papers.

The variables whose values were most frequently imputed were the volumes of timber or growing stock and the area and types of forest cover. Estimating changes in cover—types of changes, their rates, and intensities—were discussed in two articles.

For 2010 to 2019, satellite or aerial photography were the auxiliary data used in eight of the 13 papers to study various forest attributes, including timber volume estimates (twice), tree cover (twice) and land use change. The precision of various models and estimators was examined in eight of the papers.

Remote Sensing

The second leading journal was the open-access MDPI journal *Remote Sensing*. This journal began publication in 2009. It classifies articles and special issues by broad sections, one of which is "Forest Remote Sensing." In that section, 389 articles have been published since 2009, including articles in 37 special issues related to the section (https://www.mdpi.com/journal/remotesensing/sections/Forest_Remote_Sensing). Sixty-two of the articles in the category Forest Remote Sensing have "forest inventory" as a key word. Only one mentioned small area/domain estimation.

Latifi and Heurich (2019) edited a special issue of 10 papers titled, "*Remote Sensing Based Forest Inventories from Landscape to Global Scale*." Two of the papers discussed SAE-related questions. Durante et al. (2019) focused on a 2.8-million-acre region in southwestern Spain, combining Spanish NFI field plot data, high-precision airborne laser scans (ALS), and bio-geophysical spectral variables from MODIS.[9] Novo-Fernández et al. (2019) described estimation procedures that combined Spanish NFI data and ALS data to predict growing stock volume for three major commercial tree species growing in northwestern Spain. Details for both these papers are discussed further in the section on Spain's NFI, below. Hill et al. (2018) reported a case study from northwestern Germany, discussed further in the section on Germany's NFI, below.

Other papers having "forest inventory" as a key word had limited relevance to the SAE issue. In general, the articles tested ways of using NFI data to improve estimates from airborne or terrestrial LiDAR point clouds and data from various passive sensors. In many cases, the NFI field plot data were used to either demonstrate the utility of new sensors or sensor-platforms (e.g., small UAVs aka "drones") or to improve various types of estimates made from the remotely sensed data. Some estimates were classification calls, such as forest cover type or forest vs. non-forest. Other estimates focused on stand characteristics, such as growing stock volume or above-ground biomass volume, or stand indices, such as leaf-area-index or normalized difference vegetation index. Some articles focused on individual tree characteristics, such as tree species identification. The aim of improving classification algorithms or stand estimates based on NFI data was to make estimates and inferences with acceptable RMSEs for larger geographic areas—regions or countries—from wall-to-wall remotely sensed data or to create geospatial data layers or map products. But in general, the novelty of the research reported either arose from applying a technique developed elsewhere to a new landscape or from showing how NFI field data could be used to calibrate remotely sensed data and save time and resources in developing larger spatial scale products—the opposite of the SAE issue for which this paper was invited.

for smaller areas than the traditional field inventories. Tomppo's system considerably enhances the total information value of data sources used and also allows for ecological data to be effectively assessed. In the context of national forest assessment, it is now possible to obtain inventory data at the community and owner levels as well – which has previously not been possible without extensive field work." (http://www.metla.fi/tiedotteet/1997/wallenberg-eng.htm).

[9]Moderate Resolution Imaging Spectroradiometer sensor aboard the Terra and Aqua satellites. https://modis.gsfc.nasa.gov/about/.

Google Scholar Citations Since 2010

The first 200 "hits" returned by Google Scholar to the search terms "small area estim" and "forest" identified 21 articles published since 2010 within the scope of this review that were not published in either *Remote Sensing of Environment* or *Remote Sensing*. The *Canadian Journal of Forest Research* had one-third of them, the *Scandinavian Journal of Forest Research* and *Forest Ecology and Management* together had another third, and the last third were spread across six other journals.

The expanding impact of Nordic researchers in SAE is evident in the articles published in both the *Scandinavian Journal of Forest Research* and *Canadian Journal of Forest Research*. The former isn't surprising. The latter illustrates increased collaboration among Steen Magnussen (Canada) and researchers in Norway and Denmark. Magnussen et al. (2014) and Magnussen and Nord-Larsen (2020) illustrate the international collaboration on SAE that currently exists, with the lead author being from Canada and coauthors from Norway, Switzerland, and Denmark. The former article introduced five facets that can improve inference in SAE: (1) model groups; (2) test of area effects; (3) conditional EBLUPs,[10] (4) model selection; and (5) model averaging. Two contrasting case studies with data from the Swiss and Norwegian NFIs were used to demonstrate the five facets. The latter article used data from the Danish forest inventory to demonstrate how spatial model strata for post-stratification (e.g., for SAE) can be identified from design-based model-assisted inference with either lasso or finite mixture modeling methods.

Other articles published the past 3 years that illustrate the Nordic and Canadian collaboration include Rahlf et al. (2021), Strîmbu et al. (2021), Breidenbach et al. (2020), Astrup et al. (2019), and Haakana et al. (2019a,b). Rahlf et al. (2021) found that maps based on NFI data augmented by ALS data can be used in lieu of maps developed from forest management inventory (FMI) data to estimate timber volumes in mature spruce stands—potentially saving the cost of doing an FMI. Strîmbu et al. (2021) dealt with the issue of inconsistency that arises when one attempts to aggregate parameter estimates for SAEs to a larger domain and the sum differs from the directly estimated domain parameter. Breidenbach et al. (2020) used Sentinel-2[11] mosaics along with NFI data to model and map Norwegian conifer types. The models were then used to create species-specific range maps for smaller geographic areas, such as municipalities. Astrup et al. (2019) described how photogrammetric point cloud data were combined with NFI point cloud data to produce a 16 × 16 m raster map with selected modeled attributes that could be used in FMIs. The two articles by Haakana et al. (2019a,b) focused on using post-stratification as an alternative way to use auxiliary information to estimate parameters for municipalities from Finland's NFI data.

[10] EBLUP is an acronym for Empirical Best Linear Unbiased Predictor.

[11] Sentinel-2 is a European Space Agency mission of two polar-orbiting satellites monitoring variability in land surface conditions. Their wide swath width (290 km) and high revisit time supports monitoring of Earth's surface changes. https://sentinel.esa.int/web/sentinel/missions/sentinel-2.

ResearchGate Citations Since 2010

The first 50 citations returned during a search ("small area estima" and "forest") had 20 articles since 2010 and 6 from 2000 to 2009 that matched the sideboards of this review. Seventeen of the articles had been previously identified in searches of *Remote Sensing of Environment, Remote Sensing, Scandinavian Journal of Forest Research* and *Canadian Journal of Forest Research*. Of the other 24 articles outside the sideboards of this review, 18 dealt with research on North American forests and 6 with research about forests in Asia and Oceana.

Two articles providing overviews were found in the ResearchGate search; Kangas et al. (2018) and Jiang and Rao (2020). Kangas and her 10 coauthors reviewed the state of science in Nordic country NFIs and how remotely sensed data are being used to augment NFI and FMI data and reduce uncertainties in parameter estimates—nationally and sub-nationally. More importantly, the article also lays out a roadmap for future research and development work and proposes a common research program for the Nordic countries focused on six identified problem areas. Although not focused on forest inventories, the overview of SAE methods by Jiang and Rao (2020) is a good current synopsis of the current state of statistical research. It is a good entry point for readers desiring an organizing framework for the many different SAE methods.

DISCUSSION

Online searches using Google Scholar and ResearchGate for recent research pertaining to SAE of forest conditions in countries outside the United States shed some additional light on the state of science. There was a notable lag, often several years, between journal publication and when the search engines reported it. Besides *Remote Sensing of Environment* or *Remote Sensing*, three other journals have published a growing number of articles on SAE, notably the *Canadian* and *Scandinavian Journals of Forest Research* and *Forest Ecology and Management* (e.g., McRoberts, 2012; McRoberts et al., 2017). This suggests that researchers desiring to follow SAE advances should focus first on those five journals, before relying on broader search engines to find new international research on SAE pertaining to forests.

The emergence of journal policies to publish accepted journal on-line prior to articles appearing in printed volumes often results in two different years for citations. For example, Haakana et al. (2019b) was published on-line on 10 December 2019 but didn't appear in print until the April 2020 issue. The citations in this article use the on-line publication date rather than the in-print date.

Status of Small Area Estimation in National Forest Inventories: Global Overview

Barrett et al. (2016) summarized the operational use of remotely sensed data in NFIs, based on the responses of 45 countries' experts (representing 65% of global forest area) to a questionnaire

circulated as part of the COST Action FP1001.[12] They found that remotely sensed data—from many different sensors—were widely used to enhance estimates for many parameters but called for further research on ways to improve uncertainty estimation by better integrating remotely sensed data and field data.

Thirteen countries used spatial datasets (e.g., digital elevation models, soils, and geology maps, ownership data) from other sources beyond the NFI program to enhance forest map data layers that were constructed from satellite-borne sensors and field data. The most common map layers produced were land cover/land use, forest cover type, and species group maps.

Where insufficient field data were available, nine countries reported that imputation models were used to integrate other spatial datasets and available field data to predict ground attributes (from which maps can be created) and calculate local estimates. Many different prediction techniques were used. Eight countries used supervised parametric techniques (maximum likelihood, discriminant analysis, and linear, non-linear, and logistic regression). Seven countries used non-parametric methods (*k*-NN) and unsupervised approaches (neural networks, isodata, or k-means). But only 12 countries attempted to estimate uncertainty for map attributes they estimated or imputed, and only nine countries attempted to include uncertainty estimates for area statistics associated with forest area vs. non-forest area maps.

From the articles reviewed that were published in the last 5 years (2016 to 2021), it's apparent that the focus of forest-related SAE research outside the USA is shifting. Three major threads have emerged:

1. **Using model-based approaches and NFI data to make estimates at smaller spatial scales—provinces, forest management units, municipalities**. The focus is on demonstrating that existing well-funded and well-designed NFIs can supplant less-well funded FMIs for many regional/local purposes. While stand-level inventories still have their place in planning management activities, the NFI-based SAEs show promise in helping offset lack of support and funding for FMIs. In some cases, NFI-based models are being used to sort out priorities for stand/compartment exams.
2. **Increased focus on methods for reducing uncertainties of estimates and improving precision of estimates**. Several articles described using simulation approaches, based on NFI data, to explore alternative estimation procedures that reduce uncertainties/improve precision. Creating simulations is faster and less expensive than gathering plot data to test alternative SAE approaches.
3. **Estimating above ground biomass (AGB) volume has become a prominent parameter of interest for SAEs, eclipsing interest in estimating timber volumes**. Recently adopted global policies, such as greenhouse gas reporting and REDD+, have driven this shift. Public and governmental interest in understanding better the roles of forests in sequestering carbon and tracking forest carbon stocks and fluxes at the sub-national level have also played a role. Changes in the objectives of grant programs have been used to shift the focus of SAE research.

[12]COST is a European framework for improving ***Co***operation in ***S***cience and ***T***echnology. COST action FP1001 focused on improving information about potential supplies of wood.

Switzerland's National Forest Inventory: Overview of SAE Research Since 2010

Recent work done in Switzerland's NFI program was summarized by Pulkkinen and Zell (2019).

Background

The Swiss foundation on sampling theory for forest inventory rests upon the *design-based Monte-Carlo approach*, where sampling is carried out for an *infinite population* of the points within a region of interest (see Chapters 4 and 5 in Mandallaz, 2008). Target parameters are *spatial means* computed as tree-population totals (sums of tree characteristics over all the trees within the region) divided by the area of the region. The spatial means are equal to the expectations of the *local densities* of the target variables over the uniform distribution of points within the region. The local density of a target variable is defined at each point of the region as the ratio of the Horvitz-Thompson estimate of the tree population total of the variable, based on a probability sample of trees taken at the point, to the area of the region.

In the population of the points within the region, a *two-phase sampling* approach is typically used. First, many uniformly randomly located sample points are drawn independently of each other and auxiliary information is collected at these points. Second, a simple random sample from the first-phase points is used to locate field plots where data are collected on the target variables (i.e., their local densities). In a *three-phase sampling* approach, there are two nested phases for collecting auxiliary information before field plots are identified. In practice, points are located at the intersections of systematic grids whose starting point and orientation of the largest grid being considered are chosen randomly. Therefore, the variance estimators derived from the assumption of uniform random locations can be considered generally conservative. The Swiss NFI follows this design-based Monte Carlo paradigm with its *two-phase simple random sampling for post-stratification* estimation, where the auxiliary information is used to do the post-stratification.

The Swiss NFI is now in its fifth cycle. The methodology of the fourth cycle (2009–2017) was detailed by Fischer and Traub (2019). The 5th cycle continues the continuous mode adopted for the fourth cycle by systematically measuring 1/9th of the field plots over the whole country each year. The field plots are located on a 1.41 $\times$ 1.41 km grid, whereas the auxiliary information, mostly based in digital aerial images, is currently available on a 100 $\times$ 100 m super-grid.

For small-area estimation, the *design-based model-assisted approach* is used. Design-based inference relies on probability samples for validity and its estimators of population parameters are generally unbiased. Design-based inference relies on three assumptions: a probability sample incorporating some form of randomization is used; each population unit has one and only one possible value; and selection of population units into

the sample is based on positive and known probabilities of selection (McRoberts et al., 2019). Brewer (2013) equated *design-based* with *randomization-based* inference, and contrasted it to *model-based* inference, which he equated with *prediction-based* inference.

Model-assisted means that a model is used to support estimation following probability sampling (Ståhl et al., 2016). A prominent example of model-assisted survey sampling in a forest inventory context was highlighted by Gregoire et al. (2016), who described how the point cloud of airborne LiDAR height measurements can be linked by statistical regression to estimates of forest biomass from a ground sample of forested plots whose trees have their heights measured and above-ground biomass volumes computed from equations. Pulkkinen and Zell (2019) reported that end-users and stakeholders in the Swiss NFI prefer the design-based, model-assisted approach, albeit using digital aerial photography rather than airborne LiDAR, over a model-based/model-dependent approach.

Development of New Design-Based Model-Assisted Small Area Estimators

At ETH Zurich, Mandallaz and his students have developed new design-based model-assisted estimators that involve: (i) *non-exhaustive auxiliary data* (auxiliary data not available wall-to-wall but coming from a sample); and (ii) *external models* of any type (models constructed in the data that are entirely independent of the current field plot data of the small area) or *internal linear models* (linear models fitted to the data containing the current field plot data of the small-area). The estimators consist of the means of the: (i) model predictions over the auxiliary data points (null/first-phase sample); and (ii) model residuals over the field plots (second-phase sample) within the area. When using an internal model, the idea is to fit a model "globally" in a large region containing the area of interest, and to apply it "locally" to the specific area. In this case, like in the classical regression estimators developed for finite populations, the variation of the model parameter estimates over (hypothetical) repeated samples is considered in the variance estimators of the new small-area estimators. Further, the uncertainty due to employing estimated auxiliary variable means instead of (unavailable) true means is incorporated in the variance estimators. The researchers also present an approach where the internal linear model is extended with the indicator variable(s) of the area(s), thus eliminating the residual-dependent part of the estimator, which greatly simplifies the calculation of the variance.

Mandallaz (2013) introduced the new small-area estimators for two-phase sampling, with both exhaustive and non-exhaustive auxiliary data and including the special case of cluster sampling. He illustrated the estimators with a small case study and with a simple simulation example. Mandallaz et al. (2013) completed this work by presenting the estimators for the case where some auxiliary variables are available exhaustively (wall-to-wall) and others non-exhaustively. They tested the estimators with a simulation example like the one in the earlier paper but with a larger case study using data from the Swiss NFI. Hill et al. (2018) applied the estimators for timber volume estimation in forest management units of two levels (forest districts and sub-districts) in the German state of Rhineland-Palatinate using data from the German NFI.

Mandallaz (2014) extended the two-phase small-area estimators with partially exhaustive auxiliary data to three-phase sampling, where the auxiliary variable values come from nested null- and first-phase samples. He illustrated the estimators with a simulation example like those in the earlier papers.

Massey et al. (2014) applied these three-phase estimators to estimate timber volumes in the five production regions (summing up to the entire country) of the fourth Swiss NFI, when only three annual panels (out of nine) of field plot data were available. The reduced second-phase sample size was compensated by using the full field plot data from the third inventory as the first-phase auxiliary data, in addition to the usual aerial photography used as the null-phase auxiliary data. Steinmann et al. (2013) applied the two-phase synthetic and difference estimators (involving external models) for forest area and timber volume estimation in the Swiss canton of Aargau using data from the Swiss NFI. Two doctoral theses have resulted from this research (Massey, 2015; Hill, 2018). Hill et al. (2021) have implemented the estimators discussed above in the R package *forestinventory*.

Construction of SAE System for Swiss NFI

In an ongoing project, Pulkkinen, Lanz, and Zell are developing an operational system for producing estimates of several target parameters for small areas/domains in the Swiss NFI. Auxiliary information comes from several sources, the most important being a vegetation height model estimated from a digital elevation model of tree canopy height and a LiDAR-based terrain elevation model. The small-area estimators included in the system are the design-based model-assisted estimators discussed above, with (i) internal linear models or (ii) external models of any type, and with estimated auxiliary variable means. When internal models are used, they are built/fitted separately for each small area/domain. Currently, the system estimates forest area, total growing-stock volume, and total growing-stock biomass above ground for the cantons, forest districts and municipalities in Switzerland.

Norway's National Forest Inventory: Small Area Estimation on Multiple Scales

Breidenbach et al. (2019) presented the status of SAE research and use in the Norwegian NFI. national forest inventory field plot inventory data are combined with 3D remotely sensed data to estimate forest characteristics at different spatial scales. ALS and image matching are currently used as auxiliary information to create the NFI's forest resource map SR16, a raster map with a pixel size of 16 × 16 m (Astrup et al., 2019). While model-dependent methods were used on the scale of pixels and forest stands (Breidenbach et al., 2015), model-assisted estimators were used on the scale of municipalities and larger area of interests (Breidenbach and Astrup, 2012).

Developing Forest Resource Map SR16

Development of SR16 tested new methods for using ALS to make stand-level estimates and comparing those estimates with

independent data from a FMI. Astrup et al. (2019) described the development and utility of the SR16 in greater detail. They used photogrammetric point cloud data with ground plots from the Norwegian NFI. First, an existing forest mask was updated using object-based image analysis methods. Within the updated forest mask, a 16×16 m raster map was developed with Lorey's height (hL),[13] volume, biomass, and tree species as attributes. All attributes were predicted with generalized linear models that explained about 70% of the observed variation and had relative RMSEs of about 50%. The raster map was then segmented into stand-like polygons that internally were relatively homogenous with respect to tree species, volume, site index, and hL. When SR16 was used as auxiliary information to NFI field plot data and a model-assisted estimator, the precision was on average 2–3 times greater than estimates based on field data only. In conclusion, SR16 was useful for improved estimates from the Norwegian NFI at various scales. The mapped products may be useful as additional information in forest management Inventories (FMIs).

Applying SR16 to Small Area Estimation

One of the biggest challenges for the Norwegian NFI is satisfying the interests of stakeholders in forest attribute information for small sub-populations, such as municipalities or protected areas (Breidenbach and Astrup, 2012). Auxiliary information that is correlated with attributes of interest can improve the precision of estimates. Two examples have been recently reported in the literature. In the first one, Breidenbach and Astrup (2012) used the height and volume information in SR16 to improve the estimates of mean above-ground biomass for small areas. In the second (Breidenbach et al., 2019), ALS and SR16 data layers were used to improve the precision of information for FMIs. FMI data required local adjustments to obtain the desired precision. Mixed-effects models were fit, using fine-scale ALS data. SR16 data layers used to make SAEs were compared to FMI stand-level estimates. The RMSD between FMI and SR16 estimates of timber volume on stand-level ranged between 11 and 17%. While no systematic deviation was visible for stands in mature pine forest types, SR16 data underestimated timber volume in mature spruce forests by 12%, especially in ALS projects where the NFI data did not cover the full range of explanatory variables. They concluded that the accuracy of SR16 map data layers may be sufficient for most small-scale forest owners and for some strata for larger forest enterprises. Accuracy can be improved, and systematic errors removed by integrating auxiliary information where a limited number of NFI plots do not cover the range of explanatory variables within an ALS coverage area.

[13] Lorey's height (hL) is a mean height estimate that is weighted by basal area, which allows the larger trees to contribute more to the mean. It is a commonly used mean height estimator outside the USA. Lorey's height is computed as the sum of tree height multiplied by tree basal area for all trees, divided by the basal area of the stand. Because variable radius plot sampling (Bitterlich or prism sampling) selects trees proportional to their basal area, the mean height of trees included in one or more prism sample counts gives an estimate of hL.

Germany/s National Forest Inventory: Small Area Estimation at the District Level

Hill et al. (2018) described a double-sampling extension of the German NFI to make design-based SAE at the forest district level. They used an ALS-estimated canopy height model and a tree species classification map based on satellite data as auxiliary data with a regression model to produce timber volume predictions.

The German NFI is based on a nationwide 4 × 4 km grid. But some states (Rhineland-Palatinate in Hill et al., 2018) have intensified the sample to a 2 × 2 km grid. At each grid point, field crews collect data from a cluster of four sample plots, arranged in a square with 150 m sides. The number of actual plots measured in a cluster can vary between one and four depending on the forest/non-forest decisions made by the crew. At each sample point, trees to tally are identified using a BAF 4 m^2/ha prism/relascope, and included if their DBH is >7 cm.

Wagner et al. (2017) used SAE methods to estimate spruce timber reserves in the Rhineland-Palatinate's forest districts. The state forest inventory and an ALS-based canopy height model provided the data. A new spline-based SAE method was proposed. It provided stable estimates that met specialized constraints. Results were compared with existing spruce timber estimates.

Rationale for SAE Research

Rhineland-Palatinate is one of the two most densely forested German states, with 8,400 km^2 of forest comprising 42% of the land area. Two characteristics dominate Rhineland-Palatinate forests. Mixed forest stands dominate (82% of the forest area). Public ownership dominates private ownership—27% are state-owned forests and 46% are municipally owned vs. 27% that are privately owned. The state forest agency has a mandate to sustainably manage state and municipal forests, including planning, harvesting, and selling wood. Therefore, the state has been further sub-divided into 45 districts (averaging 43,777 ha), and 405 sub-districts (averaging 4,624 ha). A key question for the state agency is where and how to gather information suitable for managing at the state, district, and sub-district levels. While the NFI information is helpful at the state level, estimators at the district and sub-district level derived from the NFI have unacceptable RMSEs for planning and implementing management activities. Many states have solved this problem by establishing forest district-level inventories (FDI) with much greater sampling intensities than the NFI (e.g., the quadruple intensification of NFI in Rhineland-Palatinate). But FDIs are costly, and many states are facing increasing restrictions on budgets and personnel. Therefore, states are seeking more cost-efficient inventory methods, among them SAE methods.

Researchers from ETH Zurich and the Rhineland-Palatinate State Forest Service partnered to test SAE approaches for cost-efficiency. They considered three types of design-based regression estimators suggested by Mandallaz (2013) and Mandallaz et al. (2013): Pseudo-small, extended pseudo-synthetic, and pseudo-synthetic. Auxiliary data were a canopy height model from nationwide ALS and a tree species classification map to be used for regression estimation within tree species strata. A double-sampling approach was used, for five reasons discussed in detail

in Hill et al. (2018). At the district level, results showed that both the pseudo-small and the extended pseudo-synthetic estimators led to substantial reductions in estimation error compared to the standard one-phase estimator. But the sub-district level was too small geographically and had too few sampling points to achieve the same estimated error reduction as at the district level. However, estimation errors at the sub-district level were still smaller than the standard one-phase estimator (20 vs. 40%). But the authors acknowledged that further research is needed to determine whether the achieved reductions in error levels are enough to support forest planning decisions.

A complicating factor in the case study was that the ALS data were of various ages—some relatively recent, others a decade old. Beyond the obvious issue of tree height growth over a decade, a tougher challenge was that laser sensor technology advanced rapidly over the decade, resulting in much denser point clouds for recent years compared to older years. As older scans are replaced by newer scans, the power of the auxiliary information will improve—both in terms of consistent tree canopy height models across the landscape and within the tree species strata.

The methods introduced by Hill et al. (2018) are being tested on NFI data from other states. In 2019, data from Thuringia were being tested and data from Mecklenburg-Pomerania was next. The intent is to have these new estimation features operational by 2023, at the latest, after completion of the 2021/2022 inventory cycle.[14]

Finland's National Forest Inventory: Efficiency of Post-stratification for Small Area Estimation

Tomppo (1990, 1991) was the global pioneer in combining NFI data, satellite data, and *k*-NN for making estimates. His research formed the intellectual foundation for Finland's NFI, and for SAE work in many other countries, including the United States.

Haakana et al. (2019a,b) reported on recent research in Finland, using southern Finland provinces and municipalities within provinces as test regions both for making point estimates (e.g., growing stock volume by tree species groups) and evaluating variances estimated by alternative methods. They found that post-stratification, based on remotely sensed data, even if old and incomplete, improves efficiency in estimating selected variables at the provincial and smaller municipality levels when compared to results from making estimates using only current NFI data. Work by Tomppo, McRoberts, and Magnussen was extensively cited.

The two papers explored several options for obtaining auxiliary information to use in **post-stratification**. Sweden was cited as an example where official statistics are based solely on field plots, but estimates are developed using design-based post-stratification, based on *k*-NN maps or other map products.

Haakana et al. (2019a) presented a case study on estimating growing stock volumes by tree species groups. The auxiliary information was derived from NFI volume maps available for provinces in southern Finland. These maps were developed from the data gathered in the previous NFI iteration (2005–2008) and LANDSAT 5 TM imagery from 2007. Full-coverage raster maps with 20 m pixels were created by combining satellite images, digital map data, and NFI sample plot data and then using the *k*-NN method to estimate growing stock volume, by species group, for each pixel in a forest land mask. Procedures described by Tomppo et al. (2012) were used.

One of the challenges discussed was the use of older volume maps from the prior NFI iteration combined with older LANDSAT 5 TM data. The primary reason for using older maps based on older remotely sensed data was to use independent auxiliary data. But during the intervening time, many forest management activities, such as thinnings and final harvests, occurred, which reduced the correlation between the older auxiliary data and the current NFI data. But the reduced correlation and potential reduction in estimation efficiency weren't quantified—just recognized—because updating the prior information was thought too costly for the project.

Post-stratification by mean volume improved the precision of both area and volume estimates for forest area and growing stock volume compared to using NFI data alone. Relative efficiencies ranged between 2.3 and 3.5. As expected, post-stratification resulted in a smaller decrease in mean relative standard error for the smaller areas than for the larger areas. This result held both for the forest area variable as well as for total growing stock volume and volume by tree species stratum (pine, spruce, birch, and other deciduous strata). Further, the small area estimates from post-stratification were robust compared to the field plot data estimates because the largest variances improved more than the average variances.

Haakana et al. (2019b) acknowledged that the *k*-NN method can provide a model-based estimator for small geographic areas, but not a designed-unbiased estimator for RMSE. Thus, in this article they focused specifically on municipalities to explore the lower limits in geographic size that could still yield estimates of forest area and growing stock volume with adequate precision. They explored the differences in estimation efficiency and error estimates for various sizes of areas—ranging from 5,700 to 921,600 ha—made possible by post-stratification.

The major conclusions of the two articles were that: (1) utilizing old forest resources maps in a fully operational approach for national level estimation improved estimates; and (2) although post-stratification enabled forest area and growing stock to be estimated more accurately for much smaller geographic areas than with field plot data alone, post-stratification should be limited to the smallest municipalities where model-based estimation is still needed. Haakana et al. (2019a) acknowledged that precision could be further improved by updating maps to account for thinning, regeneration cuttings, and final harvests; segmenting maps and remotely sensed data into homogenous segments; and by having improved boundary files for municipality land use classes. But overall, these opportunities didn't detract from the overall results.

Katila and Heikkinen (2020) reviewed the time-series of *k*-NN estimates over two decades for municipalities, based in NFI data. Their interest was in the variation among estimates from different time periods—exceeding 10% in mean volume—which they believed indicated a systematic error in SAEs. They combined

[14]Personal communication with Dr. Sebastian Schnell, Thünen Institute of Forest Ecosystems, Eberswalde, Germany.

NFI estimates from three points in time—2011, 2013, and 2015—and found that multi-temporal data fusion made small but consistent improvements in the estimates.

France's National Forest Inventory: Three-Dimensional Auxiliary Data

Since 2005, the French NFI has used a two-phase sampling design on a 1 km grid. Each year, one-tenth of the plots are photo-interpreted for land use and land cover in the first phase. Then in a second phase, a sample is drawn as a function of land cover types, resulting in about 6,500 plots being measured each year nationwide.

Vega et al. (2021) introduced a new estimation algorithm to balance between statistical precision and spatial scale. The algorithm identifies the smallest possible groups of domains satisfying prescribed sampling density and estimation error. The research used NFI data from oak-dominated areas in the Sologne and Orléans areas of central France, covering 157 municipalities of varying sizes. Auxiliary data were a national forest cover type map, a canopy height model from digital aerial photographs, and LANDSAT imagery. The algorithm depends on the statistical strength between the field attributes—growing stock volume and basal area in this case—and on auxiliary variables and the spatial heterogeneity of the forests. Results illustrate the balance between desired precision and the spatial scale required to attain that precision in attribute estimates.

Fortin (2020) explored the problem caused by annually sampling only a portion of the NFI population of plots and the impact on variance of point estimates for a geographic area when plot data are from 1 to 10 years old. Fortin proposed overcoming the difference in time since last remeasurement by using an individual tree forest growth model (MATHILDE) to update older plot measurements to account for growth since last measurement. But this seemingly simple solution leads to a hybrid inferential model where uncertainty arises not only from the sample design but also from the growth model used to update measurements (Kangas et al., 2019). Fortin tested the updating approach on French NFI data from the Lorraine region and concluded that under certain conditions, using a forest growth model can increase the precision of inventory estimates.

Irulappa-Pillai-Vijayakumar et al. (2019) used three-dimensional (3D) variables from photogrammetric-estimated canopy height models, a forest type map, vegetation indices, and LANDSAT 8 spectral bands as auxiliary data to lend strength to French NFI data for a 733,500-ha region in central France that is 48% forested. Adding complexity was the fact that much of the forest in the region was a mixed broadleaved species cover type that was more diverse in species composition and therefore in form, structure, and fragmentation than the typical conifer forest. The objective of the research was to test whether multivariate *k*-NN imputations could improve the precision of estimates for 11 forest attributes beyond the precision based solely on NFI data.

The NFI data came from 755 plots measured from 2010 to 2014. Irulappa-Pillai-Vijayakumar et al. (2019) goes into considerable detail about the significant effort invested and difficulties encountered in: (1) transforming digital aerial photography and ALS into 3D digital terrain models that could be used to estimate canopy height models for two different time periods (2008 and 2014); and (2) using the estimated changes in height between 2008 and 2014 to estimate changes in other forest attributes, such as stand density, basal area and several different types of volume. Finally, auxiliary data for all 11 variables were converted to a spatial resolution of 30 m to conform with the spatial resolution of the TM sensor aboard LANDSAT.

Results were that volume attributes had the greatest reduction in errors. Using 3D change estimates contributed to the substantial increase in precision and improved neighbor selection within the *k*-NN method. The authors reported that these results open possibilities for improving forest attribute estimation for smaller areas. Their downscaling work continues.

Spain's National Forest Inventory: Integrating NFI Field Plot Data With Airborne LiDAR Data

Four recent papers have discussed research in Spain, including Condés and McRoberts (2017), Esteban et al. (2019), Durante et al. (2019), and Novo-Fernández et al. (2019).

Condés and McRoberts (2017) reported new methods to update NFI-based estimates when the year of the most recent NFI survey doesn't match the required year for international reporting requirements. Their main aim was to develop an unbiased method to update NFI estimates of mean growing stock volume (m^3/ha) using models to predict annual plot-level volume change, and to estimate the associated uncertainties. Because the final large area volume estimates were based on plot-level model predictions rather than field observations, hybrid inference was necessary to accommodate both model prediction uncertainty and sampling variation. Specific objectives were to compare modeling approaches, to assess the utility of Landsat data for increasing model prediction accuracy, to select the most accurate method, and to compare model-based and design-based uncertainty components. For four forest types, data from the 2nd and 3rd Spanish NFI surveys together with site variables and Landsat imagery were used to construct models to predict NFI information for the year of the 4th NFI survey. Data from the 3rd and 4th surveys were used to assess the accuracy of the model predictions at both plot-level and large area spatial scales. The most accurate method used a set of three models: one to predict the probability of volume removals, one to predict the amount of volume removed, and one to predict gross annual volume. Incorporation of Landsat-based variables in the models increased prediction accuracy. Differences between large area estimates based on plot-level field observations for the 4th NFI survey and estimates based on the model predictions were minimal for all four forest types. Further, the standard errors of the estimates based on the model predictions were only slightly greater than standard errors based on the field observations. Thus, model predictions of plot-level growing stock volume based on field and satellite image data as auxiliary information can be used to update large area NFI estimates for reporting years for which spectral data are available, but field observations are not.

Esteban et al. (2019) described an approach to model-assisted inference, using a random forests (RF) approach. RF has recently emerged as a popular approach because it's able to select and rank many predictor variables and it relies on an ensemble of trees as a strategy to improve model robustness. Random forest consists of a combination of decision trees where each decision tree contributes a single prediction for each population unit with the final prediction for each unit calculated as the mean over the RF decision tree predictions. Although RF has been used by others, little literature is available on model-based mean square error (MSE) estimation for population parameters with this algorithm. The study had three objectives:

1. Construct RF models to predict response variables (volume and above-ground biomass) and changes in the response variables for population units (ALS cells);
2. Compare multiple bootstrap estimators of the model based MSE of the estimate of the population mean; and
3. Construct change maps and change uncertainty maps.

Two study areas were used, one in the La Rioja region of Spain and the other in Våler municipality, southeastern Norway. The Spanish data included plot-level volume datasets acquired at different times for different plots as well as corresponding multi-temporal ALS (2010–2016) and multi-spectral data. The Norwegian data included plot-level biomass datasets for two times for the same plots and temporally consistent ALS data (1999, 2010). The authors concluded that RF models adequately described the relationship between field plot measurements of volume and biomass per unit area and remotely sensed data. They also found that model-assisted and model-based estimators based on RF predictions produced similar estimates of population means and change estimates and smaller MSEs than expansion estimators. Some insights into two bootstrapping approaches were provided too.

Durante et al. (2019) focused on a 2.8-million-acre region in Spain, combining Spanish NFI field plot data, fine-precision ALS, and bio-geophysical spectral variables from MODIS. The novelty of the study was testing a two-stage upscaling approach where above-ground biomass estimates from ALS data were first calibrated with NFI field plot data from 242 NFI field plots, then used to train a machine-learning method that could be applied to MODIS-estimated indices and topographic factors to develop wall-to-wall maps of above-ground biomass for the region. In one sense, this is the reverse of usual SAE approaches, borrowing strength from the NFI field plot data to improve biomass estimates made from laser point clouds and then link the improved biomass estimates to MODIS data to create a regional map. The authors again highlighted the difficulties created by lack of precision in field plot center coordinates (5–15 m nominal accuracy) compared to ALS data. The biomass model was based on four types of information: (1) 2 m resolution ALS data; (2) sketches of field plot layout; (3) high resolution ortho-imagery from the Spanish National Plan for Aerial Orthophotography; and (4) total height, species type, and location of each tree in the field. Earlier work in case studies in western Finland by Maltamo et al. (2009) and Norway by Nelson et al. (2012) were cited.

Novo-Fernández et al. (2019) carried out similar work in northwestern Spain. The area studied contains forest plantations that contribute 58% of the annual national timber harvest and thus are important to commercial enterprises producing panelboard, sawn lumber, and pulpwood. Dominant species are *Eucalyptus globulus* Labill, *Pinus pinast*er Ait., and *Pinus radiata* D. Don. Therefore, the main objective of this study was to generate a fine-resolution raster database with information about key forest yield variables such as total over bark volume (m^3/ha) and total aboveground biomass (t/ha), by species. Secondary objectives—necessary to achieve the first objective—included: (1) development of a procedure to harmonize the Spanish NFI and ALS data; (2) selection of the best empirical models of relationships between field measures and ALS-derived metrics, by comparing a parametric machine learning technique (multiple linear regression) and several well-known non-parametric techniques; and (3) to estimate spatially-continuous maps of yield variables. The same methodology has been used in Austria, Denmark, Sweden.

SUMMARY

Small Domain Estimation research in the forest sector has focused almost entirely on *spatial* domains to the exclusion of other domains, hence the term SAE has replaced SDE in the forest inventory literature. SAE research and applications are underway in many European countries to improve estimates—reduce the RMSE or confidence intervals—of forest attributes based on sample data collected on NFI field plots.

Airborne LiDAR data are becoming increasingly popular as auxiliary data, especially where country-wide laser scanning has replaced country-wide aerial photography as the raw data for national topographical mapping, transportation, or other agencies.

Design-based model-assisted SAE inference methods are being used in several countries, but pure model-based or hybrid inference methods are also being explored. Each methodology has advantages in specific situations. Regardless of methodology, the *k*-NN approach is the current "standard," although with tweaks here and there. Various two-stage or double-sampling approaches are popular for post-stratification.

The intellectual leadership in SAE research in the forest sector is broadening. In the 1990s and first decade of the twenty-first century, only a few researchers—notably Erkki Tomppo and Ronald McRoberts—had published more than two articles on the topic. Since 2010, the forest SAE literature documents increased trans-national collaboration by many more authors and coauthors in advancing the use SAE. This review found that the influence of Daniel Mandallaz and his students from ETH-Zurich is growing.

McRoberts pointed out[15] that as scientific disciplines mature, they inevitably move through a three-step sequence of phases:

1. Descriptive studies (estimating means and variances);
2. Predictive studies (creating models and maps); to

[15]Personal communication, 17 Nov 2019.

3. Inferential studies (determining confidence intervals for population parameters, testing of hypotheses).

National forest inventories moved early into the inferential phase, but remote sensing is just now moving into that phase. The point is that as remote sensing research and applications in the forest sector mature, they will inevitably become more rigorous statistically and characterized by use of more sophisticated statistical techniques. Greater statistical sophistication will of necessity entail greater attention to uncertainty issues in estimates, models, and predictions. There is some evidence from this review of international research that this progression of phases is occurring; as it also is currently progressing in the United States.

The increase since 2010 in research and applications of SAE methods is being driven largely by NFI stakeholders' needs for information about forests at sub-national and sub-state/sub-province spatial scales. Two driving forces stood out in the literature reviewed. First, the costs—financial and staffing—of forest management unit inventories are pinching the budgets of state/provincial forest managers. They are searching for cost-cutting measures and using NFI data to make more precise estimates at sub-national levels is emerging as a viable solution.

Second, international carbon-accounting reporting expectations are growing. Carbon stocks and fluxes now need to be estimated from spatially and species-specific forest inventory data rather than simply applying broad-based, generalized, per-area carbon estimates to forest cover type area estimates. Hence, greater emphasis on obtaining above-ground forest biomass estimates specific to forest-cover-types or species for discrete regions of a country (accounting for site differences, such as soils, geology, topography, and land use patterns). If smoothly functioning markets emerge that compensate forest landowners for carbon sequestered in forests, agencies, and landowners want to have site- and species-specific, statistically reliable information available to support their payment contracts.

Three final points. None of the recent literature reviewed cited the need for or use of SAE estimates to satisfy forest certification criteria. Second, as interest in forest carbon markets continues to grow, it will be interesting to see if SAE estimates of forest carbon stocks and fluxes become acceptable to market investors. The emergence of interest in estimating AGB portends this issue. Third, although this paper focused on forest-sector literature dominated by authors and applications outside the United States, there is much SAE underway inside and outside the forest sector within the United States. Researchers for the U.S. Census Bureau (2021) can provide useful entrees to SAE outside the forest sector, just as similar SAE research outside the forest sector is reported by the European Union and national statistical agencies of individual European countries.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

The manuscript was prepared using my company's funding.

ACKNOWLEDGMENTS

The author thanks Ronald McRoberts for reviewing the original paper and providing many helpful technical revisions and several additional citations from recent European studies. Whatever errors remain is solely the responsibility of the author.

REFERENCES

Astrup, R., Rahlf, J., Bjørkelo, K., Debella-Gilo, M., Gjertsen, A., and Breidenbach, J. (2019). Forest information at multiple scales: development, evaluation, and application of the Norwegian Forest Resources Map SR16. *Scand. J. For. Res.* 34, 484–496. doi: 10.1080/02827581.2019.1588989

Barrett, F., McRoberts, R. E., Tomppo, E., Cienciala, E., and Waser, L. (2016). A questionnaire-based review of the operational use of remotely sensed data by national forest inventories. *Rem. Sen. Environ*. 174, 279–289. doi: 10.1016/j.rse.2015.08.029

Bechtold, W. A., and Patterson, P. L. (eds.). (2005). *The Enhanced Forest Inventory and Analysis Program—National Sampling Design and Estimation Procedures*. Gen. Tech. Rept. SRS-80. Asheville, NC: USDA Forest Service. Southern Research Station. p. 85.

Brackstone, G. J. (1987). "Small area data: policy issues and technical challenges," in *Small Area Statistics*, eds R. Platek, J. N. K. Rao, C. E. Särndal, and M. P. Singh (New York, NY: Wiley), 3–20.

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

Breidenbach, J., McRoberts, R. E., and Astrup, R. (2015). Empirical coverage of model-based estimators for remote sensing assisted estimation of stand-level timber volume. *Rem. Sens. Environ*. 173, 274–281. doi: 10.1016/j.rse.2015.07.026

Breidenbach, J., Rahlf, J., Hauglin, M., and Astrup, R. (2019). "Small area estimation on multiple scale—with a focus on stand-level estimates," *Presentation at: A Century of National Forest Inventories: Informing Past, Present, and Future Decisions* (Sundvolden: Norsk Institutt For Bioøkonomi (NIBIO)).

Breidenbach, J., Waser, L., Debella-Gilo, M., Schumacher, J., Rahlf, J., Hauglin, M., et al. (2020). National mapping and estimation of forest area by dominant tree species using Sentinal-2 data. *Can. J. For. Res*. 51, 365–379. doi: 10.1139/cjfr-2020-0170

Brewer, K. (2013). Three controversies in the history of survey sampling. 12-001-X. *Surv. Methodol.* 39, 249–262. Available online at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2013002/article/11883-eng.htm (accessed June 18, 2021).

Canadian Centre for Remote Sensing (CCRS). (2019). *Fundamentals of Remote Sensing*. Ottawa, ON: Natural Resources Canada. p. 258. Available online at: https://www.nrcan.gc.ca/maps-tools-publications/satellite-imagery-air-photos/tutorial-fundamentals-remote-sensing/9309 (accessed June 18, 2021).

Chirici, G., Mura, M., McInerney, D., Py, N., Tomppo, E., Waser, L., et al. (2016). A meta-analysis and review of the literature on the K-nearest neighbors technique for forestry applications that use remotely

sensed data. *Rem. Sens. Environ*. 176, 282–294. doi: 10.1016/j.rse.2016.02.001

Condés, S., and McRoberts, R. E. (2017). Updating national forest inventory estimates of growing stock volume using hybrid inference. *For. Ecol. Manage.* 400, 48–57. doi: 10.1016/j.foreco.2017.04.046

Durante, P., Martín-Alcón, S., Gil-Tena, A., Algeet, N., Tomé, J., Recuero, L., et al. (2019). Improving aboveground forest biomass maps: from high-resolution to national scale. *Rem. Sens*. 11:795. doi: 10.3390/rs11070795

Esteban, J., McRoberts, R., Fernández-Landa, A., José Luis Tomé, J., and Næsset, E. (2019). Estimating forest volume and biomass and their changes using random forests and remotely sensed data. *Rem. Sens.* 11:1944. doi: 10.3390/rs11161944

Fischer, C., and Traub, B. (eds.). (2019). *Swiss National Forest Inventory – Methods and Models of the Fourth Assessment*. Cham: Springer Nature Switzerland AG. p. 431.

Fortin, M. (2020). Updating plots to improve the precision of small area estimates: the example of the Lorraine region, France. *Can. J. For. Res*. 50, 648–658. doi: 10.1139/cjfr-2019-0405

Ghosh, M. (2020). Small area estimation: its evolution in five decades. *Stat. Trans.* 21, 1–22. doi: 10.21307/stattrans-2020-022

Ghosh, M., and Rao, J. N. K. (1994). Small area estimation: an appraisal. *Stat. Sci*. 9, 55–93.

Gregoire, T. G., Næsset, E., McRoberts, R. E., Ståhl, G., Andersen, H. E., Gobakken, T., et al. (2016). Statistical rigor in LiDAR-assisted estimation of aboveground forest biomass. *Rem. Sens. Environ*. 173, 98–108. doi: 10.1016/j.rse.2015.11.012

Guldin, R. W. (2018). How today's professionals prefer to find the science they need to do their jobs. *J. For*. 116, 451–459. doi: 10.1093/jofore/fvy036

Gutman, G. (2010). *Optical Remote Sensing: Basics, Data Processing, Applications*. Washington, DC: NASA. p. 61. Available online at: https://www.lcluc.umd.edu/sites/default/files/lcluc_documents/gutman_lcluc_8-2010_training_0.pdf

Haakana, H., Heikkinen, J., Katila, M., and Kangas, A. (2019a). Efficiency of post-stratification for a large-scale forest inventory – case Finnish NFI. *Ann. For. Sci*. 76:9. doi: 10.1007/s13595-018-0795-6

Haakana, H., Heikkinen, J., Katila, M., and Kangas, A. (2019b). Precision of exogenous post-stratification in small area estimation based on a continuous forest inventory. *Can. J. For. Res*. 50:359–370. doi: 10.1139/cjfr-2019-0139

Hill, A. (2018). *Integration of Small Area Estimation Procedures in Large-Scale Forest Inventories*. Thesis for the degree of Doctor of Sciences, Eidgenössische Technische Hochschule Zürich, Zürich, Switzerland. p. 125.

Hill, A., Mandallaz, D., and Langshausen, J. (2018). A double-sampling extension of the German national forest inventory for design-based small area estimation on forest district levels. *Rem. Sens.* 10:1052. doi: 10.3390/rs10071052

Hill, A., Massey, A., and Mandallaz, D. (2021). The R Package forestinventory: design-based global and small area estimations for multiphase forest inventories. *J. Stat. Softw*. 97, 1–40. doi: 10.18637/jss.v097.i04

Irulappa-Pillai-Vijayakumar, D. B., Renaud, J.-P., Morneau, F., McRoberts, R. E., and Vega, C. (2019). Increasing precision for French forest inventory estimates using the *k*-NN technique with optical and photogrammetric data and model-assisted estimators. *Rem. Sens*. 11, 991–1010. doi: 10.3390/rs11080991

Jiang, J., and Rao, J. (2020). Robust small area estimation: an overview. *Ann. Rev. Stats*. 7, 337–360. doi: 10.1146/annurev-statistics-031219-041212

Kangas, A., Astrup, R., Breidenbach, J., Fridman, J., Gobakken, T., Korhonen, K., et al. (2018). Remote sensing and forest inventories in Nordic countries—a roadmap for the future. *Scand. J. For. Res*. 33, 397–412. doi: 10.1080/02827581.2017.1416666

Kangas, A., Räty, M., Korhonen, K. T., Vauhkonen, J., and Packalen, T. (2019). Catering information needs from global to local scales – potential and challenges with national forest inventories. *Forests* 10:800. doi: 10.3390/f10090800

Katila, M., and Heikkinen, J. (2020). Reducing error in small-area estimates of multi-source forest inventory by multi-temporal data fusion. *Forestry* 93, 471–480. doi: 10.1093/foresj/cpz076

Krapavickaite, D., and Rancourt, E. (2019). Letters from the editors. *Surv. Statist*. 79:3. Available online at: http://isi-iass.org/home/wp-content/uploads/N79-2019-01-ISSN.pdf

Latifi, H., and Heurich, M. (eds.). (2019). Multi-scale remote sensing-assisted forest inventory: a glimpse of the state-of-the-art and future prospects. *Rem. Sens.* 11:1260. doi: 10.3390/rs11111260

Lavrakas, P. J. (ed.). (2008). "Small area estimation," in *Encyclopedia of Survey Research Methods*, Vol 2. (Thousand Oaks, CA: SAGE Publications).

Magnussen, S., Mandallaz, D., Breidenbach, J., Lanz, A., and Ginzler, C. (2014). National forest inventories in the service of small area estimation of stem volume. *Can. J. For. Res*. 44, 1079–1090. doi: 10.1139/cjfr-2013-0448

Magnussen, S., and Nord-Larsen, T. (2020). Forest inventory inference with spatial model strata. *Scand. J. For. Res*. 36, 43–54. doi: 10.1080/02827581.2020.1852309

Maltamo, M., Packalén, P., Suvanto, A., Korhonen, K. T., Mehtätalo, L., and Hyvönen, P. (2009). Combining ALS and NFI training data for forest management planning: a case study in Kuortane, Western Finland. *Eur. J. For. Res*. 128:305. doi: 10.1007/s10342-009-0266-6

Mandallaz, D. (2008). *Sampling Techniques for Forest Inventories*. Boca Raton, FL: Chapman and Hall/CRC. p. 256.

Mandallaz, D. (2013). Design-based properties of some small-area estimators in forest inventory with two-phase sampling. *Can. J. For. Res.* 43, 441–449. doi: 10.1139/cjfr-2012-0381

Mandallaz, D. (2014). A three-phase sampling extension of the generalized regression estimator with partially exhaustive information. *Can. J. For. Res.* 44, 383–388. doi: 10.1139/cjfr-2013-0449

Mandallaz, D., Breschan, J., and Hill, A. (2013). New regression estimators in forest inventories with two-phase sampling and partially exhaustive information: a design-based Monte Carlo approach with applications to small-area estimation. *Can. J. For. Res*. 43, 1023–1031 doi: 10.1139/cjfr-2013-0181

Massey, A. (2015). *Multiphase Estimation Procedures for Forest Inventories Under the Design-Based Monte Carlo Approach*. [Thesis for the degree of Doctor of Sciences, Eidgenössische Technische Hochschule Zürich, Zürich, Switzerland. p. 85.

Massey, A., Mandallaz, D., and Lanz, A. (2014). Integrating remote sensing and past inventory data under the new annual design of the Swiss National Forest Inventory using three-phase design-based regression estimation. *Can. J. For. Res*. 44, 1177–1186. doi: 10.1139/cjfr-2014-0152

McRoberts, R. E. (2012). Estimating forest attribute parameters for small areas using nearest neighbors techniques. *For. Ecol. Manage*. 272, 3–12. doi: 10.1016/j.foreco.2011.06.039

McRoberts, R. E., Chen, Q., and Walters, B. F. (2017). Multivariate inference for forest inventories using auxiliary airborne laser scanning data. *For. Ecol. Manage*. 401, 295–303. doi: 10.1016/j.foreco.2017.07.017

McRoberts, R. E., Næsset, E., Saatchi, S., Liknes, G. C., Walters, B. F., and Chen, Q. (2019). Local validation of global biomass maps. *Int. J. Appl. Earth Obs. Geoinform*. 83:101931. doi: 10.1016/j.jag.2019.101931

Nelson, R., Gobakken, T. Næsset, E., Gregoire, T., Ståhl, G., Holm, S., et al. (2012). Lidar sampling—Using an airborne profiler to estimate forest biomass in Hedmark County, Norway. *Remote Sens. Environ*. 123, 563–578. doi: 10.1016/j.rse.2011.10.036

Novo-Fernández, A., Barrio-Anta, M., Recondo, C., Cámara-Obregón, A., and López-Sánchez, C. A. (2019). Integration of national forest inventory and nationwide airborne laser scanning data to improve forest yield predictions in north-western Spain. *Rem. Sens*. 11:1643. doi: 10.3390/rs11141693

Pfeffermann, D. (2002). Small area estimation – new developments and directors. *Intl. Stat. Rev*. 70, 125–143. doi: 10.1111/j.1751-5823.2002.tb00352.x

Pfeffermann, D. (2013). New important developments in small area estimation. *Stat. Sci*. 28, 40–68. doi: 10.1214/12-STS395

Pulkkinen, M., and Zell, J. (2019). *Overview of Research Carried out on Small-Area Estimation Around Swiss NFI Since 2010*. Birmensdorf: Swiss Federal Institute for Forest, Snow, and Landscape Research (WSL). p. 3.

Rahlf, J., Hauglin, M., Astrup, R., and Breidenbach, J. (2021). Timber volume estimation based on airborne laser scanning—comparing the use of national forest inventory and forest management inventory data. *Ann. For. Sci*. 78:49. doi: 10.1007/s13595-021-01061-4

Rahman, A. (2008). *A Review of Small Area Estimation Problems and Methodological Developments*. Discussion Paper 66. Canberra, ACT: Univ. of Canberra, National Centre for Social and Economic Modeling (NATSEM). p. 56. Available online at: https://researchoutput.csu.edu.au/en/publications/a-review-of-small-area-estimation-problems-and-methodological-dev (accessed June 18, 2021).

Rao, J. N. K. (2003). *Small Area Estimation*. New York, NY: Wiley.

Rao, J. N. K., and Molina, I. (2015). *Small Area Estimation, 2nd Edn.* New York, NY: Wiley. p. 480.

Saei, A., and Chambers, R. (2003). *Small Area Estimation: A Review of Methods Based on the Application of Mixed Models*. Methodology Working Paper M03/16. Southampton, UK: University of Southampton. Southampton Statistical Sciences Research Institute. 36 p. Available online at: https://eprints.soton.ac.uk/8166/ (accessed June 18, 2021).

Ståhl, G., Saarele, S., Schnell, S., Holm, S., Breidenbach, J., Healey, S. P., et al. (2016). Use of models in large-area forest surveys: comparing model-assisted, model-based, and hybrid estimation. *For. Ecos*. 3:5. doi: 10.1186/s40663-016-0064-9

Steinmann, K., Mandallaz, D., Ginzler, C., and Lanz, A. (2013). Small area estimations of proportion of forest and timber volume combining Lidar data and stereo aerial images with terrestrial data. *Scand. J. For. Res*. 28, 373–385. doi: 10.1080/02827581.2012.754936

Strîmbu, V., and Ørka, H., Næsset, E. (2021). Consistent forest biomass stock and change estimation across stand, property, and landscape levels. *Can. J. For. Res*. 51, 1–11. doi: 10.1139/cjfr-2020-0203

Tomppo, E. (1990). "Designing a satellite image-aided national forest survey in Finland," in *The Usability of Remote Sensing for Forest Inventory and Planning: Proceedings of an SNS/IUFRO Workshop*, *26-28 February 1990. Report No. 4*. (Umeå: Swedish University of Agricultural Sciences, Remote Sensing Laboratory), 43–47.

Tomppo, E. (1991). "Satellite image-based national forest inventory for Finland," *Proceedings of the Symposium on Global and Environmental Monitoring, Techniques, and Impacts, International Archives of Photogrammetry and Remote Sensing 28, Vol. 7, Part 1* (Victoria, BC:), 419–424. Available online at: https://goobi.tib.eu/viewer/image/856669164/1/ (accessed June 18, 2021).

Tomppo, E., Katila, M., Mäkisara, K., and Peräsaari, J. (2012). *The Multi-Source National Forest Inventory of Finland—Methods and Results 2007*. Working Papers of the Finnish Forest Research Institute 227. p. 233. Available online at: http://www.metla.fi/julkaisut/workingpapers/2012/mwp227-en.htm (accessed June 18, 2021).

U.S. Census Bureau (2021). *Small Area Estimation*. Available online at: https://www.census.gov/topics/research/stat-research/expertise/small-area-est.html (accessed June 18, 2021).

Vega, C., Renaud, J.-P., Sagar, A., and Bouriaud, O. (2021). A new small area estimation algorithm to balance between statistical precision and scale. *Intl. J. Appl. Earth Obs. Geoinf*. 97:102303. doi: 10.1016/j.jag.2021.102303

Wagner, J., Münnich, R., Hill, J., Stoffels, J., and Udelhoven, T. (2017). Non-parametric small area models using shape-constrained penalized B-splines. *J. Royal Stat. Soc*. 180, 1089–1109. doi: 10.1111/rssa.12295

Conflict of Interest: RG owns Guldin Forestry LLC, a sole-proprietorship forestry consulting firm through which he conducts forest research, forest resource assessments, and analyses of forest policy issues for clients in the United States of America and internationally. He declares that he received no funding from any source outside his firm to prepare this paper. He alone was involved in the study design; the collection, analysis, and interpretation of the data; and the writing of this article.

frontiers in Forests and Global Change

ORIGINAL RESEARCH
published: 20 October 2021
doi: 10.3389/ffgc.2021.745916

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

Hailemariam Temesgen[1†], Francisco Mauro[1†], Andrew T. Hudak[2], Bryce Frank[3], Vicente Monleon[4], Patrick Fekety[5], Marin Palmer[6] and Timothy Bryant[6]*

[1] Forest Biometrics and Measurements Laboratory, Department Forest Engineering, Resources and Management, Oregon State University, Corvallis, OR, United States, [2] US Forest Service, Rocky Mountain Research Station, Moscow, ID, United States, [3] US Bureau of Land Management, Oregon/Washington State Office, Portland, OR, United States, [4] US Forest Service, Pacific Northwest Research Station, Corvallis, OR, United States, [5] Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, CO, United States, [6] USDA Forest Service, Pacific Northwest Region, Portland, OR, United States

OPEN ACCESS

Edited by:
Philip Radtke, Virginia Tech, United States

Reviewed by:
Johannes Rahlf, Norwegian Institute of Bioeconomy Research (NIBIO), Norway
Qianqian Cao, Virginia Tech, United States

***Correspondence:**
Hailemariam Temesgen
temesgen.hailemariam@oregonstate.edu

†These authors share first authorship

Specialty section:
This article was submitted to Forest Management, a section of the journal Frontiers in Forests and Global Change

Received: *22 July 2021*
Accepted: *30 September 2021*
Published: *20 October 2021*

Citation:
Temesgen H, Mauro F, Hudak AT, Frank B, Monleon V, Fekety P, Palmer M and Bryant T (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. Glob. Change 4:745916. doi: 10.3389/ffgc.2021.745916

Stands are the primary unit for tactical and operational forest planning. Forest managers can use remote-sensing-based forest inventories to precisely estimate attributes of interest at the stand scale. However, remote-sensing-based inventories typically rely on models relating remote-sensing information to forest attributes for fixed area plots with accurate coordinates. The collection of that kind of ground data is expensive and time-consuming. Furthermore, remote-sensing-based inventories provide precise descriptions of the forest when the remote-sensing data were collected, but they inevitably become outdated as the forest evolves. Fay–Herriot (FH), models can be used with ground information from variable radius plots even if the plot coordinates are unknown. Thus, they provide an efficient way to update old remote-sensing-based inventories or develop new ones when fixed radius plots are unavailable. In addition, FH models are well described in the small-area estimation literature and allow reporting estimation uncertainties, which is key to incorporating quality controls to remote-sensing inventories. We compared two scenarios developed in the Willamette National Forest, OR, United States, to produce stand-level estimates of above-ground biomass (AGB), and Volume (V) for natural and managed stands. The first, Case 1, was developed using auxiliary data from a recent lidar acquisition. The second, Case 2, was developed to update an old remote-sensing-based inventory. Results showed that FH models allowed for improvements in efficiency with respect to direct stand-level estimates obtained using only field data for both case scenarios and both typologies of stands. Average improvements in efficiency in natural stands were 37.36% for AGB and 33.10% for Volume for FH models from Case 1 and 20.19% for AGB and 19.25 for V for Case 2. For managed stands, average improvements for Case 1 were 2.29 and 19.92% for AGB and V, respectively, and for Case 2, improvements were 15.55% for AGB and 16.05% for V.

Keywords: stand-level models, lidar, above-ground biomass, carbon monitoring system, uncertainty

INTRODUCTION

Stands are the primary unit for tactical–operational planning and management. A stand is an area or polygon with a relatively homogeneous forest structure and different from surrounding areas in terms of structure, composition, or management objectives. The size of a forest stand typically ranges from about 1 to 20–40 ha, and obtaining stand-level information is critical to inform management and planning decisions (Breidenbach et al., 2018; Mauro et al., 2019). Traditional forest inventories produce stand-level information using field surveys or stand exams where it is common to use variable radius plots (VRPs). These field surveys allow obtaining estimates for different variables of interest for the forest managers and assessing the quality of those surveys using methods described in the forest inventory literature.

Remote sensing inventories typically follow an area-based approach (ABA), where fixed area plots and remote sensing data are combined to produce maps with estimates of forest attributes at resolutions in the range of 10–30 m (Næsset, 2002). This methodology has been extensively used with lidar (e.g., Maltamo et al., 2004; González-Ferreiro et al., 2012; Babcock et al., 2015; Fekety et al., 2018), data from other sensors such as Landsat (LeMay et al., 2008; Pflugmacher et al., 2012) or Sentinel I and II (Forkuor et al., 2020), or different combinations of sensors (e.g., Vafaei et al., 2018; Forkuor et al., 2020). This methodology is well known and produces, in a very efficient manner, estimates in high-resolution grids (i.e., 10–30 m resolution) for a large number of forest attributes. These estimates can be summarized to generate stand-level maps for forest planning tasks. Besides, several studies have conducted small area estimation analysis showing that, with this methodology it is possible to obtain not only stand-level estimates of forest attributes but also measures of uncertainty for those stand-level estimates (Mauro et al., 2016, 2019; Breidenbach et al., 2018; Frank et al., 2020). Stand-level measures of uncertainty are a desirable output of any inventory method because they can be used as a measure of quality control. Reported uncertainties can be used to identify stands with more unreliable estimates that can be targeted in further field measurements efforts, saving resources for field data collections. Furthermore, even when additional ground measurements are not an option, stand-level measures of uncertainty are useful and can be incorporated in decision making processes and sensitivity analyses.

While the ABA method has been extensively developed during the last decade, it presents several drawbacks for operational inventories. This methodology's main problem is that it is based on using fixed-radius plots with accurate coordinates. The collection of that kind of ground information is costly on a per plot basis or stand when stands are the sample units (Hummel et al., 2011). Fixed-radius plot inventories are efficient at the level of a whole landscape or project area (Hudak et al., 2014), which is typically stratified to distribute the sample plots across the range of stand structure conditions without regard to stand boundaries. However, for stand-level inventory, collecting fixed-radius plot data with highly accurate GPS coordinates requires more resources per sampled stand than typical stand exams based on VRP. This is because in the later, field plot coordinates are not recorded or are obtained using less expensive low-grade GPS equipment. Recent studies have demonstrated that it is possible to use VRP combined with remote sensing data in several ways. One possibility is to optimize the basal area factor (BAF) used in the VRP to the stand structure variation (Deo et al., 2016), or to use VRP and a constant BAF, using arbitrary but consistent support areas for the remote sensing predictors throughout the study area and operate as in the traditional ABA method (Grafström et al., 2017). While these methods are very interesting for operational inventories because they allow using VRP data, they do not eliminate the need to obtain accurate coordinates for the VRP. Another option that fits better with standard practices for stands exams is the use of Fay–Herriot (FH), models (Fay and Herriot, 1979). These models are sometimes referred to as stand-level models in forestry contexts and allow combining remote sensing data with different ground measurements in stands, eliminating the need for precise coordinates for ground measurements.

While traditional ABA models are developed considering field plots as the primary modeling unit, FH models operate at a coarser scale. FH models are developed with stands as the primary element. This implies several departures from traditional ABA models. One difference is that auxiliary information for FH models needs to be associated with stands for operational inventories (Goerndt et al., 2011; Mauro et al., 2017; Green et al., 2019) or with larger-scale domains such as counties for national inventories (Coulston et al., 2021). For example, stand-level summaries of lidar variables have been used in previous studies using FH models in stand-level forest inventories in Europe and the United States (Magnussen et al., 2017; Mauro et al., 2017; Ver Planck et al., 2018). But the most critical difference between FH and traditional ABA models is that ground information for the modeling units of FH models is typically incomplete. Fixed radius plots used in traditional ABA models are exhaustive and all or most of the trees within the plots are measured. This allows treating forest attributes (i.e., response variables) computed for the plots as known quantities. However, in operational settings stands are never fully measured; instead, they are sampled with many field plots that can vary between stands. This implies that the response variables used to develop stand-level FH models are subject to sampling errors that need to be accounted for in the modeling stage. FH models include a variance component to account for these sampling errors and can be seen as measurement error models where the response used for modeling has an inherent uncertainty because it comes from a sample and not from a complete measurement.

The coarse resolution of stand-level FH models can be a drawback for certain applications. However, FH models have advantages in terms of flexibility and data requirements over ABA methods. The most interesting properties of FH models are: (1) that they can be developed with any ground measurement from which it is possible to obtain unbiased estimators for stand-level attributes and their associated variances (i.e., VRP, transects, and sector plots) and (2) that they eliminate the need to record precise plot coordinates in the field (Goerndt et al., 2011; Ver Planck et al., 2018). Thus, FH models can use VRP data and plots without accurate GPS coordinates, making them a very appealing

alternative for operational forest inventories based on lidar or other remote sensing auxiliary information sources. Despite their potential, very few applications of stand-level FH models exist in forest inventory literature and are focused on developing a new inventory using available auxiliary information. In this manuscript, we aim to analyze two possible scenarios where FH models can be used to combine remote sensing data and VRP data from stand exams. These scenarios or cases are:

1. **Case 1:** Developing a new stand-level inventory using recently collected lidar auxiliary information for an area where no fixed radius plot data is available.
2. **Case 2:** Updating an old remote-sensing-based inventory combining the same ground data used in Case 1 with remote sensing datasets developed at regional scales and with no access restrictions. The old remote sensing-based inventory was developed 5 years prior to the ground data collection. The auxiliary information included climate data, topographic variables, and spectral changes in Landsat images. These auxiliary variables aimed at capturing possible changes in the study area during the years between the old-remote sensing inventory and the updating date.

Both cases under analysis use stand exams based on VRP from the US Forest Service Field Sampled Vegetation (FSVeg) database but they can be directly replicated in many other areas managed by the US Forest Service or in other regions in the world.

MATERIALS AND METHODS

Study Area

The study area comprises 31,209 ha inside the Willamette National Forest, OR, United States covered by different remote sensing datasets that include a recent lidar acquisition and a 30 m resolution map with above-ground biomass (AGB) predictions (**Figure 1**). Details on these datasets are provided in sections "Case 1: Fay–Herriot Models for New Inventories" and "Case 2: Fay–Herriot Models to Update Inventories." Elevations range from 450 to 1700 m above sea level. Two forks of the Santiam river cross the study area from East to West and have numerous tributaries that form a complex drainage network where slopes do not have a dominant orientation. Conifers dominate vegetation with Douglas-fir, *Pseudotsuga menziesii* (Mirb) Franco, the most abundant species, and other conifers such as noble fir, *Abies procera* Rehder, silver fir, *Abies amabilis* Douglas ex J.Forbes, western hemlock, *Tsuga heterophylla* (Raf.) Sarg., and western red cedar, *Thuja plicata* Donn ex D.Don, as secondary species with a much lower abundance. Hardwood species have a minor presence with red alder, *Alnus rubra* Bong, and golden chinquapin, *Chrysolepis chrysophylla* (Douglas ex Hook.) Hjelmq., as the most important species in this group.

The study area contains 1616 stands with different management goals. Stand boundaries are the result of a continuous effort performed by forest managers in the study area and is based on the management history, structure, and composition of the forest. Stands are classified according to their management objectives into "Natural" and "Managed" (**Figure 2**). While the terms natural and managed can be the subject of lengthy discussions, we will keep this terminology as it is used in the FSVeg database. There are 696 natural stands and 920 managed stands. Natural stands occupy approximately two-thirds of the area. They are typically of larger size (i.e., average size = 31.04 ha, median size = 11.61 ha) than managed stands (i.e., average size = 9.90 ha, median size = 7.94 ha) (**Figure 2**). Managed stands have a past history of silviculture entry and often include artificial regeneration. In most cases, even-aged structures are subject to thinning and logging operations. Natural stands are subject to less intense management and tend to have larger dimensions and a larger internal variability in forest structure and ages (**Figure 2**).

Sampled Stands and Field Sampled Vegetation Ground Data

In total, 37 natural and 238 managed stands in the study area were sampled in 2018 by field crews that used VRP with BAFs that changed depending on the stand characteristics. Natural and managed sampled stands were selected by forest managers in the region using a randomize procedure and also expert knowledge to ensure that most prevalent forest types were present in the sample. The proportion of natural stands sampled (i.e., 5.31%) was about five times smaller than the proportion of managed stands sampled (i.e., 25.86%).This reflects the larger information needs for the managed stands derived from their more intensive sylviculture. VRP were randomly located within the stands by the field crews. The number of VRP collected in the 37 sampled natural stands was 157 and the number of plots in the 238 sampled managed stands was 943 plots. The number of field plots in the sampled stands varied from 2 to 18, but 3, 4, and 5 were the most frequent number of field plots per stand (**Figure 3**). The field plot density, for the entire study area (i.e., including sampled and unsampled stands), was 0.017 plots ha^{-1} for natural stands (1 plot every 58.01 ha) and 0.044 plots ha^{-1} for managed stands (1 plot every 22.90 ha). For each VRP, the species, diameter at breast height (*dbh*), height (*ht*), and the live or dead status of each selected tree were recorded. Field crews used standard devices to measure *dbh* (i.e., caliper or logger's tape) and *ht* (i.e., hypsometer or laser rangefinder). Finally, the BAF used in the plot allowed computing an expansion factor for each tree in the plot.

Parameter of Interest

For both Case 1 and Case 2, we considered estimating, for every stand in the study area, the total of AGB, and merchantable volume (V), for the year 2018, both expressed on a per unit area basis. Thus, for every stand, the unknown parameter of interest was

$$\mu_i = \frac{1}{A_i} \sum_{t=1}^{N_i} AGB_{ti} \tag{1}$$

when considering AGB and

$$\mu_i = \frac{1}{A_i} \sum_{t=1}^{N_i} V_{ti} \tag{2}$$

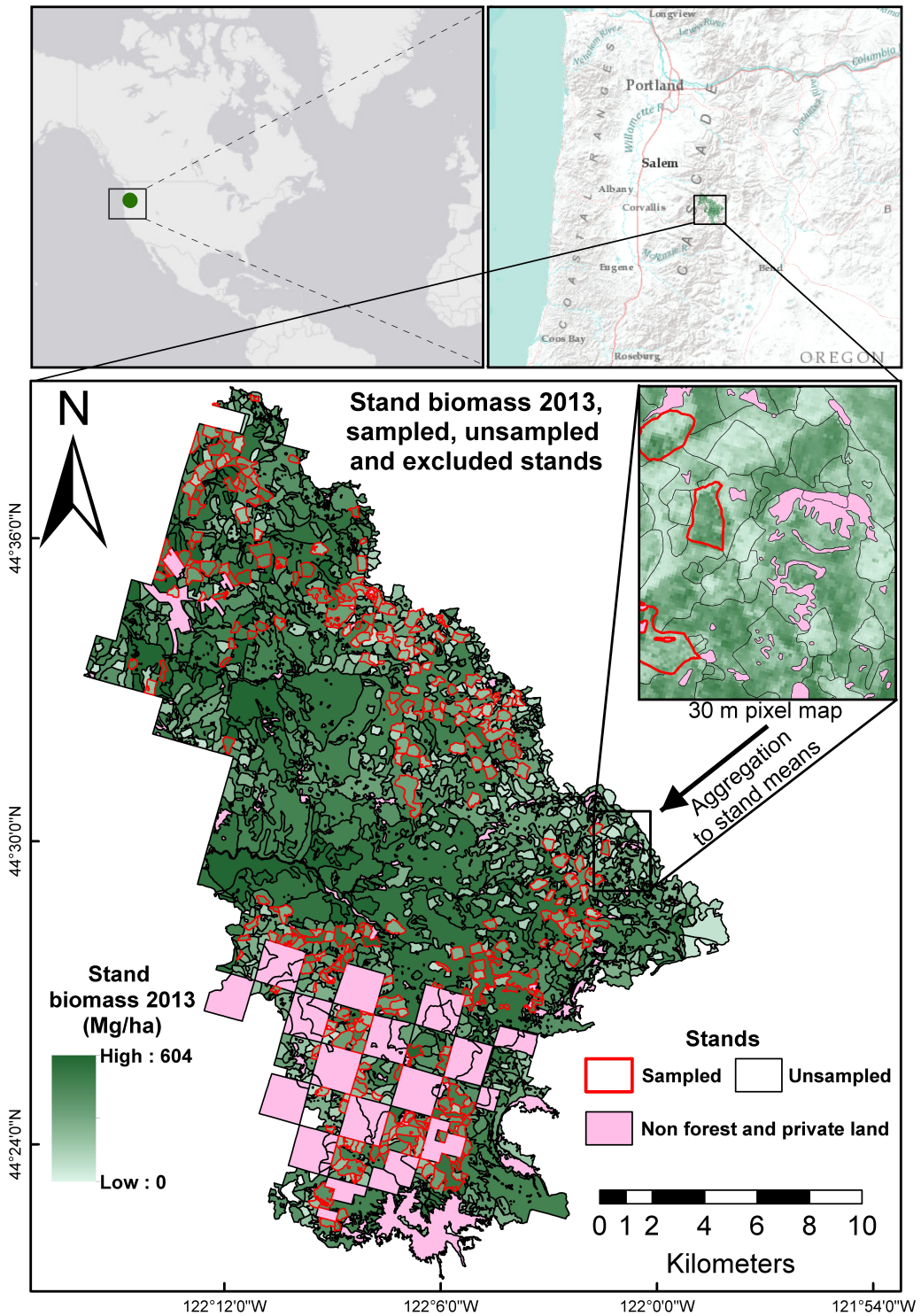


FIGURE 1 | Location of the study area. In pink are stands excluded from the analysis because they were not covered by the remote sensing inventory from 2013. Stands in green color with red outline are sampled stands; and green stands with black outline are unsampled stands.

when considering V. In equations 1, 2 AGB_{ti} and V_{ti} are the AGB and V of the t-th tree in the i-th stand, and N_i and A_i are, respectively, number of trees and the area of the i-th stand. It is important to note that while μ_i, AGB_{ti}, V_{ti}, and V_{ti} were all unknown quantities, the stand area was known for every stand in the study area.

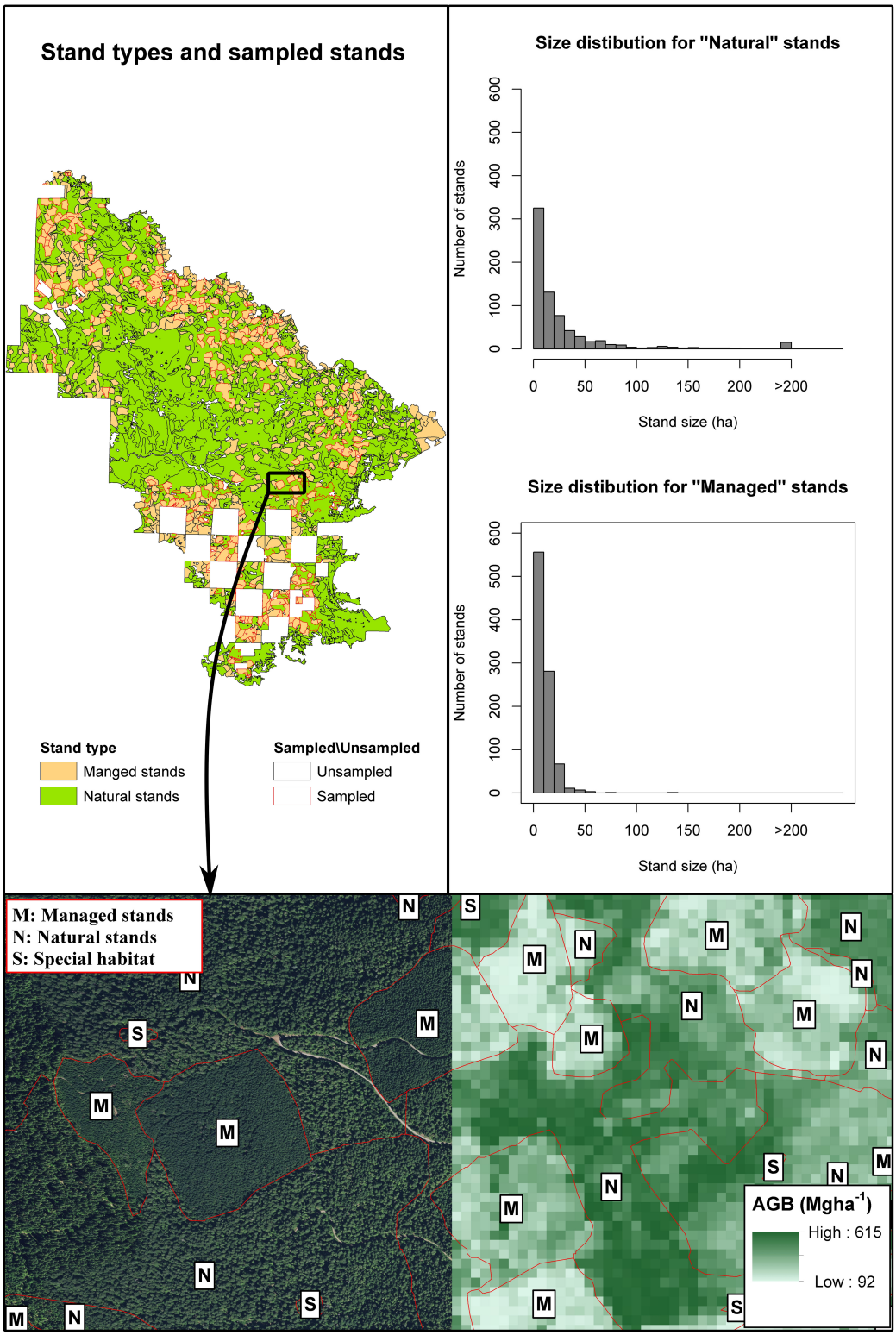


FIGURE 2 | Upper left panel displays the location of natural and managed stands. Upper right and middle right panels display the size distribution within the study area for natural and managed stands, respectively. Bottom panel, sampled area showing the orthophoto on the western side and the CMS AGB map on the eastern side of the image. Managed stands are labeled with the letter M, natural stands are labeled with the letter N, and special habitat areas (small size non-forested polygons within stands) are labeled with letter S.

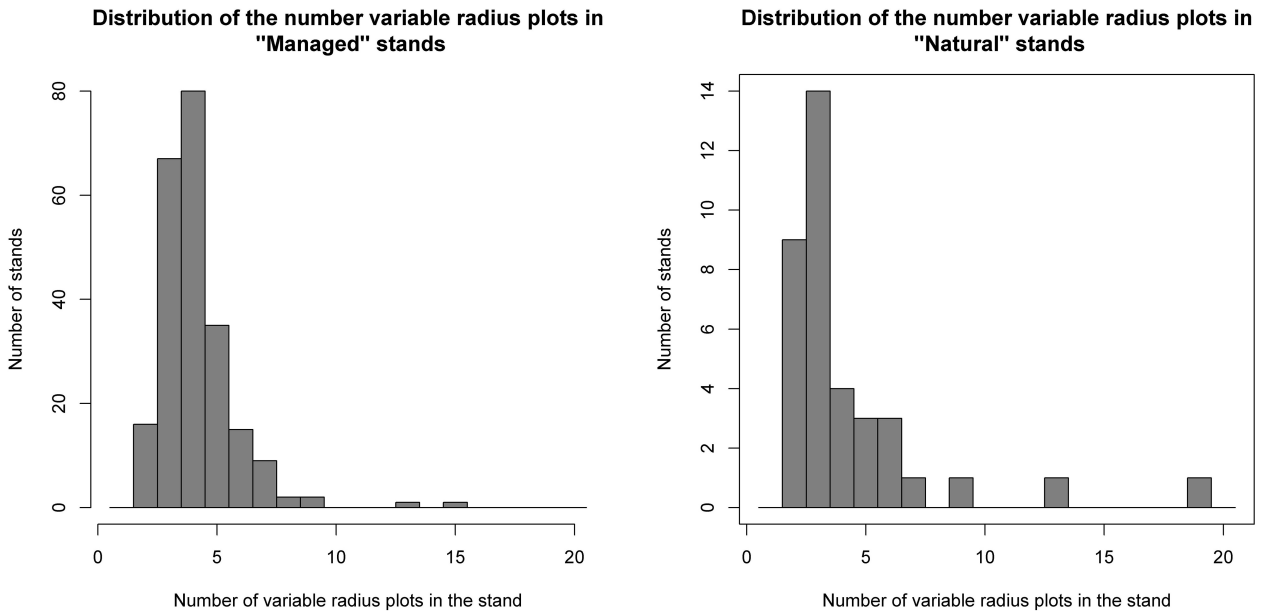


FIGURE 3 | Distribution of the number of VRP per sampled stand in natural and managed stands.

Direct Above-Ground Biomass and Volume Estimators and *mse* Estimators

We used the Forest Vegetation Simulator (FVS), to compute an estimate, $\widehat{\mu}_{gij}$, of each parameter of interest for each VRP using the Horvitz–Thompson (HT), estimator

$$\widehat{\mu}_{gij} = \sum_{t=1}^{n_{ij}} \frac{y_{ijt}}{EF_{ijt}}. \quad (3)$$

In equation 3, y_{ijt} represents either AGB_{tij} or V_{tij} for the t-th tree measured in j-th VRP in the i-th sampled stand and EF_{ijt} represents their respective expansion factors. The number of measured trees in the j-th VRP in the i-th stand is n_{ij} and the subindex g in $\widehat{\mu}_{gij}$ indicates that it is a direct estimate based on the ground data.

For each sampled stand, VRP estimates $\widehat{\mu}_{gij}$ from the n_i plots measured in the stand were averaged to produce a final direct ground estimate $\widehat{\mu}_{gi*}$ of AGB and V

$$\widehat{\mu}_{gi.} = \frac{1}{n_i} \sum_{j=1}^{n_i} \widehat{\mu}_{gij}. \quad (4)$$

A summary of the stand estimates based on the VRP data is presented in **Table 1**.

The HT estimator is unbiased and each VRP is assumed to provide an independent sample drawn under a sampling design that remains constant for all VRP in the stand. Thus, for a given stand, all $\widehat{\mu}_{gij}$ were considered to be realizations of a random variable with mean μ_i and unknown variance σ^2_{ei0} . The final direct estimate for the stand, $\widehat{\mu}_{gi*}$, is the average of n_i independent and identically distributed random variables. Therefore, $\widehat{\mu}_{gi*}$ is also a random variable with mean μ_i and its variance, σ^2_{ei} equals $\frac{\sigma^2_{ei0}}{n_i}$. This allows establishing the following relation, equation 5, between the stand estimate $\widehat{\mu}_{gi*}$, the unknown parameter of interest μ_i and the sampling error e_i

$$\widehat{\mu}_{gi.} = \mu_i + e_i \quad (5)$$

For any two stands, sampling errors are assumed to be independent of each other. Furthermore, due to the unbiasedness property of the HT estimator, errors are assumed to be distributed with zero mean and variance σ^2_{ei}. The variance σ^2_{ei0} is unknown, but an unbiased estimator can be obtained pooling together the estimates of all VRP in a given stand as

$$\widehat{\sigma}^2_{ei0} = \sum_{j=1}^{n_i} \frac{(\widehat{\mu}_{gij} - \widehat{\mu}_{gi*})^2}{n_i - 1}. \quad (6)$$

Based on equation 6 the variance and the mean square error of $\widehat{\mu}_{gi*}$ is estimated using

$$mse\left(\widehat{\mu}_{gi.}\right) = \widehat{\sigma}^2_{ei} = \frac{\widehat{\sigma}^2_{ei0}}{n_i}. \quad (7)$$

Estimators in equations 4, 7 are typically used in stand-level inventories using only ground data when reporting estimates and measures of uncertainty for sampled stands.

Stand-Level Fay–Herriot Models and Estimators

Stand-Level Fay–Herriot Models

Stand-level FH models explicitly acknowledge that the stand-level information on the parameter of interest is subject to sampling errors. The first component in an FH model postulates a relation between the true and unknown parameters of interest for stands

TABLE 1 | Summary of stand-level estimates based on VRP data.

Variable	Stand type	Weighted mean	Arithmetic mean	Min	Max	SD
AGB (Mg ha^{-1})	Natural	279.24	259.05	61.99	526.32	95.76
	Managed	107.25	103.94	7.00	222.39	34.04
V (m^3 ha^{-1})	Natural	1016.85	944.53	195.40	1890.23	362.41
	Managed	325.11	312.02	11.19	661.49	120.56

Weighted mean are the means of the stand-level estimates based on VRP data with weights proportional to the stand area. SD is the standard deviation of the VRP estimates of the sampled stands. AGB stands for above-ground biomass, and V for volume, V.

and the available auxiliary information through a regression model

$$\mu_i = \mathbf{x}_i^t\beta + v_i. \tag{8}$$

In equation 8, v_i is the model error that is assumed to be normally distributed with mean 0 and variance σ_v^2 [i.e., $v_i N(0, \sigma_v^2)$], β is a vector of model coefficients where the first element is the model intercept and $\mathbf{x}_i$ is a vector of stand-level auxiliary variables where the first element equals 1 when β includes an intercept term (see sections "Case 1: Fay–Herriot Models for New Inventories" and "Case 2: Fay–Herriot Models to Update Inventories" for a description of the auxiliary variables used for Case 1 and Case 2, respectively). These models cannot be fit because the true values of μ_i are unknown. In practice, only the direct ground estimates $\widehat{\mu}_{gi.}$ are available, however, both, μ_i and $\widehat{\mu}_{gi*}$ are related through the sampling model indicated in equation 5. When the regression model (8) and the sampling model (5) are combined, assuming that v_k and e_l are independent for all k and l, we obtain the basic FH model (9)

$$\widehat{\mu}_{gi.} = \mathbf{x}_i^t\beta + v_i + e_i. \tag{9}$$

Fay–Herriot models explicitly acknowledge the presence of the sampling errors and require information on the variances $\widehat{\sigma}_{ei}^2$ of the direct ground estimates. These variances can be estimated from the VRP data, equation 7, and then used with the known auxiliary information for the stands $\mathbf{x}_i^t$ and the direct estimates $\widehat{\mu}_{gi.}$ to estimate the remaining model parameters, i.e., β and σ_v^2. These models are typically fit using restricted maximum likelihood, REML, under the implicit assumption that sampling errors are normally distributed, i.e., $e_i \sim N(0, \frac{\sigma_{ei\,0}^2}{n_i})$.

Fay–Herriot Estimators

Once FH models are fitted, they can be used to obtain stand-level estimates and their corresponding uncertainty metrics. For sampled stands, estimates based on the FH model, $\widehat{\mu}_{FH,i}$, are obtained using the empirical best linear unbiased predictor, EBLUP,

$$\widehat{\mu}_{FHi>1VRP} = \gamma_i\widehat{\mu}_{gi} + (1-\gamma_i)\,\mathbf{x}_i^t\widehat{\beta}. \tag{10}$$

For unsampled stands and stands with only one VRP, estimates, $\widehat{\mu}_{FH,i}$, are obtained as synthetic estimates entirely based on the fitted model

$$\widehat{\mu}_{FHi\leq1VRP} = \mathbf{x}_i^t\widehat{\beta}. \tag{11}$$

For sampled stands, the EBLUP, equation 10, is a weighted average of the direct estimator obtained using only the ground information and the synthetic estimator. The weight and the degree of shrinking of $\widehat{\mu}_{FHi}$ toward the synthetic estimator $\mathbf{x}_i^t\widehat{\beta}$, is controlled by the parameter

$$\gamma_i = \frac{\widehat{\sigma}_v^2}{\widehat{\sigma}_v^2 + \widehat{\sigma}_{ei}^2}, \tag{12}$$

in the following manner. For stands where the direct estimates are reliable and have small errors compared to the unexplained variance of the fitted models (i.e., $\widehat{\sigma}_v^2 > \widehat{\sigma}_{ei}^2$), γ_i is close to 1, and $\widehat{\mu}_{FH,i}$ is approximately equal to the ground estimate for the stand. That is, in stands with low sampling errors, the direct ground estimate is "trusted" more than the model and $\widehat{\mu}_{FHi} \cong \widehat{\mu}_{gi*}$. For stands where direct estimates are unreliable, $\widehat{\sigma}_v^2 = \widehat{\sigma}_{ei}^2$ the parameter γ_i is close to 0, and most weight and confidence will be put in the synthetic prediction $\widehat{\mu}_{FHsi} \cong \mathbf{x}_i^t\widehat{\beta}$. For unsampled stands or stands with only one VRP, γ_i cannot be computed because it is not possible to obtain the variance of the direct estimator, $\widehat{\sigma}_{ei}^2$, with less than two VRP. Therefore, for stands with less than two VRP, all weight needs to be put in the model, and then the stand-level estimates based on the FH model are synthetic.

For stands with two or more VRP, for models fitted using REML, an approximately unbiased estimator of the mean square error of $\widehat{\mu}_{FHi}$ is

$$mse\left(\widehat{\mu}_{FHi>1VRP}\right) = g_{i\,1}\left(\widehat{\sigma}_v^2\right) + g_{i\,2}\left(\widehat{\sigma}_v^2\right) + 2g_{i\,3}\left(\widehat{\sigma}_v^2\right). \tag{13}$$

This mean square error estimator has three components $g_1\left(\widehat{\sigma}_v^2\right)$, $g_2\left(\widehat{\sigma}_v^2\right)$, and $2g_3\left(\widehat{\sigma}_v^2\right)$ indicated in equations 14–16:

$$g_{i\,1}\left(\widehat{\sigma}_v^2\right) = \gamma_i\widehat{\sigma}_{ei}^2 \tag{14}$$

$$g_{i\,2}\left(\widehat{\sigma}_v^2\right) = (1-\gamma_i)^2\mathbf{x}_i^t\left\{\sum_{i:n_i>2}\frac{\mathbf{x}_i^t\mathbf{x}_i}{\widehat{\sigma}_v^2+\widehat{\sigma}_{ei}^2}\right\}^{-1}\mathbf{x}_i \tag{15}$$

$$g_{i\,3}\left(\widehat{\sigma}_v^2\right) = \widehat{\sigma}_{ei}^4(\widehat{\sigma}_v^2+\widehat{\sigma}_{ei}^2)^{-3}\overline{V}(\widehat{\sigma}_v^2). \tag{16}$$

The term $\overline{V}\left(\widehat{\sigma}_v^2\right)$ in equation 15 is the inverse of the Fisher information matrix for the model (9). Details on $\overline{V}\left(\widehat{\sigma}_v^2\right)$ can be found in Rao and Molina (2015, p. 136). This estimator has a bias whose order of magnitude is $o\left(m^{-1}\right)$, where m is the number of sampled stands. Thus in applications where a large number of stands are sampled it can be expected to provide almost unbiased estimates of the mean square error of $\widehat{\mu}_{FHi}$. For unsampled stands or stands with only one plot, an estimator of the mean square error of $\widehat{\mu}_{FHi}$ can be obtained using equation 17 (Rao and Molina, 2015, p. 139)

$$mse\left(\widehat{\mu}_{FHi\leq1VRP}\right) = \mathbf{x}_i^t\left\{\sum_{i:n_i\geq2}\frac{\mathbf{x}_i^t\mathbf{x}_i}{\widehat{\sigma}_v^2+\widehat{\sigma}_{ei}^2}\right\}^{-1}\mathbf{x}_i + \widehat{\sigma}_v^2. \tag{17}$$

Note that we only use the subindexes $>1VRP$ and $\leq1VRP$ in equations 10, 11, 13, 17 to explicitly state the formulas to use depending on the number of VRP in the stand. In the remaining sections these subindexes will be omitted to simplify the notation;

and $\widehat{\mu}_{FHi}$ and $mse\,(\widehat{\mu}_{FHi})$ will refer to the estimator and *mse* estimator needed depending on the number of VRP in the stand. The root mean square error, *rmse*, and relative root mean square error, *rrmse*, for estimates based the FH models were computed as $rmse\,(\widehat{\mu}_{FHi}) = \sqrt{mse\,(\widehat{\mu}_{FHi})}$ and $rrmse\,(\mu_{FHi}) = \frac{rmse(\widehat{\mu}_{FHi})}{\widehat{\mu}_{FHi}}$, respectively.

Comparisons Between Fay–Herriot Estimators and Ground-Based Estimators

Comparisons between models for Case 1 and Case 2 for a given variable were based on the ratio of the estimated model variances $\widehat{\sigma}^2_{v\ case\ 2}/\widehat{\sigma}^2_{v\ case\ 1}$. To compare the uncertainty of stand-level estimates from the FH models, we used both $rmse\,(\widehat{\mu}_{FHi})$ and $rrmse\,(\mu_{FHi})$. Finally, improvements with respect to estimates based only the field data were measured using the relative efficiency. This metric was only computed for stands with two or more VRP.

$$\triangle_{eff\ i} = 1 - \frac{rmse\,(\widehat{\mu}_{FHi})}{rmse\,(\widehat{\mu}_{gi})} \tag{18}$$

Models for Case 1 and Case 2

Case 1: Fay–Herriot Models for New Inventories

For the first case scenario, auxiliary variables were computed from a recent lidar data collection completed in the fall of 2016. The lidar data were acquired using a Leica ALS70-HP lidar system mounted on a fixed-wing platform flying at an average altitude of 1965 m above the ground level with a nominal speed of 110 knots. The scanning angle was 30°, and the nominal pulse density 4.2 pulses per m^2.

A 30 m resolution grid was cast over the study area and lidar metrics including (1) percentiles and summaries (i.e., means, standard deviations, and moments) of the distribution of elevations above the ground of the lidar returns, (2) proportions of points in different height strata, and (3) topographic metrics were computed for each pixel using FUSION (Mc Gaughey, 2019). In total 134 variables were available. For each stand, we computed the mean and standard deviation of the pixel-level values of these metrics. The result was a total of 268 (i.e., 134 means and 134 standard deviations) stand-level metrics. These stand-level metrics were considered descriptors of the stands' structure for the FH models for Case 1.

Case 2: Fay–Herriot Models to Update Inventories

The second case scenario consists of updating an old remote-sensing-based inventory using VRP ground measurements and FH models. For this case, auxiliary variables are stand-level predictions from a previous map and Landsat-based indexes of disturbances for the period between the old map and the date for which updated estimates were sought.

For our analyses, we used the 30 m resolution AGB map developed by Fekety and Hudak (2019) for 2013 as an old remote-sensing-based inventory. This map, CMS1-AGB map hereafter, was created in the context of the NASA Carbon Monitoring System project described in Hudak et al. (2020) using a two-step process. The first step consisted of using fixed radius plots from a set of lidar acquisitions across the northwestern United States that did not include the study area, to develop traditional ABA models where AGB was expressed as a function of lidar, topographic, and climate metrics. This model was developed at a regional scale and the plots used in the training stage included forested areas with structures and species compositions that were similar to those observed in the study area. A sample of lidar predictions in those lidar acquisitions was later used to develop a regional model to predict AGB across the forested region of the northwest United States. This regional model was primarily based on a climate metrics and Landsat time-series and was used to generate annual predictions of AGB for the period 2000–2016 at a 30 m resolution (Hudak et al., 2020). Pixel level predictions from the 2013 CMS1-AGB map were aggregated at the stand level to produce stand-level means and standard deviations of AGB predictions for 2013. These values are descriptors of the state of the forest at the moment of completion of the old inventory and are not considered to be true stand values for 2013 but approximated ones that can be used as auxiliary variables for the FH models for Case 2.

To account for changes between 2013 and 2018, we introduced additional auxiliary variables potentially correlated with growth, removals, or disturbances between 2013 and 2018 in the stands of the study area. For every stand in the study area, we computed changes between 2013 and 2018, for stand-level means, standard deviations, and modes of: (1) the red, green, blue, near-infrared, and short wave infrared one and two Landsat 8 bands, (2) band ratios including the normalized difference vegetation index (Rouse et al., 1974), NDVI, and normalized burn ratio index (Key and Benson, 2006), NBR, and (3) the brightness, wetness, and greenness tasseled cap components (Kauth and Thomas, 1976). Landsat scenes used to compute Landsat predictors correspond to the worldwide reference system path-rows 45–29 and 46–29. Median values of each Landsat band for the period going from the first of June to the 30th of September of the corresponding year were obtained and used to compute the derived indexes for each year. Stand-level means, standard deviations, and modes for bands and indexes were computed and the differences between the values obtained for 2018 and 2013 were used as auxiliary variables for the FH models. Finally, for each stand, we computed the number and proportion of pixels identified as disturbed during the period 2013–2018 by the landscape change monitoring system (LCMS) map, and the average disturbance value of all pixels identified as disturbed within the stand. The identification of disturbed pixels in LCMS is based on time series analysis of Landsat images to segment spectral trajectories. Segmentations and disturbance identification are performed with an essemble of algorithms [i.e., LandTrendr (Kennedy et al., 2010), VeRDET (Hughes et al., 2017), and CCDC (Zhu and Woodcock, 2012)]. The magnitude of the disturbances were derived as 2013–2018 changes in the relativized differenced normalized burn ratio RdNBR (Miller and Thode, 2007). In total, 38 predictors were available for Case 2. Two were the mean and standard deviation of the 2013 CMS1-AGB predictions, 18 were stand level summaries of changes in Landsat bands, 15 were stand level summaries of changes in Landsat spectral indexes and the last three were the number, proportion, and average magnitude of the disturbance metrics reported by LCMS.

Model Selection

For both scenarios, the final number of auxiliary variables available for the modeling was large (i.e., 268 for Case 1 and 38 for Case 2) and a model selection step was necessary. The model selection was performed for each combination of case scenario (i.e., Case 1 vs. Case 2), stand-type (i.e., natural vs. managed), and response variable (i.e., AGB vs. V) separately. The model selection consisted of a first step in which we used an automatic variable selection approach using best subsets regression and the R-package leaps (Lumley, 2020). In this step, we directly regressed direct estimates for AGB and V against each case's stand-level auxiliary variables to select candidate combinations auxiliary variables. Selected combinations of auxiliary variables had lengths that ranged from 1 to 6 variables. For each number of variables, the five combinations with the lowest adjusted R^2 when directly regressing against the direct ground estimates were kept. This resulted in a list of 30 candidate combinations of predictors for each case scenario, stand type, and response variable. We obtained the corresponding FH models for each candidate in these lists using the R package sae (Molina and Marhuenda, 2015) using REML. Finally, the modeler selected the model to use for each case scenario, stand type and response variable, based on the estimated model error variance, $\widehat{\sigma}_v^2$, the significance of the $\widehat{\beta}$ coefficients,

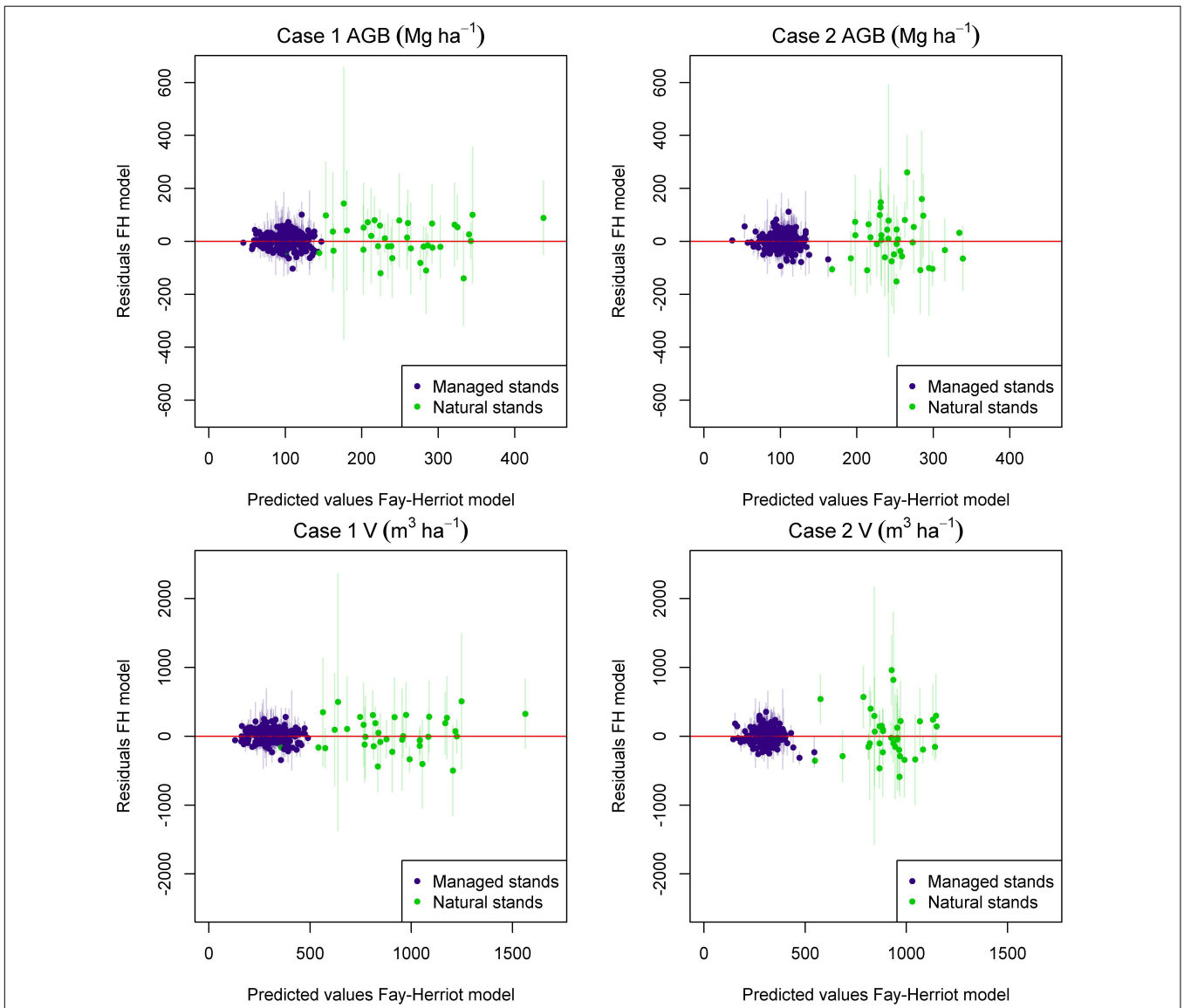


FIGURE 4 | Predicted vs. residuals plots for FH models for above-ground biomass (AGB) and volume (V) for Case 1 and Case 2. Whiskers around each point with a width of 1.96 times the standard deviation of the direct ground estimate are included to reference the uncertainty of the field estimates associated with each data point. Residuals were computed as $\widehat{\mu}_{gi*} - \mathbf{x}_i^t\widehat{\beta}$, with $\widehat{\mu}_{gi*}$ the direct ground estimate for the stand and $\mathbf{x}_i^t\widehat{\boldsymbol{\beta}}$ prediction entirely based on the model (prediction before computing the EBLUP).

the Bayesian information criterion, BIC, and predicted vs. observed diagrams.

RESULTS

Selected Models

Different patterns were observed regarding the selected models for Case 1 and Case 2 and for natural and managed stands. Regardless of the case and variable of interest, residuals for both cases tended to be centered around zero, and no significant departures with respect to the model assumptions were observed. The variability of these residuals was substantially larger for natural stands than for managed stands. Models for Case 2 tended to provide a shorter range of predicted values when compared with the models for Case 1 (**Figure 4**). In general, models for natural stands had a smaller number of predictors. This result was expected. The number of sampled stands is substantially smaller for natural stands than for managed stands, therefore estimated model coefficients for natural stands tend to have larger standard errors and less coefficients appeared as significant in the fitted models. The intercepts for models for natural stands for Case 1 were not significant and were removed. For both managed and natural stands, models for Case 1 had lower values of $\widehat{\sigma}_v^2$. For managed stands, $\widehat{\sigma}_v^2$ for AGB and V of the FH models for Case 2 were 26 and 42% larger than $\widehat{\sigma}_v^2$ for Case 1, respectively. For natural stands, we observed a 4.14 and 3.27-fold increase in $\widehat{\sigma}_v^2$ for AGB and V when comparing the values obtained for Case 2 with those obtained for Case 1 (**Tables 2**, **3**). This indicates FH models based on a recent lidar acquisition explain more variance than the models for Case 2 (**Tables 2**, **3**). When comparing models for managed and natural stands obtained under a given case scenario, we observed that for Case 1, $\widehat{\sigma}_v^2$ for AGB and V in natural stands was 2.35 and 3.48 times larger than in managed stands, respectively. For Case 2, models for natural stands explained almost no variance, and for AGB and V, $\widehat{\sigma}_v^2$ was 7.73 and 8.02 times larger than the one obtained in managed stands.

Stand Level Estimates

Selected models for Case 1 and Case 2 were used to obtain stand-level estimates and their associated mean squared errors for sampled and unsampled stands (**Figure 5**). For both cases and response variables, estimates based on FH models for

TABLE 2 | Summary of selected models for Case 1 and Case 2 for above-ground biomass, AGB (Mg ha^{-1}), and volume, V (m^3 ha^{-1}).

Stand type	Variable	Case	Auxiliary variable	$\widehat{\beta}$	std.error	*t*-Value	*p*-Value	$\widehat{\sigma}_v^2$	$\frac{\widehat{\sigma}^2_{vCase2}}{\widehat{\sigma}^2_{vCase1}}$
Natural	AGB	Case 1	Mean(Cov48to100m)	662.48	247.56	2.68	7.45E−03	1323.95	4.14
			Mean(1st_elev_mode)	6.08	1.25	4.87	1.09E−06		
			Sd(elev_ave)	15.27	5.59	2.73	6.28E−03		
		Case 2	Intercept	276.34	19.91	13.88	8.63E−44	5479.22	
			Diff_Sd(NDVI)	−3.71	1.78	−2.09	3.69E−02		
			Diff_Sd(Blue)	3.58	1.52	2.35	1.86E−02		
	V	Case 1	Mean(Cov48to100m)	2266.71	947.29	2.39	1.67E−02	23339.52	3.27
			Sd(elev_ave)	64.61	21.45	3.01	2.59E−03		
			Mean(1st_elev_mode)	20.07	4.81	4.17	3.07E−05		
		Case 2	(Intercept)	814.81	120.66	6.75	1.45E−11	76354.06	
			Diff_Mode(B)e	0.79	0.40	1.97	4.83E−02		
			Diff_Mode(Blue)	−2.97	1.41	−2.11	3.49E−02		
Managed	AGB	Case 1	Intercept	50.21	20.83	2.41	1.60E−02	563.33	1.26
			Mean(1st_cov_ab_mean)	7.16	1.56	4.58	4.70E−06		
			Mean(all_1st_cov_ab_mean)	−4.89	1.32	−3.69	2.21E−04		
			Mean(prop_6-9m)	−369.68	81.16	−4.55	5.24E−06		
		Case 2	Intercept	171.28	10.32	16.59	8.24E−62	708.58	
			Sd(PRED_AGB)	−0.58	0.10	−5.89	3.76E−09		
			Diff_Mean(W)	−0.66	0.15	−4.31	1.62E−05		
			Diff_Mean(NDVI)	−1.15	0.40	−2.85	4.39E−03		
			Diff_Mean(NBR)	1.92	0.50	3.85	1.19E−04		
	V	Case 1	Intercept	143.53	60.51	2.37	1.77E−02	6699.72	1.42
			Mean(1st_cov_ab_mean)	29.96	5.69	5.26	1.43E−07		
			Mean(all_1st_cov_ab_mean)	−20.76	4.85	−4.28	1.87E−05		
			Mean(prop_9-12m)	−1577.41	240.47	−6.56	5.39E−11		
		Case 2	Intercept	507.54	31.56	16.08	3.36E−58	9519.56	
			Diff_Sd(PRED_AGB)	−1.75	0.35	−5.03	4.89E−07		
			Diff_Mean(swir1)	0.49	0.12	4.25	2.16E−05		

Auxiliary variables were computed applying a function to rasterized layers (lidar metrics, Landsat bands, and predicted biomass) to summarize pixel level values and produce stand level metrics. Resulting metrics are indicated using the following naming convention Function(layer). Functions and layers are described in ***Table 3****.*

TABLE 3 | Stand-summarizing functions applied to the 30 m resolution layers of auxiliary variables and description of metrics included in the selected models.

Summarizing functions		Layers		
		Case	Acronym	Description
Mean	Mean of pixel level values within the stand	Case 1	1st_cov _ab_mean	Number of first returns above mean\total number of first returns
Sd	Standard deviation of pixel level values within the stand		all_1st_cov _ab_mean	Number of returns above mean\total number of first returns
			Cov48to100m	Number of first returns with heights between 48 and 100 m\total number of first returns
Mode	Mode of pixel level values within the stand		prop_9-12m	Proportion of returns between 6 and 9 m
			prop_6-9m	Proportion of returns between 9 and 12 m
Diff_Mean	Difference between 2013 and 2018 means of pixel level values within the stand, i.e., Diff_Mean(Layer)= Mean(Layer-2018)-Mean(Layer-2013)		1st_elev_mode	Mode of elevation of first returns
			elev_ave	Average of elevation returns
		Case 2	PRED_AGB	CMS predicted biomass 2013
Diff_Sd	Difference between 2013 and 2018 standard deviations of pixel level values within the stand, i.e.,Diff_Sd(Layer)= Sd(Layer-2018)-Sd(Layer-2013)		NBR	Normalized burn ratio
			NDVI	Normalized difference vegetation index
			SWIR1	Band 6. Short-wave infrared, 1.57–1.65 μm
			W	Wetness tasseled cap index
Diff_Mode	Difference between 2013 and 2018 modes of pixel level values within the stand, i.e., Diff_Mode(Layer)= Mode(Layer-2018)-Mode(Layer-2013))		B	Brightness tasseled cap index
			Blue	Band 1. Blue band, 0.441–0.514 μm

managed stands tended to be smaller than estimates for natural stands and the same pattern was observed for the corresponding *rmse* (**Figure 5**). When considering *rrmse*, managed stands showed larger relative uncertainties. This is partly caused by the fact that managed stands stock substantially less AGB and V.

For both AGB and V, estimates and *rrmse* obtained for Case 1 tended to agree spatially with estimates and *rrmse* for Case 2 (**Figure 5**). For natural stands, Spearman rank correlation between estimates for Case 1 and Case 2 was 0.22 (p-value = 1.04×10^{-8}) for AGB and 0.12 (p-value = 1.39×10^{-3}) for V. The low agreement for natural stands seems to be caused by the low explanatory power of the models for Case 2 in natural stands. For managed stands, Spearman rank correlation between estimates for Case 1 and Case 2 were 0.69 (p-value $< 10^{-6}$) for AGB and 0.50 (p-value $< 10^{-6}$) for V, and when each map was grouped into 10 deciles, these categories tended to coincide. The same occurred with the estimated *rmse* maps (see **Supplementary Figure 1**).

Efficiency Improvements in Sampled Stands

For both types of stands, FH models for Case 1 and Case 2 provided improvements for direct ground estimates for AGB and V. Estimates from Case 1 were consistently more precise than those from Case 2 (**Figure 6**). This result was expected after observing the values obtained for $\widehat{\sigma}_v^2$ for the different models. In general, improvements in efficiency and differences between cases were larger for natural stands (**Figure 6**). For natural stands, improvements in efficiency for Case 1 had an average $\triangle_{eff\ i}$ of 37.36% for AGB and 33.10% for V (**Table 4**). For Case 2, the average of $\triangle_{eff\ i}$ was 20.19% for AGB and 19.25% for V. For managed stands, the average of $\triangle_{eff\ i}$ for Case 1 was 20.29% for AGB and 19.91% for V, and for Case 2, the average of $\triangle_{eff\ i}$ was 17.55% for AGB and 16.05% for V (**Table 4**). The smaller values of $\triangle_{eff\ i}$ in managed stands is explained by the larger homogeneity of this type of stands for which many of the direct ground estimates were already precise, leaving little room for improvements to the FH models. Differences in $\triangle_{eff\ i}$ between cases for managed stands were smaller than the differences for natural stands. This seems to be the consequence of both the low explanatory power of the auxiliary variables for Case 2 in natural stands and the smaller room for improvements in managed stands.

DISCUSSION

This study presents and analyzes two possible case scenarios where FH models can be used to assist forest inventories with remote sensing information. We compared results for different stand typologies and case scenarios. We start this section by discussing the differences between cases and stand typologies and then address general issues related to the use of FH models in forest inventories.

Differences Between Case 1 and Case 2

When both scenarios were compared, estimates for Case 1 had, in general, lower errors than estimates from Case 2. The differences between cases were more important for natural stands than for managed stands. Multiple studies have shown that forest structural attributes correlate better with lidar auxiliary information than auxiliary variables from optical sensors.

TABLE 4 | Summary of stand-level estimates, uncertainties, and improvements in efficiency for FH models by type of stands (i.e., natural vs. managed stands), response variable (i.e., above-ground biomass, AGB, and volume, V), and case scenario (i.e., Case 1 and Case 2).

Stand type	Variable	Case	Sampled/ unsampled	Mean estimate	SD estimates	Mean *rmse*	Mean *rrmse* (%)	Mean Δ_{effi} (%)
Natural	AGB (Mg ha^{-1})	Case 1	Sampled	248.23	75.58	31.01	6.49	37.36
			Unsampled	212.56	75.47	40.35	18.73	
		Case 2	Sampled	250.59	67.54	44.18	8.45	20.19
			Unsampled	242.96	44.10	78.70	14.90	
		Ground	Sampled	259.05	95.76	64.52	12.67	
	V (m^3 ha^{-1})	Case 1	Sampled	906.90	278.28	119.22	1.88	33.10
			Unsampled	778.54	265.99	166.82	6.01	
		Case 2	Sampled	914.56	252.84	158.80	2.29	19.25
			Unsampled	933.89	188.17	298.00	4.08	
		Ground	Sampled	944.53	362.41	230.60	3.46	
Managed	AGB (Mg ha^{-1})	Case 1	Sampled	101.63	27.40	13.48	17.33	20.29
			Unsampled	64.44	27.17	25.03	105.61	
		Case 2	Sampled	102.06	27.51	14.12	18.18	17.55
			Unsampled	86.10	34.69	28.23	105.82	
		Ground	Sampled	103.94	34.04	19.00	26.45	
	V (m^3 ha^{-1})	Case 1	Sampled	302.90	99.86	45.74	7.88	19.92
			Unsampled	207.41	78.25	86.23	30.77	
		Case 2	Sampled	303.57	98.61	48.88	8.25	16.05
			Unsampled	284.89	77.99	99.96	15.57	
		Ground	Sampled	312.02	120.56	64.50	11.18	

Auxiliary information for Case 1 proceeded from lidar. For Case 2, we used a previous remote sensing-based inventory the CMS1-AGB map, which heavily relies on metrics derived from 30-year climate normals, topographic, and Landsat variables, to which we added proxies for disturbances directly derived from Landsat images. This explains that Case 1 outperforms Case 2 for all response variables and stand types. Nevertheless, estimates from Case 2 are more efficient than direct ground estimates and regardless of the case, the rank correlations between estimates for Case 1 and Case 2 for managed stands indicated that both methods agree in the way the sort stands according to the predicted AGB or V. These results indicate that Case 2 is also useful for managed stands, and that certain management decisions, for example, concentrating harvest activities in the 10% of the managed stands with more volume, would tend to coincide regardless of which map (i.e., Case 1 or Case 2) is used to inform those decisions (see **Supplementary Figure 1**).

Two remarks should be made about the initial map for Case 2. On the one hand, the CMS1-AGB map has spatial and temporal coverage that cannot be matched by previous maps based only on lidar. Thus, this map can be used to develop similar stand-level inventories anywhere in the western United States. Furthermore, the multitemporal component of this map allow for possible applications of FH models to estimate changes and monitor vegetation dynamics that are not an option using single date lidar data. On the other hand, the CMS1-AGB map is expected to provide predictions with more noise than similar maps based on the ABA method and lidar data. This implies that results obtained in this study for Case 2 might improve substantially when the previous inventory is an ABA lidar-based inventory. Many countries have developed nationwide lidar acquisitions or are on the verge of completing such data collections, and national forest inventories can provide the necessary fixed area plots to use the ABA to develop maps based on lidar at national or regional scales. The effort required to develop these maps is large, and re-mapping is not expected to happen with a high frequency. This indicates that a potential niche of application of FH models and Case 2 is updating national or regional level ABA maps.

Differences Between Natural and Managed Stands

When comparing natural and managed stands, we observed that the former had larger estimated model variances, resulting in stand-level estimates with larger uncertainties in absolute terms. These differences are explained by the fact that natural stands are inherently more complex and variable than managed stands. Part of that complexity is not captured by predictors computed at the stand level. Relative uncertainties (i.e., *rmse*) were lower for natural stands than for managed stands. The higher stocking levels cause that in natural stands. Improvements in efficiency for natural stands were larger than those observed for managed stands, especially for Case 1. Finally, the differences in $\triangle_{effi}$ for Case 1 and Case 2 were relatively small for managed stands (i.e., about 3% difference between average values of $\triangle_{effi}$) but large for natural stands (i.e., approximately 15% difference between average values of $\triangle_{effi}$). The interaction of three different factors can explain this differentiated behavior. The first is that managed stands are relatively homogeneous units, and their direct ground estimates were more reliable than those obtained in natural stands. Thus, the potential for improved

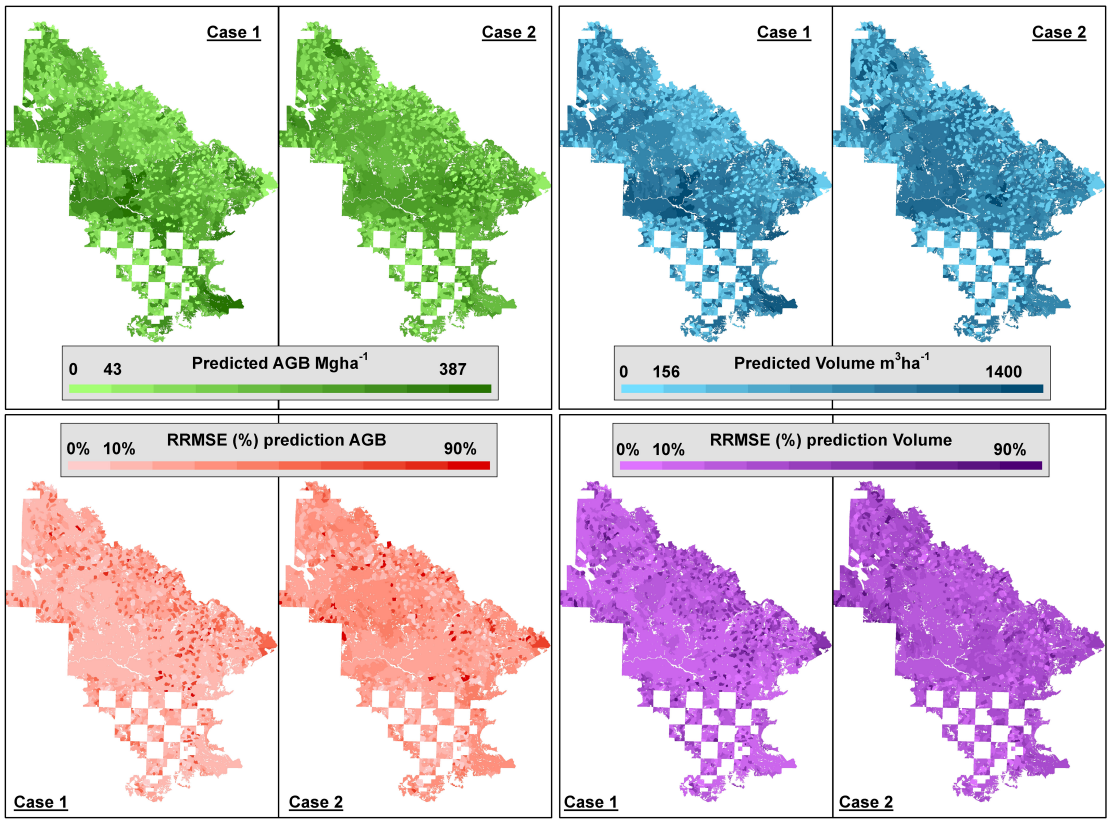


FIGURE 5 | Stand-level predictions of above-ground biomass (AGB) and volume (V) for all stands in the study area (sampled and unsampled) and stand-level RMSE based on the FH models for Case 1 and Case 2.

managed stands was more limited and made differences between cases smaller. Another factor is that the larger stocking levels are frequently associated with remote sensing predictions with larger uncertainties (Magnussen et al., 2014; Mauro et al., 2016; Breidenbach et al., 2018). Large uncertainties in the previous inventory must result in a poorer characterization of the initial state of the stands, which partially explains the performance drop for Case 2 in natural stands. The auxiliary information for Case 2 is primarily based on metrics derived from 30-year climate normals to capture steep AGB (and V) gradients, which largely compensated for signal saturation of Landsat variables (Hudak et al., 2020), albeit without sensitivity to local variation in stand structure. On the other hand, the lidar data used for Case 1 neither saturates in forested areas with closed canopies nor is insensitive to structure variation between or within stands, thus elevating the performance of the FH models for Case 1 compared to Case 2.

General Considerations for the Use of Fay–Herriot Models in Forest Inventories

For all cases, response variables and stand types analyzed in this study, FH models allowed for gains in efficiency with respect to direct ground estimates. These results are concurrent with previous research using FH models in forest inventories (Goerndt et al., 2011; Magnussen et al., 2017; Mauro et al., 2017; Breidenbach et al., 2018; Ver Planck et al., 2018) and confirm that FH models: (1) allow using ground measurements that are easier to obtain than those used in ABA approaches and (2) results in efficiency improvements when compared to methods based only on ground data. Thus, while research efforts on using FH models are still necessary, there is substantial evidence that these models can play an essential role in operational forest inventory applications.

To the best of our knowledge, FH models have not been used in any operational forest inventory, and it somehow surprises how little attention FH models have received in the literature. While some research applications of FH models exist, the study of this type of model has been negligible compared to applications using conventional ABA approaches. The dominance of the traditional ABA approach can be explained by (1) its ability to produce high resolution maps with predictions of forest attributes and (2) its typically better predictive performance than FH models (Mauro et al., 2017; Breidenbach et al., 2018; Green et al., 2019). However, developing ABA models is not always a possibility. There are many scenarios where FH models can be a very appealing alternative; for example, only stand-level inventory data may be available. During the last decades, forest inventories have been consistently less constrained by the availability of useful auxiliary information, but the costs

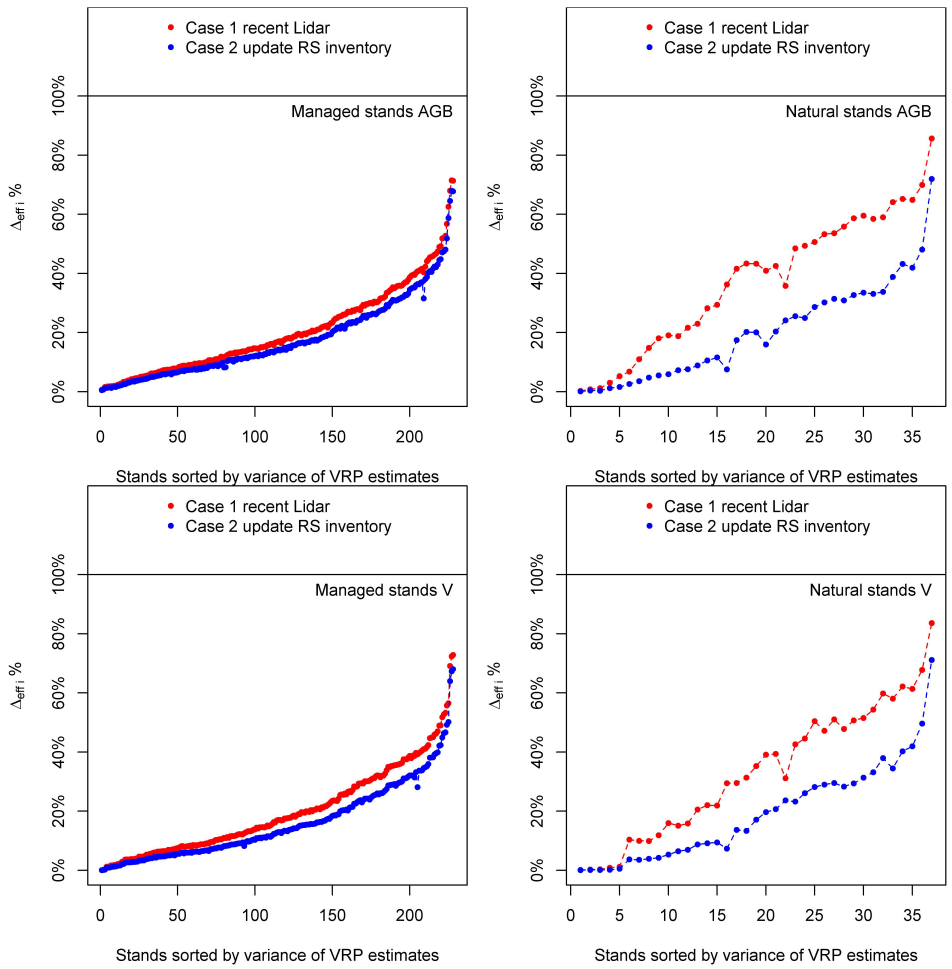


FIGURE 6 | Reduction of uncertainty achieved by Fay–Herriot models with respect to an inventory based only on ground data(Δ_{effi}) for Case 1 and Case 2 for above-ground biomass (AGB) and volume (V).

associated with ground data collection have increased, or at least not decreased at comparable rates. Simply put, ground data are too valuable to ignore, and FH models allow for an effective combination of those valuable datasets with different sources of auxiliary information.

A critical difference between traditional ABA models and FH models is the auxiliary information used by each technique. Workflows for preprocessing auxiliary information for traditional ABA models are well established and documented, with multiple tools available to implement these processing steps (i.e., Mc Gaughey, 2019; Roussel et al., 2020); this is not the case for FH models. In this study, we used as predictors stand summaries of: (1) gridded products (i.e., gridmetrics rasters) generated with FUSION (Mc Gaughey, 2019), (2) previously mapped estimates of forest attributes, and (3) changes in Landsat imagery from LCMS or computed using Google earth engine (Gorelick et al., 2017). Summarizing the entire point clouds within the stands under analysis is an alternative used in previous studies to compute lidar-based predictors (Ver Planck et al., 2018). Both options are valid from a methodological perspective as they provide standardized ways to compute auxiliary variables. Their effectiveness can differ if one preprocessing technique provided auxiliary variables that correlated better with the target responses than the other. However, as far as we know, no study to date has analyzed the differences in performance and tradeoffs of these two methods to generate stand-level predictors for FH models. Thus, this is an area where future research can help in establishing standardized processing workflows for lidar-assisted forest inventories using FH models.

This study presents two case scenarios in which basic FH models are used with VRP and demonstrates that FH models are a suitable alternative to use available auxiliary information to improve the efficiency of the estimation process. Our analysis presents a baseline for stand-level FH models and could be improved in different ways. One way is developing models that account for spatial correlations like those developed by Ver Planck et al. (2018). Another option is to use FH variants where the model variance is not constant (Breidenbach et al., 2018). A third option is to use

multivariate FH models where correlations between different response variables can be considered to improve the results of univariate models (Benavent and Morales, 2016; Frank, 2020). In all cases, one factor that must be constantly considered is that estimates obtained from FH models are always based on a model (i.e., "model-based"). Thus, extrapolations entail high risks of producing biased estimates and model validation steps are critical to ensure that the fitted models correctly describe the populations under study.

Based on our findings and previous results (Goerndt et al., 2011; Breidenbach et al., 2018; Ver Planck et al., 2018), we envision that niche of application of stand-level FH models is not a replacement of traditional ABA methods but a complement for situations in which the time and resources available for ground data collection are limited or fixed radius plots with precise locations to develop ABA models are otherwise unavailable. This niche is larger than it might seem *a priori* for several reasons. One reason is that obtaining accurate coordinates for the ground measurements is not a constraint for FH models. For example, only the identifier of the stand where each ground observation was taken was necessary to develop this study using FSVeg data. Another reason, and probably the most compelling one, is that FH models can be applied with data from VRP or other sampling techniques such as sector plots (Iles and Smith, 2006) or transects (Warren and Olsen, 1964; Woodall and Monleon, 2008). This flexibility indicates numerous applications for fast inventories and monitoring problems in which FH models can be the preferred alternative. These applications include, but are not limited to, annual inventories for timber sales or fast updates of inventories after events like floods or wildland fires.

Improved AGB and V benefit both private companies and public land management agencies. Given the extremely high cost of establishing ground plots and the increasing demand for accurate biomass and carbon stock assessment, the inventory solution will require the innovative use of combined sources of remotely sensed and other auxiliary data. FH models based on VRP allow using remotely sensed information combined with an operative ground truth data collection and enable cost-effectively estimating forest attributes. In this study, the FH models have shown to be a viable and flexible option to estimate AGB or V and maximize the utility of both the ground inventory and environmental datasets. Moreover, different information for forest management planning is required at different levels or scales. For tactical planning, reasonably precise and unbiased estimates of forest variables for individual stands or polygons are already obtained using VRP because of its low cost and sampling efficiency at the stand level. Thus, FH models are an alternative for many established inventory programs to integrate their VRP data with lidar or other remote-sensing datasets to obtain more efficient and better information for sustainable forest management.

CONCLUSION

The main conclusion obtained when comparing estimates from FH models for Case 1 and Case 2 indicated that estimates from FH models based on a recent lidar acquisition were the most efficient alternative. For managed stands, differences between case scenarios were small, but in natural stands, FH models based on data from a recent lidar data collection produced more efficient results substantially. However, in all cases, estimates from FH models for both case scenarios and both types of forest stands were more efficient than direct ground estimates. Based on this result, we conclude that FH models are a valuable alternative for many forest inventory tasks if fixed area plots or their precise geolocations are unavailable.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because this study was developed using data from Field Sampled Vegetation (FSVeg) database. USDA Federal employees, contractors, and affiliates need to follow the steps indicated in the link below to access the data. https://www.fs.fed.us/nrm/documents/fsveg/cse_user_guides/FSVegQuickGuide.pdf.

AUTHOR CONTRIBUTIONS

HT wrote some parts of the manuscript, verified the analytical methods, critically reviewed the manuscript, and supervised the findings of this work. FM conceived of the presented idea, performed the computations, and wrote the first version of the manuscript. AH, BF, and VM critically reviewed the manuscript and provided critical feedback. PF, MP, and TB contributed to the final version of the manuscript. All authors discussed the results and contributed to the final manuscript.

FUNDING

This work was supported by Challenge Cost Share Agreement 20-CS-11062754-066 between Oregon State University and the USDA Forest Service, Pacific Northwest Region and by a NASA Carbon Monitoring System Program award (80HQTR20T0002) through a Joint Venture Agreement (20-JV-11221633-112) between the USDA Forest Service, Rocky Mountain Research Station and Oregon State University.

ACKNOWLEDGMENTS

We would like to acknowledge Cheryl Friesen, James Rudisill, and Karin Wolken that were an active part in discussions that led to the ideas presented in this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/ffgc.2021.745916/full#supplementary-material

REFERENCES

Babcock, C., Finley, A. O., Bradford, J. B., Kolka, R., Birdsey, R., and Ryan, M. G. (2015). LiDAR based prediction of forest biomass using hierarchical models with spatially varying coefficients. *Remote Sens. Environ.* 169, 113–127. doi: 10.1016/j.rse.2015.07.028

Benavent, R., and Morales, D. (2016). Multivariate Fay–Herriot models for small area estimation. *Comput. Stat. Data Anal.* 94, 372–390. doi: 10.1016/j.csda.2015.07.013

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

Coulston, J. W., Green, P. C., Radtke, P. J., Prisley, S. P., Brooks, E. B., Thomas, V. A., et al. (2021). Enhancing the precision of broad-scale forestland removals estimates with small area estimation techniques. *Forestry* 94, 427–441. doi: 10.1093/forestry/cpaa045

Deo, R. K., Froese, R. E., Falkowski, M. J., and Hudak, A. T. (2016). Optimizing variable radius plot size and LiDAR resolution to model standing volume in conifer forests. *Can. J. Remote Sens.* 42, 428–442.

Fay, R. E., 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, 269–277. doi: 10.2307/2286322

Fekety, P. A., Falkowski, M. J., Hudak, A. T., Jain, T. B., and Evans, J. S. (2018). Transferability of lidar-derived basal area and stem density models within a Northern Idaho Ecoregion. *Can. J. Remote Sens.* 44, 131–143. doi: 10.1080/07038992.2018.1461557

Fekety, P. A., and Hudak, A. T. (2019). *Annual Aboveground Biomass Maps for Forests in the Northwestern USA, 2000-2016*. Oak Ridge, TN: National Laboratory Distributed Active Archive Center, doi: 10.3334/ORNLDAAC/1719

Forkuor, G., Benewinde Zoungrana, J.-B., Dimobe, K., Ouattara, B., Vadrevu, K. P., and Tondoh, J. E. (2020). Above-ground biomass mapping in West African dryland forest using sentinel-1 and 2 datasets - a case study. *Remote Sens. Environ.* 236, 111496. doi: 10.1016/j.rse.2019.111496

Frank, B. M. (2020). *Aerial Laser Scanning for Forest Inventories: Estimation and Uncertainty at Multiple Scales*. Ph.D. thesis. Corvallis, OR: Oregon State University.

Frank, B., Mauro, F., and Temesgen, H. (2020). Model-based estimation of forest inventory attributes using lidar: a comparison of the area-based and semi-individual tree crown approaches. *Remote Sens.* 12:2525. doi: 10.3390/rs12162525

Goerndt, M. E., Monleon, V. J., and Temesgen, H. (2011). A comparison of small-area estimation techniques to estimate selected stand attributes using LiDAR-derived auxiliary variables. *Can. J. For. Res.* 41, 1189–1201.

González-Ferreiro, E., Diéguez-Aranda, U., and Miranda, D. (2012). Estimation of stand variables in Pinus radiata D. don plantations using different LiDAR pulse densities. *Forestry* 85, 281–292. doi: 10.1093/forestry/cps002

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. (2017). Google earth engine: planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* 202, 18–27. doi: 10.1016/j.rse.2017.06.031

Grafström, A., Schnell, S., Saarela, S., Hubbell, S. P., and Condit, R. (2017). The continuous population approach to forest inventories and use of information in the design. *Environmetrics* 28:e2480. doi: 10.1002/env.2480

Green, P. C., Burkhart, H. E., Coulston, J. W., and Radtke, P. J. (2019). A novel application of small area estimation in loblolly pine forest inventory. *Forestry* 93, 444–457. doi: 10.1093/forestry/cpz073

Hudak, A. T., Fekety, P. A., Kane, V. R., Kenedy, R. E., Filipelli, S. K., Falkowski, M. J., et al. (2020). A carbon monitoring system for mapping regional, annual aboveground biomass across the northwestern USA. *Environ. Res. Lett.* 15:095003.

Hudak, A. T., Haren, A. T., Crookston, N. L., Liebermann, R. J., and Ohmann, J. L. (2014). Imputing forest structure attributes from stand inventory and remotely sensed data in western Oregon, USA. *For. Sci.* 60, 253–269.

Hughes, M. J., Kaylor, S. D., and Hayes, D. J. (2017). Patch-based forest change detection from landsat time series. *Forests* 8:166. doi: 10.3390/f8050166

Hummel, S., Hudak, A., Uebler, E., Falkowski, M., and Megown, K. (2011). A comparison of accuracy and cost of LiDAR versus stand exam data for landscape management on the Malheur National Forest. *J. For.* 109, 267–273.

Iles, K., and Smith, N. J. (2006). A new type of sample plot that is particularly useful for sampling small clusters of objects. *For. Sci.* 52, 148–154. doi: 10.1093/forestscience/52.2.148

Kauth, R. J., and Thomas, G. (1976). "The tasselled cap–a graphic description of the spectral-temporal development of agricultural crops as seen by Landsat," in *Proceedings of the Machine Processing of Remotely Sensed Data*, (West Lafayette, IN: Purdue University).

Kennedy, R. E., Yang, Z., and Cohen, W. B. (2010). Detecting trends in forest disturbance and recovery using yearly landsat time series: 1. landtrendr — temporal segmentation algorithms. *Remote Sens. Environ.* 114, 2897–2910. doi: 10.1016/j.rse.2010.07.008

Key, C. H., and Benson, N. C. (2006). "Landscape assessment (LA)," in *FIREMON: Fire Effects Monitoring and Inventory System*, eds D. C. Lutes, R. E. Keane, J. F. Caratti, C. H. Key, N. C. Benson, S. Sutherland, et al. (Ogden, UT: US Department of Agriculture, Forest Service, Rocky Mountain Research Station), 1–55.

LeMay, V., Maedel, J., and Coops, N. C. (2008). Estimating stand structural details using nearest neighbor analyses to link ground data, forest cover maps, and Landsat imagery. *Remote Sens. Environ.* 112, 2578–2591. doi: 10.1016/j.rse.2007.12.007

Lumley, T. (2020). *Leaps: Regression Subset Selection*. Available online at: http://CRAN.R-project.org/package=leaps. (accessed October 6, 2021).

Magnussen, S., Mandallaz, D., Breidenbach, J., Lanz, A., and Ginzler, C. (2014). National forest inventories in the service of small area estimation of stem volume. *Can. J. For. Res.* 44, 1079–1090. doi: 10.1139/cjfr-2013-0448

Magnussen, S., Mauro, F., Breidenbach, J., Lanz, A., and Kändler, G. (2017). Area-level analysis of forest inventory variables. *Eur. J. For. Res.* 136, 839–855. doi: 10.1007/s10342-017-1074-z

Maltamo, M., Eerikäinen, K., Pitkänen, J., Hyyppä, J., and Vehmas, M. (2004). Estimation of timber volume and stem density based on scanning laser altimetry and expected tree size distribution functions. *Remote Sens. Environ.* 90, 319–330.

Mauro, F., Molina, I., García-Abril, A., Valbuena, R., and Ayuga-Téllez, E. (2016). Remote sensing estimates and measures of uncertainty for forest variables at different aggregation levels. *Environmetrics* 27, 225–238. doi: 10.1002/env.2387

Mauro, F., Monleon, V. J., Temesgen, H., and Ford, K. R. (2017). Analysis of area level and unit level models for small area estimation in forest inventories assisted with LiDAR auxiliary information. *PLoS One* 12:e0189401. doi: 10.1371/journal.pone.0189401

Mauro, F., Ritchie, M., Wing, B., Frank, B., Monleon, V., Temesgen, H., et al. (2019). Estimation of changes of forest structural attributes at three different spatial aggregation levels in northern California using multitemporal LiDAR. *Remote Sens.* 11:923. doi: 10.3390/rs11080923

Mc Gaughey, R. J. (2019). *FUSION\LDV: Software for LIDAR Data Analysis and Visualization*. Washington, D.C: USDA Forest Service.

Miller, J. D., and Thode, A. E. (2007). Quantifying burn severity in a heterogeneous landscape with a relative version of the delta Normalized Burn Ratio (dNBR). *Remote Sens. Environ.* 109, 66–80. doi: 10.1016/j.rse.2006.12.006

Molina, I., and Marhuenda, Y. (2015). sae: an R package for small area estimation. *R J.* 7, 81–98.

Næsset, E. (2002). Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. *Remote Sens. Environ.* 80, 88–99.

Pflugmacher, D., Cohen, W. B., and Kennedy, R. E. (2012). Using landsat-derived disturbance history (1972–2010) to predict current forest structure. *Remote Sens. Environ.* 122, 146–165. doi: 10.1016/j.rse.2011.09.025

Rao, J. N. K., and Molina, I. (2015). "Empirical best linear unbiased prediction (EBLUP): basic area level model," in *Small Area Estimation*, ed. P. Lahiri (Hoboken, NJ: John Wiley & Sons, Inc), 123–172.

Rouse, J. W., Haas, R. H., Schell, J. A., and Deering, D. W. (1974). Monitoring vegetation systems in the Great Plains with ERTS. *NASA special publication* 351, 309.

Roussel, J.-R., Coops, N. C., Tompalski, P., Goodbody, T. R. H., Meador, A. S., Bourdon, J.-F., et al. (2020). lidR: an R package for analysis of airborne laser scanning (ALS) data. *Remote Sens. Environ.* 251:112061. doi: 10.1016/j.rse.2020.112061

Vafaei, S., Soosani, J., Adeli, K., Fadaei, H., Naghavi, H., Pham, T. D., et al. (2018). Improving accuracy estimation of forest aboveground biomass based

on incorporation of ALOS-2 PALSAR-2 and sentinel-2A imagery and machine learning: a case study of the hyrcanian forest area (Iran). *Remote Sens.* 10:172. doi: 10.3390/rs10020172

Ver Planck, N. R., Finley, A. O., Kershaw, J. A., 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

Warren, W. G., and Olsen, P. F. (1964). A line intersect technique for assessing logging waste. *For. Sci.* 10, 267–276. doi: 10.1093/forestscience/10.3.267

Woodall, C., and Monleon, V. (2008). *Sampling Protocol, Estimation, and Analysis Procedures for the Down Woody Materials Indicator of the FIA Program*. Newtown Square, PA: Northern Research Station.

Zhu, Z., and Woodcock, C. E. (2012). Object-based cloud and cloud shadow detection in landsat imagery. *Remote Sens. Environ.* 118, 83–94. doi: 10.1016/j.rse.2011.10.028

Conflict of Interest: The 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Forests and Global Change

PERSPECTIVE
published: 11 November 2021
doi: 10.3389/ffgc.2021.746439

Needs for Small Area Estimation: Perspectives From the US Private Forest Sector

Steve Prisley[1], Jeff Bradley[2], Mike Clutter[3], Suzy Friedman[4], Dick Kempka[5], Jim Rakestraw[6] and Edie Sonne Hall[7]*

[1] National Council for Air and Stream Improvement, Roanoke, VA, United States, [2] American Forest & Paper Association, Washington, DC, United States, [3] Forest Investment Associates, Atlanta, GA, United States, [4] National Alliance of Forest Owners, Washington, DC, United States, [5] Molpus Woodlands Group, LLC, Jackson, MS, United States, [6] International Paper, Statesboro, GA, United States, [7] Three Trees Consulting, Seattle, WA, United States

The commercial forest sector in the US includes forest landowners and forest products manufacturers, as well as numerous service providers along the supply chain. Landowners (and contractors working for them) manage forestland in part for roundwood production, and manufacturers purchase roundwood as raw material for forest products including building products, paper products, wood pellets, and others. Both types of organizations need forest resource data for applications such as strategic planning, support for certification of sustainable forestry, analysis of timber supply, and assessment of forest carbon, biodiversity, or other ecosystem services. The geographic areas of interest vary widely but typically focus upon ownership blocks or manufacturing facilities and are frequently small enough that estimates from national forest inventory data have insufficient precision. Small area estimation (SAE) has proven potential to combine field data from the national forest inventory with abundant sources of remotely sensed or other resource data to provide needed information with improved precision. Successful implementation of SAE by this sector will require cooperation and collaboration among federal and state government agencies and academic institutions and will require increased funding to improve data collection, data accessibility, and further develop and implement the needed technologies.

Keywords: landowner, manufacturer, sustainability, markets, carbon, precision

OPEN ACCESS

Edited by:
Aaron Weiskittel, University of Maine, United States

Reviewed by:
John Paul McTague, University of Georgia, United States
Roque Rodríguez-Soalleiro, University of Santiago de Compostela, Spain
Ben Rice, Midgard Natural Resources, United States

****Correspondence:***
Steve Prisley
sprisley@ncasi.org

Specialty section:
This article was submitted to Forest Management, a section of the journal Frontiers in Forests and Global Change

Received: *23 July 2021*
Accepted: *20 October 2021*
Published: *11 November 2021*

Citation:
Prisley S, Bradley J, Clutter M, Friedman S, Kempka D, Rakestraw J and Sonne Hall E (2021) Needs for Small Area Estimation: Perspectives From the US Private Forest Sector. Front. For. Glob. Change 4:746439. doi: 10.3389/ffgc.2021.746439

INTRODUCTION

Our purpose here is to articulate the needs of the US private forest sector for enhanced forest resource information that might be possible through the application of small area estimation (SAE), combining plot data from the national forest inventory with supplemental data from remote sensing or other geospatial sources. We use the term "small area estimation" to refer to a suite of statistical approaches to improve the precision of forest inventory estimates for small geographic areas or categories by incorporating additional data beyond the plot measurements themselves.

The private forest sector is a dominant contributor to many aspects of forestry in the US. In this article, we use the term "private forest sector" to include manufacturers of forest products and private timberland owners. The term "working forests" has been used widely to refer to forests that

are managed to provide a steady supply of wood to forest products markets while providing other ecosystem benefits. However, there are no publicly available data that can distinguish forest areas based on management intent of private owners. We therefore use the term "private timberlands" as the closest approximation to "working forests." We adopt the definition of "timberland" from the national forest inventory, which defines it as forest land capable of producing at least 20 cubic feet of wood per acre per year and not reserved from harvest.

Private timberland owners (organizations and families) own nearly 60% of forest land, provide nearly 90% of wood harvested for products, and account for more than 80% of forest volume growth (Oswalt et al., 2019). Forest products manufacturers account for approximately four percent of the total US manufacturing GDP, produce nearly $300 billion in products annually and employ approximately 1 million people. The industry meets a payroll of approximately $60 billion annually and is among the top 10 manufacturing sector employers in 45 states.

The private forest sector is also a primary contributor to natural climate solutions. Private timberlands store more than half of the forest carbon in the US, and account for nearly 75% of annual gross sequestration. Sustainable harvest of wood for products has led to increasing forest carbon stocks and increasing carbon storage in harvested wood products. Numerous studies confirm that active markets for wood provide an incentive for landowners to maintain or increase forest area and productivity (Lubowski et al., 2008; Abt et al., 2010, 2014; Costanza et al., 2016; Dale et al., 2017; Jefferies and Leslie, 2017; Birdsey et al., 2018; Kim et al., 2018).

The private sector is an extensive user of information from the national forest inventory, housed in the Forest Inventory and Analysis (FIA) program of US Forest Service Research. The FIA program conducts annual field inventory across all ownerships in the US, as well as surveys on mill production (through the Timber Product Output program) and forest owners (through the National Woodland Owner Survey). The field inventory is conducted on permanent plots across all ownerships at a sampling intensity of one plot per 2,400 hectares (5,937 acres). A subset of plots in all states is remeasured annually such that all plots in the eastern US are remeasured every 5–7 years, and plots in the west are remeasured every 10 years.

National forest inventory data is also widely used by carbon credit programs to assess baseline forest carbon levels. For example, the California Cap and Trade Program, the largest carbon market for private investors and companies in the US, uses FIA data to determine forest carbon project baselines and the associated volume of credits generated each year. The Family Forest Carbon Program[1] of the American Forest Foundation and The Nature Conservancy also uses FIA data to establish baseline carbon levels and to measure the performance of forest management practices.

Many users of FIA data have expressed expanding needs for more detailed information on smaller spatial domains while maintaining the core field measurement program (Guldin, 2020a,b). The FIA program has responded with increased research activity in SAE (for example, Brooks et al., 2016; Nagle et al., 2019; Green et al., 2020; Coulston et al., 2021). These efforts have shown that precision can be improved using SAE with plot data combined with ancillary data. While such studies have demonstrated the promise of SAE, none have led to operational production of national datasets.

While FIA is budget-constrained and is currently challenged to maintain even the present level of sample intensity (geographically and temporally), there are abundant remote sensing and geospatial data that could lend increased precision to inventory-based estimates of forest resources. Many of these ancillary data layers and sources may already be used by organizations in the forest sector: soils data, satellite and aerial imagery, canopy heights from LIDAR or aerial photogrammetry, topography, hydrography, land cover, and numerous others. The pressing needs of the private sector for higher-precision resource information argue for further research in SAE methods and delivery of tools to apply these methods.

In the following sections, we will first provide examples of reasons why the private forest sector is facing increasing needs for reliable forest resource information. Then we discuss the specific estimates from FIA data that can meet these needs, with examples of current levels of precision of these estimates. Finally, we highlight opportunities for improvements that would enhance the value of FIA data for the private forest sector as well as many other users.

EXAMPLES OF WHY INFORMATION IS NEEDED

Assessment of Forest and Carbon Sustainability

Sustainability certification and reporting are critical for forest sector companies to document their performance against sustainability standards. Forest sector organizations are increasingly investing time and effort in reporting Environmental, Social, and Governance (ESG) indicators to communicate to customers and investors that sustainability is embedded in their business practices.

For manufacturers certifying the sustainability of their supply chain, this reporting leads to an increased need for resource data from the geographic regions in which they operate, which is often met using data from FIA. For example, both the Sustainable Forestry Initiative (SFI) and Forest Stewardship Council (FSC) certification standards include requirements that companies avoid the use of wood that may come from forests that have been converted to non-forest use (deforestation). This requires regular monitoring of forest land use changes within operating regions. Unfortunately, FIA-based estimates of forest area change for smaller regions (e.g., small states or woodbaskets) may fail to reach

[1] www.forestfoundation.org/family-forest-carbon-program

precision targets defined by certification standards, such as demonstrating with statistical significance that forest area is not declining.

For private timberland owners with detailed data on their own holdings, FIA data may be used to obtain factors for tree species in a region to convert inventory volumes to carbon stocks, and to estimate carbon in pools other than live trees. Furthermore, trade associations representing segments of the forest sector make extensive use of FIA data to communicate the sustainability of forests and their contributions to meeting environmental goals.

Other certification requirements involve attention to the quantity of wood harvested relative to the quantity grown. Such growth/drain analyses are common, but require estimates of change over time. Because harvest is a relatively rare event across large landscapes, the sample size for estimates of harvest are small, leading to higher uncertainties. Demonstrating that growth exceeds removals with statistical significance is often difficult for some areas using FIA plot data alone.

New guidelines are being developed for companies to report on value chain effects of their products on terrestrial carbon dynamics. One proposed metric involves carbon stock changes on lands from which they obtain raw materials. In a forestry context, this would require carbon stock estimates at a regional or woodshed level from two successive inventories and expressing that stock change relative to the quantity of wood harvested. Without employing SAE approaches, such estimates have high uncertainties.

Wood Markets

Landowners and manufacturers are engaged in markets for roundwood from forests, as sellers and buyers, respectively. For both, it is essential to understand the market dynamics in their operating regions to plan effectively. This entails knowing the relationship between forest area change, forest growth, mortality, and harvest within a geographic area. These dynamics are critical to evaluating long-term resource availability and sustainability.

Land use change is a longer-term driver of wood markets and can affect the availability and cost of wood in rapidly developing areas. Similarly, economic disruptions to local wood markets can occur when established mills cease or reduce operations, or when new mills begin operations. Catastrophic events such as fire, hurricanes, or drought can quickly and dramatically alter local resource availability. Therefore, companies must monitor wood market conditions within their operating areas, requiring information on harvest levels, mortality, land ownership changes, forest area changes, and forest growth rates.

Forest Carbon Markets

The potential for forests in the US to contribute to natural climate solutions has spawned interest and activity in forest carbon offset markets. Such markets are designed to incentivize forest owners to increase average forest carbon stocks through payments for carbon offsets. To produce real climate benefits, forest carbon offset markets need to account for (a) carbon stored in products as well as forests, (b) additionality (benefits above and beyond business-as-usual behavior), (c) leakage (emissions that occur due to increased harvests outside a project that compensate for reduced harvests within a project), and (d) substitution (higher emissions resulting from the use of carbon-intensive substitute products such as concrete or steel in place of wood-based building products). Addressing these considerations requires data on initial forest carbon stocks for project areas, forest growth rates, levels of harvest associated with a "business-as-usual" or "standard practice" baseline, eventual use of harvested wood within the region (proportion of harvest going to lumber, panels, paper, fuel, etc.), and market factors related to leakage and substitution (such as supply and demand elasticities). National forest inventory data can meet some of these information needs for large areas, but some estimates will lack needed precision for smaller geographic areas.

Biodiversity at Landscape Scale

Private timberland owners and manufacturers recognize the importance of conducting forest management activities in a way that conserves habitat for species of conservation concern. A first step in doing so is understanding the geographic distribution of forest conditions associated with individual species.

Forest inventory data can be used to assess the relative quality of habitat for some species by quantifying relevant aspects of stand structure. For example, in a protocol developed to assess quality of open-canopy pine forests for species of concern in the US South (Nordman et al., 2016), metrics include proportion of basal area in pine trees (of certain species) in specified diameter ranges, proportion of basal area in hardwood trees, percent canopy cover from pine species, and stand density index. Similarly, Davis et al. (2015) describe an old-growth structure index (OGSI) for the Pacific Northwest derived from inventory metrics such as density of live trees above a diameter threshold, density of standing dead trees above a diameter threshold, percent cover by down dead wood of certain size, and an index of tree diameter diversity.

While protocols and indices such as these can be applied to FIA data and be extremely useful in broad-scale monitoring of structural diversity at a landscape scale, the categorical domains can be very narrow (e.g., trees per hectare greater than 100 cm diameter). Obtaining estimates of uncertainty for indices involving multiple metrics can quickly become intractable, and uncertainties will almost certainly be high even across large geographical areas. Davis et al. (2015) noted sources of uncertainties but were not able to quantify general levels of uncertainty in results.

If suitable ancillary data are available (from FIA or other publicly accessible sources) to lend strength to some of the estimates needed for these indices, then SAE approaches may prove valuable in quantifying uncertainties and improving precision of estimates related to biodiversity.

TYPE OF FOREST RESOURCES DATA NEEDED

Clearly, the forest sector needs current, reliable data on the forest resources they manage or depend on for raw materials.

Data needs include forest area and change over time, estimates of relevant resource quantities (e.g., wood volume, biomass, carbon) and the current rates of change of those quantities. These rates of change include categories such as forest growth, mortality, and harvest. Information regarding the geographic distribution of resources is important, especially as it relates to transportation networks, manufacturing facilities, population centers, and features that affect management practices, such as steep slopes, soil erodibility, wetlands, and habitat for species of conservation concern.

Typically, resource information needs are limited geographically to the operating regions for individual organizations. For timberland owners, it would be the areas in and around their forest holdings. For manufacturers, it would be areas within a sourcing region for each of their facilities. These regions are often small enough that there are insufficient numbers of FIA samples to provide inventory-based estimates with reasonable precision in categories of interest.

To illustrate the levels of uncertainty of commonly used estimates available from FIA plot data for typical operating areas, we developed estimates for an 80 km (50 mile) and 160 km (100 mile) radius around an arbitrary location in the US South near the Georgia-Alabama border (**Table 1**). In this example, we consider either a landowner or a manufacturer interested in softwood sawtimber available from private timberlands in the operating area. Therefore, relevant information would include area in pine forest types, pine forest area by age class (for modeling future supplies), softwood sawtimber growing stock volumes, and growth and removals of softwood sawtimber. For all variables, we retrieved summary information using the USFS EVALIDator tool (USDA Forest Service, 2021) accessing 2019 inventories for Georgia and Alabama.

Ninety-percent confidence intervals on softwood sawtimber volumes on private lands are ± 12.7% and ± 6.6% for the 80 and 160 km radius areas, respectively. Note that the sample size for harvest removals is only 23–25% of the sample size for private pine timberland area. Samples with harvests represent plots on which harvest occurred at some point during a 5–7 year period between plot measurements. This relative rarity of harvest activity leads to far greater uncertainty in estimates: 90% confidence intervals for annual harvest removals are ± 29.3 and ± 16.9% for the 80 and 160 km radius areas, respectively.

DISCUSSION

Precision Targets

If confidence intervals for needed estimates (**Table 1**) are considered low or inadequate, it is reasonable to ask what levels of precision for specific estimates are needed? Is there a threshold at which a confidence interval would be deemed "acceptable"? Unfortunately, it is extremely difficult in most cases to specify a target confidence interval that is needed. Resource information from inventory data is just one factor among many that affect private sector decisions. Managers frequently face decisions involving financial variables such as taxes and interest rates, market variables such as anticipated demand and supply, international and regional competition for raw materials, and restrictions on other key resources. Few, if any, of these factors carry estimates of uncertainty, so it is unlikely managers could specify a threshold for needed precision of resource data. Furthermore, decisions frequently must be made within a limited time; there is little room to wait for "better information" before deciding. Often, the best that can be done is to put estimated levels of uncertainty into context with other decision variables and consider risks related to uncertainty.

A possible exception is when statements about rates of change must be made with some level of confidence. For example, it may be important for certification or reporting purposes to be able to state that forest area or forest carbon stocks are not decreasing within an operating area. This implies that measured change in forest area or carbon stocks can be shown to be increasing or stable (changing at a rate not significantly different from zero), with a specified confidence. In such cases, though, the precision

TABLE 1 | Example of variables of interest, sample size, and 90% confidence interval for FIA plot-based estimates for 80 and 160 km radii around an arbitrary location.

Variables of interest	80 km radius			160 km radius		
	Estimate	Sample size (plots)	Conf. interval (%)	Estimate	Sample size (plots)	Conf. interval (%)
Area of private timberland (ha)	1,347,355	629	6.2	4,943,146	2,317	2.5
Area in pine forest types (ha)	657,714	346	9.0	2,236,049	1,188	2.5
Pine forest area by age class (ha)[a]						
0–5 years	48,962	25	34.5	208,381	112	16.5
6–10 years	66,524	33	29.7	210,478	118	16.4
11–15 years	53,468	29	33.2	241,033	133	15.3
Softwood sawtimber volume (k m^3)	58,983	283	12.7	171,898	914	6.6
Annual softwood sawtimber growth (k m^3/yr)	4,517	302	12.3	13,764	987	6.4
Annual softwood sawtimber harvest (k m^3/yr)	1,481	87	29.3	4,441	273	16.9

All estimates pertain to private timberland only.
[a]*Only three age classes listed for brevity.*

target will depend on the underlying rate of resource change, so it will differ in different geographic areas.

Privacy Concerns

The US private forestry sector has valued privacy of information and regulations that protect confidentiality of business information. Therefore, some might expect concerns within the private sector about public access to fine-grain resource information developed through SAE. However, this issue was never raised in an FIA User Group meeting on SAE attended by 74 participants, including representatives of private timberland owners (Guldin, 2020b). Furthermore, several of the authors of this manuscript work for private timberland owners or associations, and none have expressed such concerns. With widespread public access to high-resolution imagery, public records of land ownership, and numerous interpretive maps such as forest biomass distribution, it is clear that the risk of loss of privacy is outweighed by the gains possible through broader adoption of SAE applied to forest resource data.

Moving Forward

The FIA program is the logical place for expanding research and development of SAE applied to forest resource data. However, meeting the private sector resource information needs will require partnership, concerted effort, and increased investment.

The private sector is already partnering with the FIA program in a variety of ways: cooperative funding of research into SAE, cooperating with FIA by allowing access to private lands for field inventory, and responding to Timber Products Output and National Woodland Owner Survey questionnaires. The private sector also has been a strong supporter of the FIA program by advocating for increased funding for the program. The FIA program, in turn, has proven responsive to needs expressed by the private sector through Blue Ribbon Panels on FIA and annual FIA user group meetings.

There are several opportunities that could benefit not only the private forest sector, but many public and academic users of FIA data as well. These may be categorized as improving the quality and consistency of data, making data more accessible to users, and making better use of technology and ancillary datasets.

Improving the quality and consistency of data:

- Organizations within the private sector may be able to help FIA validate research products using proprietary resource data, such as assessment of accuracy of SAE products using fine-scale company inventory data;
- FIA and other units within USFS Research can focus on closing substantial data and knowledge gaps related to belowground and dead wood carbon dynamics, forest management effects on carbon cycles, soil carbon sequestration in forest ecosystems, and storage of carbon in harvested wood products;
- FIA program leadership can work to improve the nationwide consistency of field protocols and analytical approaches that will ensure credible, consistent, and timely data on forest carbon stocks and fluxes.

Making data more accessible to users:

- The FIA program could benefit from external expertise to improve the design and delivery of online tools for analysis and dissemination of data to significantly enhance accessibility and usability;
- Early engagement with the user community in the design of tools for delivery of SAE estimates would help ensure that resulting products meet user needs.

Making better use of technology and ancillary datasets:

- Because land use change is such a critical factor in forest carbon fluxes, FIA can build on successes using remote sensing-based programs such as the Landscape Change Monitoring System (LCMS) and Image-based Change Estimation (ICE) to arrive at a reliable, annually updated source of information on nationwide forest area change;
- FIA scientists can move from a research to an implementation phase for SAE applications to national forest inventory data, which will require deciding on specific ancillary datasets (such as remote sensing products) and methods that show the greatest promise.

Improvement of resource data delivery with SAE builds on the foundation of the FIA phase 2 field inventory. None of the advances recommended here should come at the expense of the core program of field inventory. This means that advances are dependent on additional funding. At every opportunity, private sector organizations should advocate for full and increased funding for the FIA program to meet these objectives.

SUMMARY

The US forest sector is highly dependent on the contributions made by private timberland owners and manufacturers. Private sector stakeholders are facing increasing demands for resource information, which could be met in part by data from the national forest inventory. Improved precision in estimates from FIA can be achieved using SAE approaches and leveraging additional datasets. Additional federal investment in research, aided by partnership efforts with the private sector, states, and educational institutions will be necessary to meet private sector information needs.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

SP developed the initial draft. JB, MC, SF, DK, JR, and ES participated equally in review and revision. All authors contributed to the article and approved the submitted version.

REFERENCES

Abt, K. L., Abt, R. C., Galik, C. S., and Skog, K. E. (2014). *Effects of Policies on Pellet Production and Forests in the U.S. South: A Technical Document Supporting the Forest Service 2010 RPA Assessment*. General Technical Report SRS-202. Asheville, NC: United States Department of Agriculture Forest Service, Southern Research Station.

Abt, R. C., Galik, C. S., and Henderson, J. D. (2010). *The Near-Term Market and Greenhouse Gas Implications of Forest Biomass Utilization in the Southeastern United States*. Durham, NC: Duke University, Nicholas Institute for Environmental Policy Solutions and Center on Global Change.

Birdsey, R., Duffy, P., Smyth, C., Kurz, W. A., Dugan, A. J., and Houghton, R. (2018). Climate, economic, and environmental impacts of producing wood for bioenergy. *Environ. Res. Lett.* 13:050201.

Brooks, E. B., Coulston, J. W., Wynne, R. H., and Thomas, V. A. (2016). Improving the precision of dynamic forest parameter estimates using Landsat. *Rem. Sens. Environ.* 179, 162–169.

Costanza, J. K., Abt, R. C., McKerrow, A. J., and Collazo, J. A. (2016). Bioenergy production and forest landscape change in the southeastern U.S. *Glob. Change Biol. Bioenergy* 9, 924-939.

Coulston, J. W., Green, P. C., Radtke, P. J., Prisley, S. P., Brooks, E. B., Thomas, V. A., et al. (2021). Enhancing the precision of broad-scale forestland removals estimates with small area estimation techniques. *For. Int. J. For. Res.* 94, 427–441.

Dale, V. H., Parish, E., Kline, K. L., and Tobin, E. (2017). How is wood-based pellet production affecting forest conditions in the southeastern United States? *For. Ecol. Manage.* 396, 143-149. doi: 10.1016/j.foreco.2017.03.022

Davis, R. J., Janet, O. L., Robert, K. E., Warren, C. B., Matthew, G. J., Zhiqiang, Y., et al. (2015). *Northwest Forest Plan–the first 20 years (1994-2013): Status and Trends of Late-Successional and Old-Growth Forests*. Gen. Tech. Rep. PNW-GTR-911. Portland, OR: U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station, 112.

Green, P. C., Burkhart, H. E., Coulston, J. W., Radtke, P. J., and Thomas, V. A. (2020). Auxiliary information resolution effects on small area estimation in plantation forest inventory. *For. Int. J. For. Res.* 93, 685–693. doi: 10.1093/forestry/cpaa012

Guldin, R.W., ed. (2020a). *Program Summary: Forest Inventory and Analysis Program National Users Group Meeting. August 17-21, 2020*. Washington, DC: Society of American Foresters, 27.

Guldin, R.W., ed. (2020b). *Program Summary: Small Area Estimation Focus Sessions at Forest Inventory and Analysis Program National Users Group Meeting. September 30 – October 2, 2020*. Washington, DC: Society of American Foresters, 21.

Jefferies, H. M., and Leslie, T. (2017). *Historical Perspective on the Relationship Between Demand and Forest Productivity in the US South*. Charlotte, NC: Forest2Market, Inc, 104.

Kim, T. J., Wear, D. N., Coulston, J., and Li, R. (2018). Forest land use responses to wood product markets. *For. Policy Econ.* 93, 45-52. doi: 10.1016/j.forpol.2018.05.012

Lubowski, R. N., Plantinga, A. J., and Stavins, R. N. (2008). What drives land-use change in the United States? A national analysis of landowner decisions. *Land Econ.* 84, 529-550. doi: 10.3368/le.84.4.529

Nagle, N. N., Schroeder, T. A., and Rose, B. (2019). A regularized raking estimator for small-area mapping from forest inventory surveys. *Forests* 10, 1–17. doi: 10.3390/F10111045

Nordman, C., White, R., Wilson, R., Ware, C., Rideout, C., Pyne, M., et al. (2016). *Rapid Assessment Metrics to Enhance Wildlife Habitat and Biodiversity Within Southern Open Pine Ecosystems,* version 1.0. March 31, 2016. Durham, NC: U.S. Fish and Wildlife Service and NatureServe, for the Gulf Coastal Plains and Ozarks Landscape Conservation Cooperative.

Oswalt, S. N., Brad, S. W., Patrick, M. D., and Scott, P. A. (2019). *Forest Resources of the United States, 2017: a technical document supporting the Forest Service 2020 RPA Assessment*. Gen. Tech. Rep. WO-97. Washington, DC: U.S. Department of Agriculture, Forest Service, Washington Office, 223. doi: 10.2737/WO-GTR-97

USDA Forest Service (2021). *Forest Inventory EVALIDator Web-Application* Version 1.8.0.01. *Forest Inventory and Analysis Program*. St. Paul, MN: U.S. Department of Agriculture, Forest Service, Northern Research Station.

Conflict of Interest: SP was employed by the National Council for Air and Stream Improvement, Inc.; JB was employed by the American Forest & Paper Association; MC was employed by Forest Investment Associates; SF was employed by the National Alliance of Forest Owners; DK was employed by Molpus Woodlands Group, LLC; JR was employed by International Paper; and ES was a principal at Three Trees Consulting.

frontiers in Forests and Global Change

ORIGINAL RESEARCH
published: 15 November 2021
doi: 10.3389/ffgc.2021.761509

Small Area Estimation of Postfire Tree Density Using Continuous Forest Inventory Data

George C. Gaines III and David L. R. Affleck*

Department of Forest Management, University of Montana, Missoula, MT, United States

Wildfire activity in the western United States is expanding and many western forests are struggling to regenerate postfire. Accurate estimates of forest regeneration following wildfire are critical for postfire forest management planning and monitoring forest dynamics. National or regional forest inventory programs can provide vegetation data for direct spatiotemporal domain estimation of postfire tree density, but samples within domains of administrative utility may be small (or empty). Indirect domain expansion estimators, which borrow extra-domain sample data to increase precision of domain estimates, offer a possible alternative. This research evaluates domain sample sizes and direct estimates in domains spanning large geographic extents and ranging from 1 to 10 years in temporal scope. In aggregate, domain sample sizes prove too small and standard errors of direct estimates too high. We subsequently compare two indirect estimators—one generated by averaging over observations that are proximate in space, the other by averaging over observations that are proximate in time—on the basis of estimated standard error. We also present a new estimator of the mean squared error (MSE) of indirect domain estimators which accounts for covariance between direct and indirect domain estimates. Borrowing sample data from within the geographic extents of our domains, but from an expanded set of measurement years, proves to be the superior strategy for augmenting domain sample sizes to reduce domain standard errors in this application. However, MSE estimates prove too frequently negative and highly variable for operational utility in this context, even when averaged over multiple proximate domains.

Keywords: forest inventory, wildland fire, forest regeneration, bias estimation, forest inventory and analysis, monitoring trends in burn severity

OPEN ACCESS

Edited by:
Gretchen Moisen, Rocky Mountain Research Station, United States Forest Service (USDA), United States

Reviewed by:
Francisco Mauro, Oregon State University, United States
John Coulston, Southern Research Station, Forest Service, United States Forest Service (USDA), United States

****Correspondence:***
George C. Gaines III
george.gaines@umconnect.umt.edu

Specialty section:
This article was submitted to Forest Management, a section of the journal Frontiers in Forests and Global Change

Received: *19 August 2021*
Accepted: *13 October 2021*
Published: *15 November 2021*

Citation:
Gaines GC III and Affleck DLR (2021) Small Area Estimation of Postfire Tree Density Using Continuous Forest Inventory Data. Front. For. Glob. Change 4:761509. doi: 10.3389/ffgc.2021.761509

INTRODUCTION

Wildfires in the western USA are increasing in frequency, size and severity and many western forests are struggling to regenerate postfire (Stevens-Rumann et al., 2017). Hot, dry climatic conditions fueled a 2020 wildfire season of unprecedented dimension, with over 1.5 million ha burned in California alone (Higuera and Abatzoglou, 2021). In the USA, securing regeneration of burned forest areas can be important for compliance with federal legislation, atmospheric CO_2 sequestration, and perpetuation of forest products availability. Accurate estimates of residual tree cover and new seedling recruitment following wildfire are thus critical for understanding postfire forest dynamics and maximizing the impact of limited resources for postfire management activities like tree planting.

Many countries now monitor forest resources using a network of sample locations distributed at a nationwide or broad, regional level. In the USA, the sample plot network administered by the United States Forest Service (USFS) Forest Inventory and Analysis (FIA) program provides nationwide ground observations of vegetation attributes, including tree regeneration (Bechtold and Patterson, 2005). In addition, the Monitoring Trends in Burn Severity (MTBS) program provides fire perimeters and burn severities for all large wildfire events from 1984 to 2018 (Eidenshink et al., 2007). Together these two sources of information provide a means of estimating postfire forest characteristics. Yet the spatial and temporal resolution of the FIA sample relative to the spatiotemporal frequency of wildland fires is expected to render traditional estimation techniques unreliable for domains defined by individual fire perimeters or collections thereof. Here we investigate the viability of direct domain estimators of postfire tree density across various domain resolutions, and compare them to indirect estimators. Indirect estimators, a class of small area estimation (SAE) techniques, borrow sample observations from proximate domains to increase effective sample sizes for domains requiring more precise estimation, or small areas.

Applications of SAE techniques have proliferated in the forestry literature, reflective of the need in public and private sectors alike to increase the spatiotemporal resolution of estimates of forest attributes without major investments in additional data collection. Examples include approaches to estimation proceeding from design-based (e.g., Breidenbach and Astrup, 2012; Hill et al., 2018), model-based (e.g., Breidenbach and Astrup, 2012; McRoberts, 2012; Coulston et al., 2021) and hybrid (e.g., Magnussen et al., 2014b) inferential paradigms. For detailed contrasts of differing inferential frameworks see Gregoire (1998) and Ståhl et al. (2016).

Breidenbach and Astrup (2012) evaluated alternative approaches to domain estimation of above-ground forest biomass using Norwegian National Forest Inventory (NFI) data. Domains consisted of 14 municipalities forming an exhaustive partition of the study area. They compared domain sample means with synthetic and generalized regression (GREG) domain estimators, as well as with empirical best linear unbiased predictor (EBLUP) composite domain estimators. The GREG and EBLUP estimators both leveraged remotely-sensed canopy height data. Both also resulted in more accurate estimates than domain sample averages, as indicated by smaller estimated variances in the case of GREG and by smaller estimated mean squared errors (MSEs) in the case of EBLUP. Notably, the MSEs estimated for the domain EBLUPs were of an unconditional nature (Datta et al., 2011), being averaged over an explicit (Gaussian) model of domain heterogeneity.

McRoberts (2012) presented model-based nearest neighbor (NN) techniques for SAE, illustrated using USFS FIA data and Landsat-derived attributes. The NN domain estimates of volume ($Mg\ ha^{-1}$) proposed were synthetic in the sense that observations from the complete population were eligible to serve as neighbors for any given location within a domain. Evaluation of the relationship between observations and NN predictions of volume for lack of fit was suggested in the model-based context as a means of assessing the presence of domain-level estimation bias.

Adopting a design-based approach, Hill et al. (2018) evaluated (two-stage) domain-level GREG estimators for application with German NFI data. They related timber volume at a plot level to LiDAR-derived variables and a species classification map, and compared a weighted domain sample average with approximately design-unbiased GREG estimators incorporating domain-specific intercepts. The GREG estimators reduced estimated variances of domain sample means by 43% in larger geographic domains and 23% in smaller domains.

Coulston et al. (2021) compared post-stratified estimators with model-based estimators of domain-level forest removals across the southeastern US. They related FIA ground data to Landsat-based tree cover loss and sawmill survey data at the area level. The model-based SAE strategies they developed for domain-level forest removals provided smaller estimated (unconditional) MSEs relative to the estimated variances of post-stratified domain estimators, at both county and multi-county domain resolutions.

More generally, several themes can be identified from the literature on small area estimation in forest inventory. The first is that most applications consider only domains with fixed spatial delineation, defined for example by administrative/political boundaries (e.g., Breidenbach and Astrup, 2012; McRoberts, 2012; Hill et al., 2018). As described below, domains of interest that arise from forest disturbances have spatial and temporal bounds that are important—both in defining the parameters of interest and in determining what measurements are within or outside the domains. Second, there are often asymmetries in how data from spatially-proximate vs. temporally-proximate (but potentially spatially-coincident) domains are used in domain estimation. Numerous studies evaluated the use of data drawn only from spatially-proximate domains, perhaps because data from other years were unavailable. Other studies have drawn on inventory data from multiple years, but only while correspondingly broadening the definition of the target estimand from an attribute specific to a point in time to one averaged over a (multi-year) period. In each of the four studies cited above, measurements spanning a multi-year period are used in a "temporally indifferent" sense (Bechtold and Patterson, 2005) to form domain estimates that explicitly or implicitly encompass a multi-year extent. A third theme is that most previous applications (including all of those cited above) leverage relationships between ground observations of the target attribute and one or more auxiliary variables. That is, they evaluate gains in accuracy that might be achieved through the incorporation of extra-domain data and of statistical relationships between the attribute of interest and other data products.

An additional theme that emerges from the SAE literature is that estimation of the bias or MSE of indirect domain estimators is challenging. Under a design-based approach, the ability to estimate the bias of domain estimators is hindered by the same constraint that motivates indirect estimation in the first place, namely a lack of sufficient data. As such, both Hill et al. (2018) and Breidenbach and Astrup (2012) eschew synthetic regression domain estimation; they focus instead on

approximately unbiased regression estimators, precluding the need for bias or MSE estimation. Under a model-based approach, domain differences are incorporated into an explicit probabilistic model. This elevates a need for model validation strategies (see e.g., McRoberts, 2012), but also allows for derivation of MSEs and of estimators thereof. Datta et al. (2011) describe alternative MSEs that can be pursued under the model-based approach, but suggest that the conditional MSE of interest under the design-based approach is least readily estimated. In line with this, many SAE studies adopting EBLUP domain estimation have employed estimators of unconditional MSEs characterizing average performance over a distribution of possible domain effects (e.g., Breidenbach and Astrup, 2012; Coulston et al., 2021).

In this study, we investigate two methods for augmenting domain samples for indirect estimation of tree attributes in disturbed areas: one method borrows explicitly in space; the other in time. Also, inasmuch as indirect estimation necessarily introduces bias, with different strategies incorporating different sources of bias, we also evaluate estimators of the MSE and bias of the indirect domain estimators. Overall, our objectives are to i) advance a framework for defining wildfire-origin domains and estimating forest attributes at specified postfire intervals; ii) evaluate the feasibility of direct estimation of postfire tree regeneration across varying domain extents using FIA data; iii) determine the advantages and limitations associated with alternative strategies for incorporating FIA data from proximate spatial and temporal domains into indirect estimators; and, iv) investigate the utility of estimators of the MSE and bias of indirect estimators. Our approach is developed in the next section and then exemplified using fire perimeters from the western coterminous US and field data from the FIA program.

FRAMEWORK FOR DOMAIN DELINEATION AND ESTIMATION

We assume that interest lies in resources distributed across a population defined over both spatial $\mathcal{X}$ and temporal $\mathcal{T}$ extents. Also, we assume the resources are monitored via a probability-based sample design that selects a finite number of locations in space $\mathbf{x} \in \mathcal{X}$ and designates each for measurement at a time $t \in \mathcal{T}$. Our research then focuses on the estimation of resource parameters over (small) domains of the population.

Domains of interest in forest management may persist over time and be defined only by their spatial extents. For example, a domain may be defined administratively, such as the State of Wyoming or the Shoshone National Forest (WY). However, the domains of interest here are those that are created by a disturbance event (or complex of disturbance events) and that thus also have a temporal component. For example, a domain may consist of all lands burned by a particular wildfire event in 1990. Such a domain has a spatial extent defined by the 1990 burn perimeter and a temporal extent running from 1990 forward. Generalizing, a domain may instead consist of all lands within the Shoshone National Forest burned by wildfires in 1990, or all lands within Wyoming that burned in wildfires between 1990 and 1999. In the latter example, the spatial extents of the constituent fires may overlap (e.g., a subset of the area burned in 1990 could burn again in 1999). This could be handled in various ways depending on research or management interests, but in the subsequent we attribute any such overlap to the most recent burn and effectively clip it from the spatial extent of the earlier burn. Thus, a domain defined by a 1990 wildfire event may have a spatial extent that is constant from 1990 to 1998, and a reduced spatial extent from 1999 onwards owing to a partial reburn event in 1999. Notably, such domains are not likely to form an exhaustive partition of the population in any given year, and in any given year not all existing disturbance-generated domains will have persisted over the same time interval.

Owing both to the potential for the spatial extent of a domain to change over time and to the fact that the resources of interest are dynamic, domain properties are referenced by a domain index d ($d = 1, 2, \ldots$) and a temporal index l ($l = 0, 1, 2, \ldots$). The latter index measures time (numbers of years) elapsed since the defining disturbance event(s). Define $\mathcal{A}(d, l) \subseteq \mathcal{A}(d, 0)$ as the spatial extent of domain d at l years post-disturbance, corresponding to the original spatial extent of the disturbance less any regions subsequently disturbed within l years. Interest centers on the spatial density of a resource attribute y at given points in time, or

$$\lambda(d, l) = \frac{1}{|\mathcal{A}(d, l)|} \int_{\mathcal{A}(d,l)} y(\mathbf{x}, l) \mathrm{d}\mathbf{x} \tag{1}$$

where $|\mathcal{A}(d, l)|$ is the area of the domain d after a lag of l years, and $y(\mathbf{x}, l)$ is the resource value at spatial coordinate $\mathbf{x}$ as it exists l years after the domain-defining disturbance event. That is, we adopt a continuous population perspective (see e.g., Grafström et al., 2017) and focus on $y(\mathbf{x}, l)$ as defining the number of live trees per unit area at location $\mathbf{x}$ in year l, which in practice necessitates counting live trees over a fixed support area, such as a circular plot. Thus, for example, if the domain d corresponds to a particular 1990 wildfire, then interest may lie in the number of live trees per unit area that are standing in 1995 [$= \lambda(d, 5)$] or that are standing in 2000 [$= \lambda(d, 10)$]. In either case, it must be recognized that the spatial extent of the domain could be different in 2000, 1995, and 1990 owing to subsequent disturbance [i.e., $\mathcal{A}(d, 10) \subseteq \mathcal{A}(d, 5) \subseteq \mathcal{A}(d, 0)$]. Moreover, if the domain d corresponds to all lands burned by wildfires in Wyoming between 1990 and 1999, then $\lambda(d, 5)$ still defines the density of the resource 5 years post-disturbance. In this case, the parameter integrates regeneration density in 1995 over areas burned in 1990 as well as regeneration density in 1999 over areas that burned in 1994. That is, as defined here, the lag index l does not denote a period of time initiating at the oldest (or most recent) disturbance event subsumed within a domain of interest, but rather a fixed interval allowed to elapse over all disturbances within a domain.

In the small area estimation terminology of Rao and Molina (2015), a direct estimator of $\lambda(d, l)$ would draw only on the set $s(d, l)$ of sample observations $y_k = y(\mathbf{x}_k, l_k)$ located in domain d and observed after a lag of l years. The size of $s(d, l)$, denoted $n(d, l)$, is assumed to be a random variable because $\mathcal{A}(d, l)$ is not an independently sampled stratum of the population. One direct estimator applicable to equal-probability inventory designs is the

domain sample mean

$$\bar{y}(d,l) = \frac{1}{n(d,l)} \sum_{k \in s(d,l)} y_k \quad (2)$$

Under simple random sampling (SRS), $\bar{y}(d,l)$ is a conditionally unbiased estimator of $\lambda(d,l)$ provided $n(d,l) > 0$ (see **Appendix A**). However, this result does not hold for other equal probability sampling designs; bias of the domain sample mean accrues from variability in $n(d,l)$ and generally decreases only as $n(d,l)$ increases (Särndal et al., 2003, pp. 176–177).

For small domains, the domain sample mean (Equation 2) and other direct estimators are expected to have high variance owing to small and variable sample sizes. Thus, we also consider indirect estimators of $\lambda(d,l)$ that utilize data from an augmented sample set $\tilde{s}(d,l) \supseteq s(d,l)$ of observations coming from within and beyond the spatiotemporal domain $\mathcal{A}(d,l)$. For example, $\tilde{s}(d,l)$ may supplement $s(d,l)$ with observations drawn from another domain d' but made at the same time-since-disturbance [i.e., by borrowing data from $\mathcal{A}(d',l)$], or from the same domain but at different lags-since-disturbance l' [from $\mathcal{A}(d,l')$], or from a combination of these extensions. Denoting the size of the augmented sample by $\tilde{n}(d,l)$, a simple indirect domain estimator that might be applied under equal probability sampling is the augmented sample mean

$$\hat{\bar{y}}(d,l) = \frac{1}{\tilde{n}(d,l)} \sum_{k \in \tilde{s}(d,l)} y_k \quad (3)$$

Implicit in the use of this estimator is the assumption that the spatial density of the attribute of interest differs little within the domain vs. over the region from which data are borrowed. Generally, this assumption becomes less tenable as that extra-domain region is expanded in space or time but, regardless, Equation (3) is a biased estimator of $\lambda(d,l)$, even under SRS. Its bias under SRS will depend on the relative size of the region from which data are borrowed and on the extent to which the spatial density of y differs over that region relative to $\lambda(d,l)$ (see **Appendix B**). At the same time, the variance of an indirect estimator such as $\hat{\bar{y}}(d,l)$ is expected to be lower than that of $\bar{y}(d,l)$ owing to the augmented sample size.

Inasmuch as indirect domain estimators are generally biased, MSE should provide a more informative statistical summary than variance. Unfortunately, useful analytical expressions (or estimators) of the MSE of an indirect domain estimator are difficult to obtain. Building on Rao and Molina (2015, p. 43) and suppressing the domain and lag indices (d and l) for brevity, the MSE of an indirect estimator $\hat{\lambda}_i$ can be written as

$$\mathrm{MSE}\left[\hat{\lambda}_i\right] = \mathrm{E}\left[\hat{\lambda}_i - \lambda\right]^2 = \mathrm{E}\left[\hat{\lambda}_i - \hat{\lambda}_u\right]^2 - \mathrm{V}\left[\hat{\lambda}_u\right] + 2\,\mathrm{C}\left[\hat{\lambda}_i, \hat{\lambda}_u\right] \quad (4)$$

where $\hat{\lambda}_u$ is an unbiased estimator of the domain parameter, $\mathrm{V}\left[\hat{\lambda}_u\right]$ is its variance, and $\mathrm{C}\left[\hat{\lambda}_i, \hat{\lambda}_u\right]$ is its covariance with $\hat{\lambda}_i$. Going further, from the basic definition of MSE (i.e., variance plus squared bias), Equation (4) can be re-arranged to provide an expression for the squared bias K of an indirect domain estimator, viz.

$$\mathrm{K}\left[\hat{\lambda}\right] = \left(\mathrm{E}\left[\hat{\lambda}_i\right] - \lambda\right)^2 = \mathrm{E}\left[\hat{\lambda}_i - \hat{\lambda}_u\right]^2 - \mathrm{V}\left[\hat{\lambda}_u\right] - \mathrm{V}\left[\hat{\lambda}_i\right] + 2\,\mathrm{C}\left[\hat{\lambda}_i, \hat{\lambda}_u\right] \quad (5)$$

The above expressions for MSE and squared bias have been used to derive several estimators for indirect domain estimation (e.g., Gonzalez and Waksberg, 1973, pp. 6; Marker, 1995, pp. 67–71; Rao and Molina, 2015, pp. 44–45). Commonly however, the covariance term in expresssions (Equations 4, 5) has been ignored. Dropping the covariance term may be justified in applications where the indirect estimator draws on a considerably larger sample than the direct estimator—for then the two estimators can be expected to have low correlation. Yet in settings where the domain sample size is an appreciable component of the data used by the indirect estimator, the covariance term cannot be expected to be negligible. Instead, it is expected to be positive, tending to $\mathrm{V}\left[\hat{\lambda}_u\right]$ as $\tilde{n}$ approaches the domain sample size and tending to 0 only as $\tilde{n}$ becomes much larger than n.

METHODS

Forest Inventory Data

This study utilizes data from the USFS annualized Phase 2 (P2) plot network spanning all lands (forested and non-forested, all ownerships) in the 11 contiguous states of the western USA (**Figure 1**). The plot network is based on an equal-intensity sampling design that began with tessellation of the landbase into approximately 2,400 ha hexagons, followed by the selection of 1 plot location per hexagon (Bechtold and Patterson, 2005). Implementation of the annualized FIA program in the western states involves the remeasurement of one of 10 interpenetrating panels of plots each year, yielding a nominal sampling intensity of approx. one plot measurement per 24,000 ha per year.

At the time this research was undertaken, FIA plot data were publicly available for measurements taken in 2018 back through the year of initial implementation (which varied by state). All FIA plots are assessed for condition (e.g., forested vs. non-forested) and the attributes measured on forested conditions permit computation of live tree density over a range of age and size classes (seedlings, saplings, and larger trees) for each of the 4 subplots comprising an FIA plot (see Bechtold and Patterson 2005). Such data also exist for some regionally intensified FIA plot grids and regional post-fire FIA plot remeasurement designs, but these were not included in the analysis as they have variable spatial and temporal measurement intensities. For various reasons (e.g., presence of seasonal water, hazardous field conditions), vegetation data are not available for every subplot; such subplots were necessarily excluded from the analysis dataset and not utilized in averaging tree densities to a plot level. However, NFI subplot condition mapping procedures permitted the incorporation of data from subplots that were only partially measurable. The numbers of measurements of (at least)

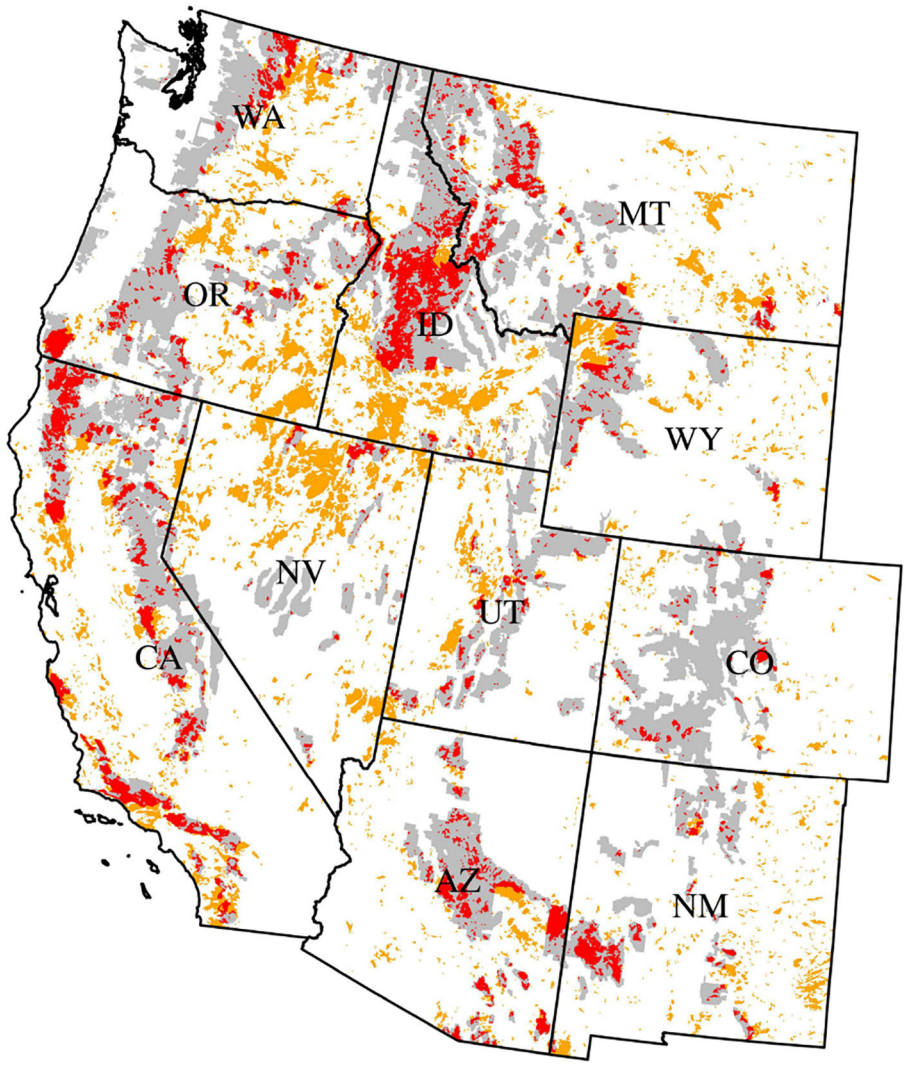


FIGURE 1 | Study area spanning 11 states in the western USA. Areas spanned by MTBS burn polygons 1984–2018 are shown in red where they overlap USFS National Forest System (NFS) lands and in orange otherwise; unburned NFS lands are shown in gray.

partially-forested FIA plots by state and year are summarized in **Supplementary Figure 1**.

Domains

This research centers on estimating post-fire tree density in forested areas of the western US experiencing wildland fire events. Thus, domains were defined using 1984–2018 burn perimeters obtained from the MTBS program (Eidenshink et al., 2007), which maps all wildland fires ≥404 ha in the western US. Also, in order to facilitate a focus on forested areas, where maintaining or re-establishing forest cover is a management objective, domains were restricted to the intersection of MTBS burn perimeters and USFS National Forest System (NFS) lands (excluding grasslands or other non-forest land designations, see **Figure 1**). Burned areas outside of these lands and burned areas on non-forested lands more generally were not considered parts of the domains of interest. Finally, US state boundaries were overlaid over the burn perimeters. This was done in part to account for differential sampling intensities over time across states (see **Supplementary Figure 1**), as well as to allow for estimation at a state-level resolution.

Given these constraints, the most finely resolved domains considered here consist of a complex of NFS lands within an individual western US state that are spanned by MTBS perimeters of a specific burn year. But also considered are aggregates of these domains taken over different time spans. Thus, allowing for a 2-year burn period, a domain can consist of NFS lands within a western US state spanned by MTBS perimeters from a given biennium; a 10-year burn period allows for domains consisting of NFS lands within an individual state spanned by MTBS perimeters from a given decade. In these instances, only

non-overlapping time spans are considered; that is, in the 10-year case, we consider decadal domain burn periods ranging from 1990–1999 to 2000–2009.

The parameters of interest for each domain are taken as the mean tree densities at specified post-burn intervals. That is, as $\lambda(d, l)$ defined by Equation (1) with $y(\mathbf{x}, l)$ denoting tree density (numbers of trees per ha) at location $\mathbf{x}$ at a temporal lag of l years post-fire. Below we consider only lags of 2 years or greater owing to the fact that data on first year germinants are not collected on FIA plots.

Domain Estimation

The FIA sample is distributed across all lands, while the domains of interest here span only burned, forested lands under NFS ownership. Therefore, the full FIA sample was first subset to plots falling within MTBS perimeters and within the states shown in **Figure 1**. The geographic coordinates of these plots and the standard cluster configuration were then used to determine the burned status of subplots. Data for subplots outside the bounds of any MTBS perimeters dating back to 1984 were dropped; measurement data for all remaining subplots were tied to the most recent MTBS burn and an associated fire-measurement lag computed. FIA condition mapping procedures then enabled elimination of subplots or portions of subplots classified to non-forest conditions (e.g., rangeland condition). Notably, subsetting to forested subplot data did not eliminate any plot measurements from our analysis set, it changed only the subplot support of those FIA plot measurements spanning multiple conditions. Finally, subplot measurement data were associated with the domains described above or with none of those domains (e.g., because a subplot was not located on NFS lands); data from the same domain and having the same lag were then aggregated to the plot level. All geospatial operations were undertaken in R (R Core Team, 2021).

Sample sizes available for direct estimation $n(d, l)$ were determined from the number of FIA plot measurements falling within the domain d of interest and at the lag l of interest. In this, and in the subsequent estimators, plot-level records were treated the same irrespective of potentially differing numbers of subplots (e.g., because some subplots were outside the domain of interest or measured at a different lag). Plot-level compilations of trees per ha (all size classes, all species) were used for direct estimation of $\lambda(d, l)$ via estimator (Equation 2). This domain sample mean $\bar{y}(d, l)$ is not an unbiased (or conditionally unbiased) estimator of $\lambda(d, l)$ under the FIA design. For instance, consider a domain known to completely encompass 10 hexagons comprising a 10-year remeasurement panel (see Bechtold and Patterson 2005) as well as portions of neighboring hexagons. Then, conditioning on a domain sample size of 1 also means conditioning on the location of the singular plot measurement coming from within one of the 10 completely spanned hexagons (and not from any of the incompletely spanned hexagons), meaning that the domain sample mean cannot be conditionally unbiased in general. Still, as with other ratio-type estimators the bias will decrease with increasing sample size. As an aside, we note that the domain sample mean (Equation 2) differs from the ratio estimation approach adopted by the FIA program. In this application, a y_k in Equation (2) is the number of trees on burned, partially-forested subplots of an FIA plot divided by the aggregate area of those burned, partially-forested subplots. The strategy advanced by Bechtold and Patterson (2005) is to instead (i) average the numbers of trees on burned forest land per unit plot area; (ii) average the areas of burned forest land per unit plot area; (iii) form a ratio of these two averages. Williams (2001) describes some of the key differences between these ratio estimators.

The standard error of $\bar{y}(d, l)$ was estimated using

$$\mathsf{SE}\left[\bar{y}(d, l)\right] = \frac{\hat{\sigma}_y(d, l)}{\sqrt{n(d, l)}} \tag{6}$$

where

$$\hat{\sigma}_y^2(d, l) = \frac{1}{n(d, l) - 1} \sum_{k \in s(d,l)} \left[y_k - \bar{y}(d, l)\right]^2 \tag{7}$$

is an estimator of the within-domain sample variance.

Direct estimates of $\lambda(d, l)$ [where $n(d, l) \geq 1$] and associated standard errors [where $n(d, l) \geq 2$] were computed for all domains and all feasible lags. Tree density could not be estimated for all possible lags on all domains, however, because the annualized FIA program began only in 2001 (and only then for some states; see **Supplementary Figure 1**). Also, at the time of this research measurements were available only through 2018. Thus, for example, mean tree densities at the 5- and 10-year lags are estimable for the domain defined as NFS lands burned in California in 2000, but only at the 5-year lag for the domain defined as NFS lands burned in California in 2010. Variability in the numbers of domains for which tree density can be estimated by burn period and lag is summarized in **Supplementary Figure 2** for domains of various burn interval lengths. It's also worth noting that for multi-year domains, lag remains constant and the applicable plot measurement years vary over the MTBS perimeters. For example in the case of the domain d comprised of NFS lands in ID burned in 2006 or 2007, the direct estimator of $\lambda(d, l)$ for $l = 10$ uses only 2016 plot measurements for areas burned in 2006 fires and only 2017 measurements over the 2007 burns. This preserves the length of time elapsed between burns and corresponding plot observations.

Every direct domain and lag estimate was compared against two types of indirect estimates. The first type augmented the domain sample size by borrowing data from a broader spatial extent. Specifically, for a given domain d and lag l, all FIA plot measurements with the same lag l and falling within MTBS perimeters intersected by a spatial buffer extended around domain $\mathcal{A}(d, l)$ were drawn into $\tilde{s}(d, l)$. Buffer distances ranging from 25 to 250 km were implemented in R (R Core Team, 2021). Note that under this procedure the augmented sample $\tilde{s}(d, l)$ can include plot data that are not within any domain of interest (i.e., in MTBS perimeters but outside the administrative state and/or NFS delineation), but only if the plot measurements were taken l years post-fire.

The second type of indirect estimate was obtained from augmented samples formed by borrowing data from a broader temporal extent. For a given domain d and lag l of interest, any

FIA plot measurements made within the spatial extent $\mathcal{A}(d,l)$ and at $l \pm \delta$ years post-fire were drawn into $\tilde{s}(d,l)$. With this approach the augmented sample $\tilde{s}(d,l)$ can include only plot data from the same domain of interest [same MTBS perimeter(s)] but measurements taken prior or subsequently to the lag of interest. Thus, for a domain d defined as all 2010 MTBS burns on NFS lands in Montana and a lag of interest of $l = 5$ years, $s(d,l)$ would consist only of plot data measured in 2015 within $\mathcal{A}(d,5)$; but $\tilde{s}(d,l)$ would consist also of plot data measured in 2015$\pm\delta$ within the spatial extent $\mathcal{A}(d,5)$ (provided $2015 - \delta \geq 2012$ because only $l \geq 2$ year data are considered, and provided $2015 + \delta \leq 2018$ because FIA measurements from 2019 or later were not available). Lag buffers δ ranging from 1 to 7 years were evaluated.

With both sample augmentation strategies, the indirect estimator (Equation 3) was applied. Furthermore, estimates of standard error were obtained similarly to direct estimation as

$$\widehat{\mathrm{SE}}\left[\hat{\bar{y}}(d,l)\right] = \frac{\hat{\tilde{\sigma}}_y(d,l)}{\sqrt{\tilde{n}(d,l)}} \tag{8}$$

where

$$\hat{\tilde{\sigma}}^2_y(d,l) = \frac{1}{\tilde{n}(d,l)-1} \sum_{k \in \tilde{s}(d,l)} \left[y_k - \hat{\bar{y}}(d,l)\right]^2 \tag{9}$$

Thus, the estimated standard error for the indirect estimator is a function of both a potentially larger sample size and of the variability within that larger sample. Relative standard error was obtained by relating $\widehat{\mathrm{SE}}$ to estimated tree density.

MSE Estimation

Equation (8) can be used to estimate the precision of the indirect estimator, but makes no attempt to account for its inherent bias; a useful indicator of this estimator's accuracy would account for both. Equation (4) led to two estimators of the MSE of the indirect domain estimators (see **Appendix B** for details). The simplest, again suppressing the domain and lag indices d and l for brevity, takes the form

$$\widehat{\mathrm{MSE}}\left[\hat{\bar{y}}\right]_1 = \left(\hat{\bar{y}} - \bar{y}\right)^2 - \frac{\hat{\sigma}^2_y}{n} \tag{10}$$

This MSE estimator is based on an approximation suggested by Rao and Molina (2015, p. 44) but employs $\bar{y}$ in place of a strictly unbiased domain estimator. It does not attempt to account for the covariance between the direct and indirect domain estimators. As such, it can be expected to be more appropriate in contexts where augmented sample sizes are consistently much larger than domain sample sizes. The other estimator evaluated here takes the form

$$\widehat{\mathrm{MSE}}\left[\hat{\bar{y}}\right]_2 = \left(\hat{\bar{y}} - \bar{y}\right)^2 - \frac{\hat{\sigma}^2_y}{n}\left[1 - 2\,\frac{n}{\tilde{n}}\right] \tag{11}$$

In this estimator the factor $\left[1 - 2\,\frac{n}{\tilde{n}}\right]$ results from the inclusion of an estimated covariance between $\hat{\bar{y}}$ and $\bar{y}$. We note that $\widehat{\mathrm{MSE}}\left[\hat{\bar{y}}\right]_2 \geq \widehat{\mathrm{MSE}}\left[\hat{\bar{y}}\right]_1$ (though neither estimator is guaranteed to be positive) and expect that $\widehat{\mathrm{MSE}}\left[\hat{\bar{y}}\right]_2$ will be more accurate when augmented samples are not substantially larger than the corresponding domain samples. Finally, as suggested by Marker (1995) we computed estimated squared bias as of the indirect domain estimator as

$$\widehat{\mathrm{K}}\left[\hat{\bar{y}}\right] = \widehat{\mathrm{MSE}}\left[\hat{\bar{y}}\right]_q - \hat{\tilde{\sigma}}^2_y(d,l) \tag{12}$$

for $q = 1, 2$.

Estimates of MSE and squared bias were computed for each domain and lag individually, and also averaged over groups of proximate domains. The latter strategy was suggested by Gonzalez and Waksberg (1973) to reduce instability in MSE or squared bias estimates. In this study, we averaged MSE and squared bias estimates over all domains within the same state and having the same burn period length (e.g., any biennium for domains with 2-year burn periods), as well as over all estimation lags.

RESULTS

Over the 11 states of the western USA shown in **Figure 1**, there were 4,778 FIA P2 plot locations falling at least partially within MTBS burn perimeters dating from 1984 to 2018. These locations provided 5,946 plot measurements from burned areas with measurement lags ranging from 2 to 35 years post-fire.

The distribution of domain sample sizes for domains of different temporal extents is shown in **Figure 2**. For domains spanning only a single burn year (e.g., all NFS lands burned in OR in 2000), sample sizes are almost so small as to prohibit direct estimation: in only 6% of cases (domains × lags) did the sample size exceed 5 observations. Even for domains spanning 4-years (e.g., all NFS lands burned in OR between 2000 and 2003), the median sample size is only 2 observations. This rises to 7 in the case of decadal domains (e.g., all NFS lands burned in OR between 2000–2009), the lowest temporal resolution considered to be of administrative utility.

Though small, and inherently random, these domain sample sizes are governed in part by the FIA sampling intensity of approximately 1 plot measurement per 2,400 ha per decade. That nominal intensity is shown as the dotted line in **Figure 2**; realized intensities are captured by the solid lines that consistently fall short of the approximately 1:24,000 nominal rate.

Figure 2 also highlights two distinct domains for reference. Shown in red is the domain comprising OR NFS lands burned between 2000 and 2009 (lags 2–9 year). At lag 2 year, this domain spanned an areal extent of 605,690 ha, but with partial reburns the extent dropped to 550,806 ha at lag 9 year. Sample sizes ranged from 15 (lag 6 year) to 28 (lag 4 year), reflecting the generally high inter-annual variation in domain sample sizes. In blue is the domain comprising ID NFS lands burned between 1990 and 1999 (lags 14–19 years). This domain spanned an area of 332,272 ha at lag 14 year and captured sample sizes ranging from 6 to 15 observations.

Restricting attention to decadal domains, the relationship between area and estimated standard error of the domain sample

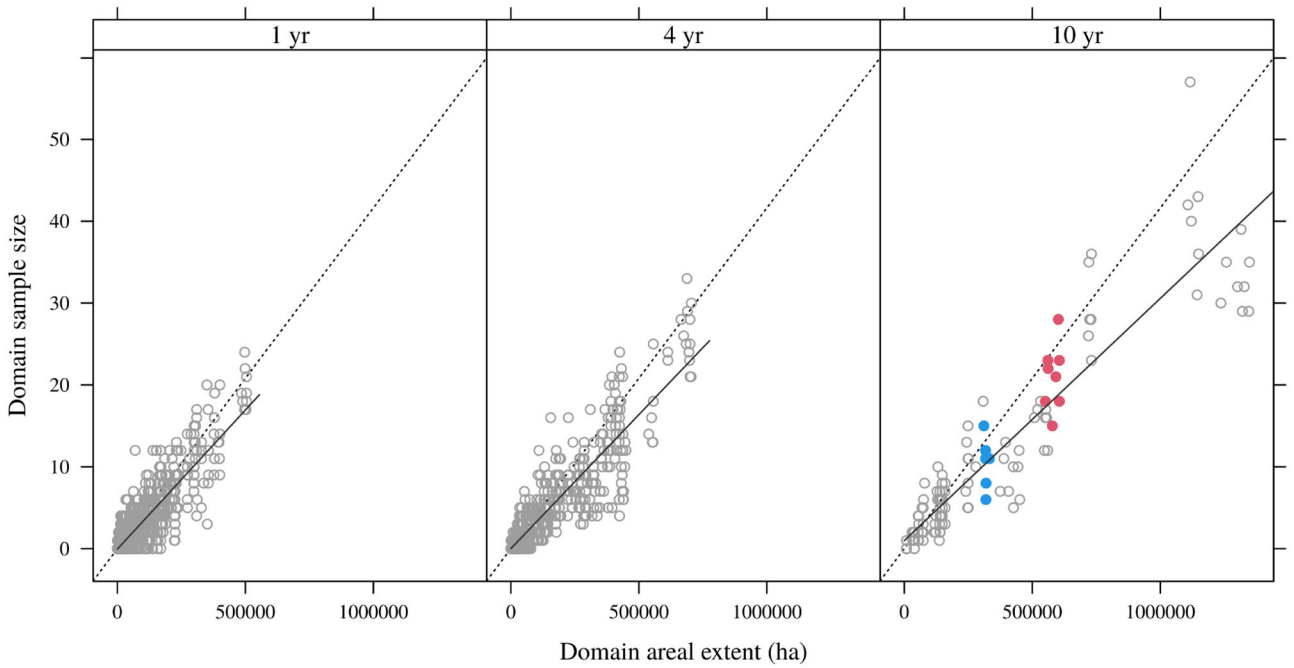


FIGURE 2 | Domain sample sizes and areas for annual, quadrennial, and decadal domains, showing data for all lags. Dotted line is the nominal FIA sampling intensity; solid lines are linear regressions for particular lag buffers. Red symbols denote various lags for the OR 2000–2009 domain; blue symbols denote estimable lags for the ID 1990–1999 domain.

means is shown in **Figure 3**. Domains with larger areal extents generally had larger sizes (see **Figure 2**) and smaller standard errors (**Figure 3**, left panel), though there is substantial variation around the latter trend. Moreover, standard errors could not even be computed for 15% of cases owing to domain sample sizes less than 2; over the remaining cases the median relative standard error was 47%. **Figure 3** also shows the relationship between estimated standard errors (where these could be computed) and domain sample means. On the natural logarithm scale, there is a strong linear association between the domain sample mean and its estimate standard error.

Borrowing data from an extended spatial extent generally augments the sample sizes available for indirect domain estimation (**Figure 4**). The dotted lines in **Figure 4** correspond to the same nominal sampling intensity as in **Figure 4**, while the solid lines now show the realized augmented sampling intensities. As expected, the larger the spatial buffer and the larger the initial domain extent, the greater the increase in sample size. However, the spatial buffering operation yields erratic results at the domain level. For the domain spanning OR NFS lands burned between 2000 and 2009 (red symbols), spatial buffering greatly and consistently increases the sample sizes available for estimation. Yet the effect is much less pronounced for the domain spanning ID NFS lands burned between 1990 and 1999.

The distribution of estimated standard errors for indirect estimates borrowing proximate spatial data, relative to those for direct estimates, is shown in **Figure 5** for 10-year domains. Although the relative standard errors of indirect estimates are larger than those for the corresponding direct estimates in some cases (even with 200 km buffers), spatially augmented samples tend to reduce relative standard errors. The extent of the shift in the distribution of standard errors is a function of the magnitude of the spatial buffer, as expected. However, the magnitude of the shift is not pronounced and the median relative standard error using a 200 km buffer is still 38%. In addition, even at a 200 km buffer, 5% of cases (10-year domains × estimable lags) have augmented sample sizes less than 2 and thus do not permit estimation of standard errors.

Relative to spatial buffering, borrowing data from an expanded temporal extent augments domain sample sizes at a consistent rate (**Figure 6**). The dashed lines in **Figure 6** represent the nominal sampling intensity of a domain augmented according to the expanded temporal range of measurements. Specifically, one would expect approximately 1 FIA plot measurement at a given lag l within a domain of 24,000 ha; by extension, in allowing for plot measurement lags of $l \pm \delta$ one would expect to collect $1 + 2\delta$ plot measurements for a domain of that size. Mean augmented sample sizes (solid lines in **Figure 6**) fall short of the expected augmented sample sizes, but the sample augmentation effect is more consistent across domains than with spatial buffering. That is, with an expanded temporal extent there is less variability in the proportionate increases in sample sizes across domains, as indicated for the highlighted OR and ID domains.

Corresponding to the more consistent sample augmentation of temporal buffering, the impacts on the distribution of

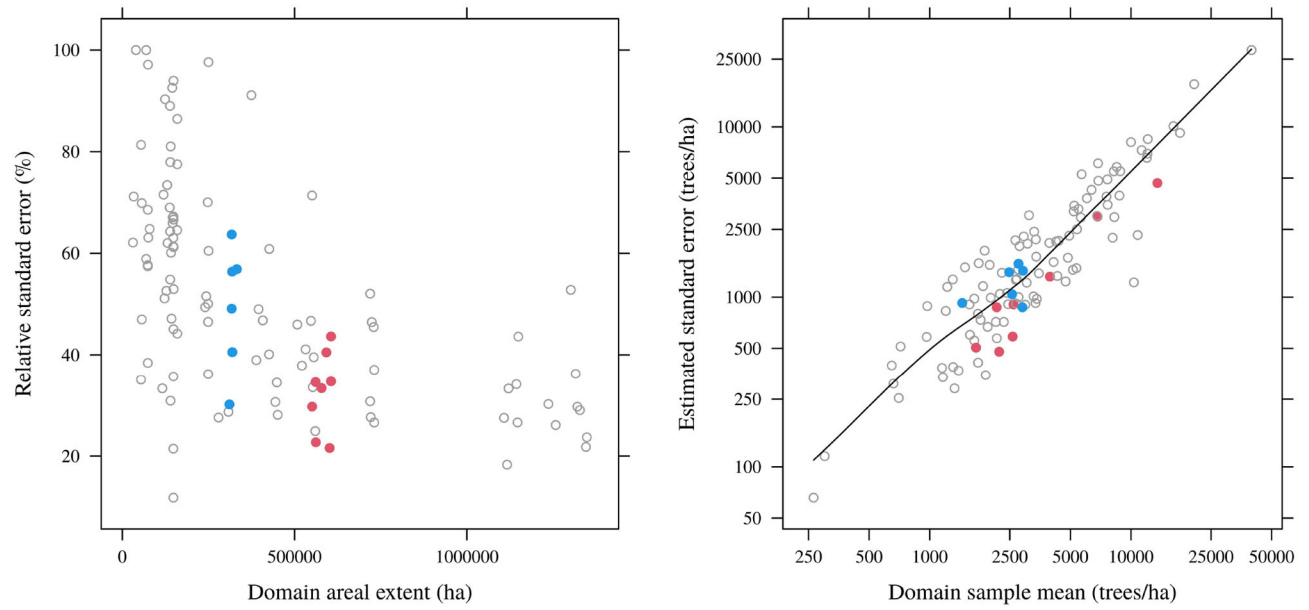


FIGURE 3 | Relationships between estimated relative standard errors of domain sample means and domain areas (left) and between estimated standard errors and domain sample means (right; log scale). Only results for 10-year domains (any lags) and domain sample sizes above 2 are shown.

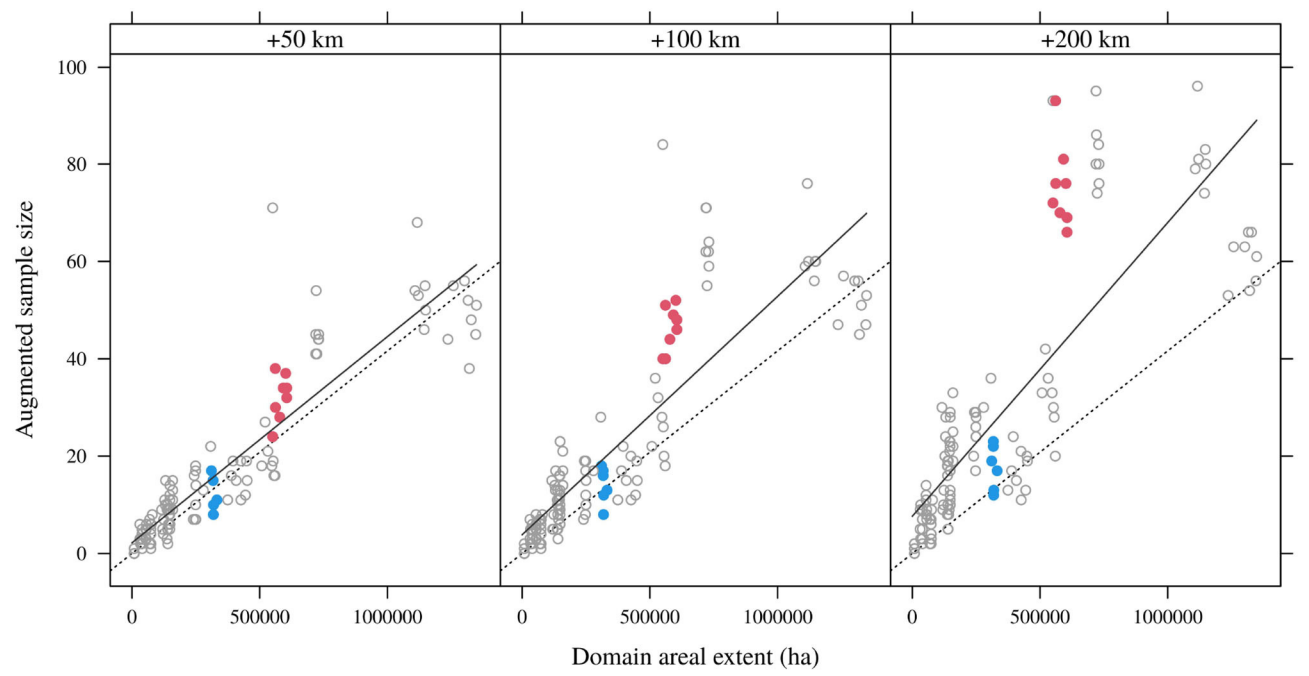


FIGURE 4 | Augmented sample sizes and domain areas for 10-year domains (all lags) and different spatial buffers (50, 100, and 200 km). Dotted line is the nominal FIA sampling intensity; solid lines are linear regressions for particular spatial buffers.

estimated standard errors of indirect estimates were larger and more consistent (**Figure 7**). Comparison to **Figure 5** also shows that relative standard errors of indirect estimates under $l \pm \delta$ borrowing are generally lower than under space borrowing. At the least intensive lag-borrowing level ($\delta = 1$ yr), they exceed the corresponding standard errors of the direct estimator much less frequently than under space borrowing, at even the largest buffer distance (200 km). Also, unlike under space borrowing (**Figure 5**). **Figure 7** shows substantial reductions in relative standard errors of both domains represented by red and blue

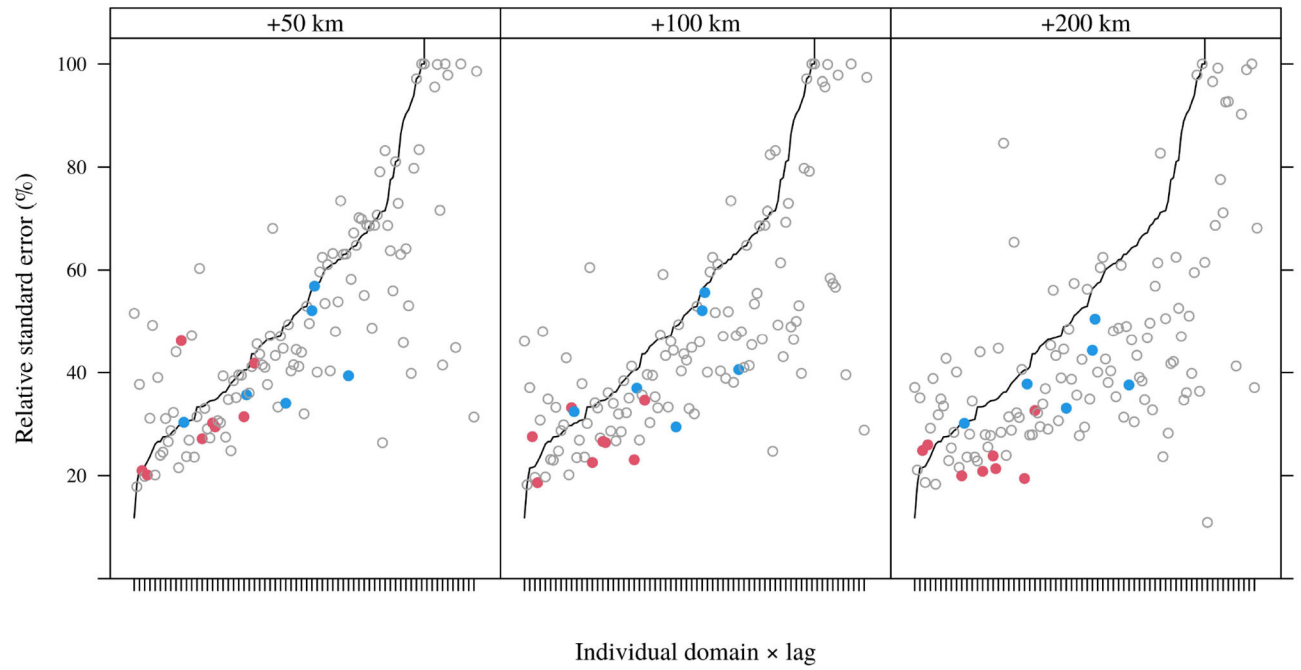


FIGURE 5 | Estimated relative standard errors of augmented sample means for 10-year domains (all lags) under for different spatial buffers; domains and lags are ordered according to relative standard errors of domain means (as represented by the black curve). Horizontal axis labels are individual domain and lag identities and have been suppressed for clarity.

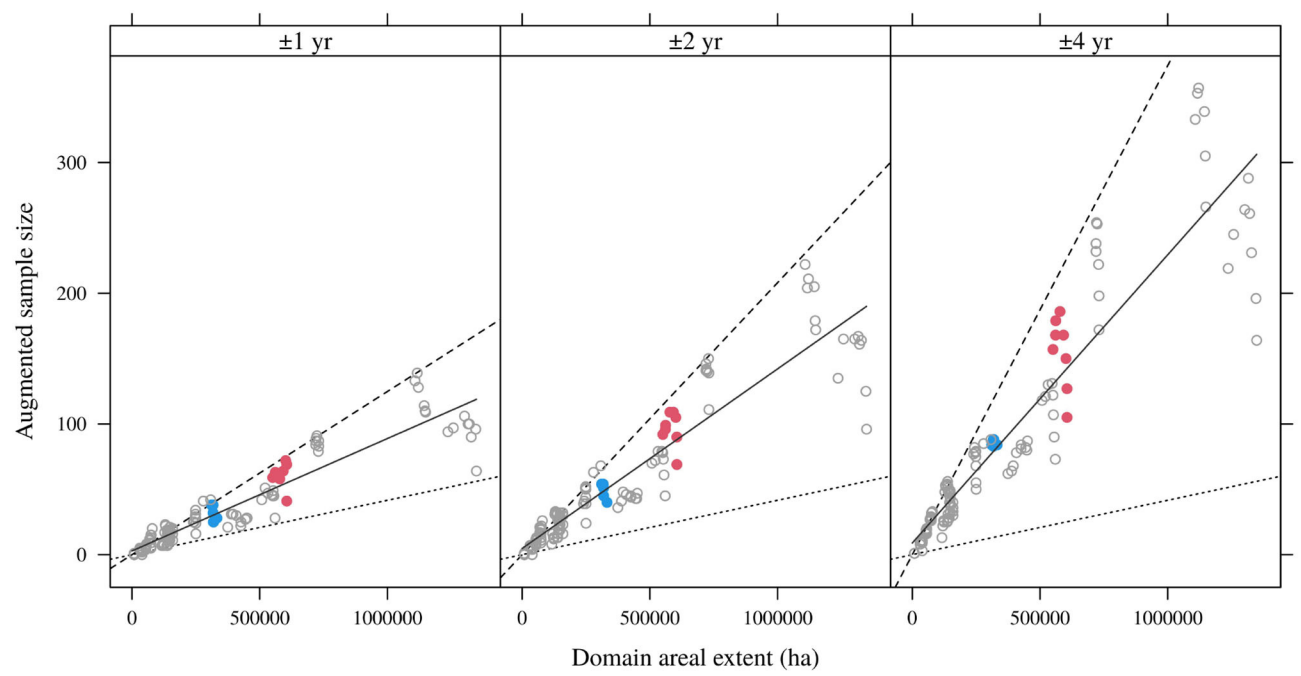


FIGURE 6 | Augmented sample sizes and domain areas for 10-year domains (all lags) and different lag buffers (δ =1, 2, 4 year). Dotted line is the nominal FIA sampling intensity $r \approx 1:24,000$; dashed lines are augmented intensities $(2\delta + 1)r$; solid lines are linear regressions for particular burn intervals.

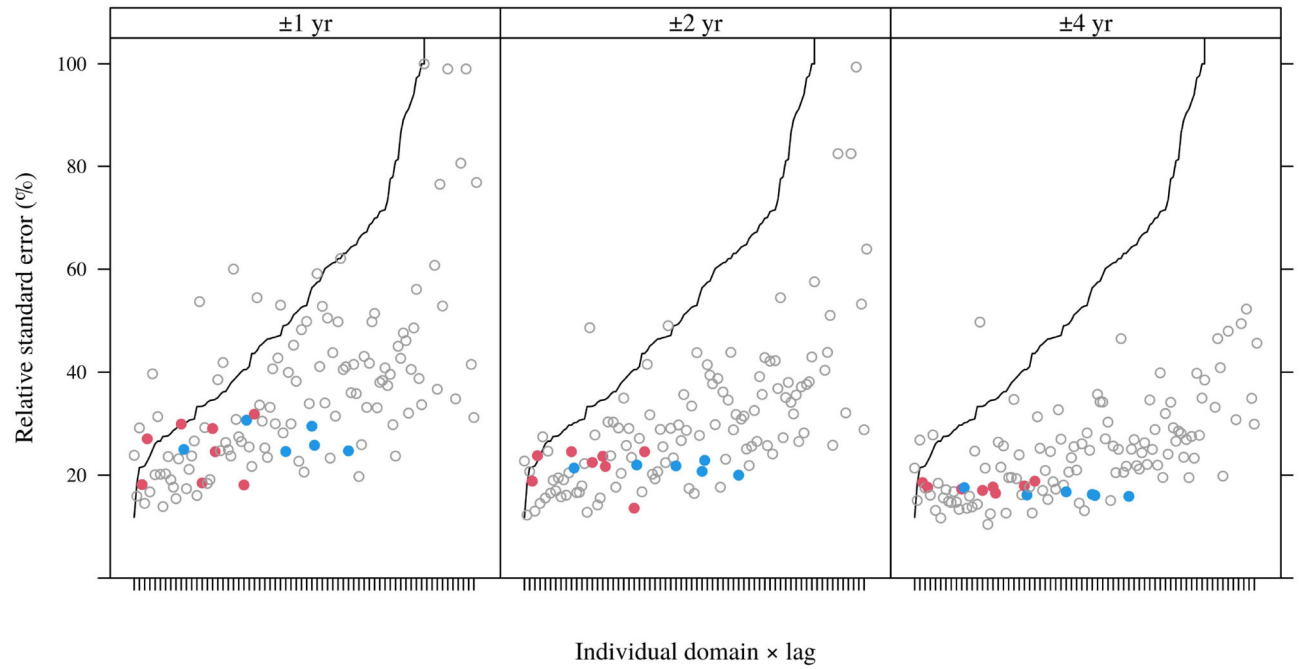


FIGURE 7 | Estimated relative standard errors of augmented sample means for 10-year domains (all lags) under for different temporal buffers; domains and lags are ordered according to relative standard errors of domain means (represented by the black curve). Horizontal axis labels are individual domain and lag identities and have been suppressed for clarity.

points, which consistently decline with increasing δ until they are approximately equal across lags for both domains at $\delta = 4$ yr.

Turning to MSE and bias estimation, for the decadal domains considered above MSE of the indirect estimators couldn't be estimated in 14% of cases (19 of 132 domains × lag combinations) regardless of temporal or spatial buffers employed. This was a result of domain sample sizes less than 2, which precluded estimation of $\hat{\sigma}^2(d,l)$ and thus of $\widehat{\mathrm{MSE}}\left[\hat{\bar{y}}\right]_1$ or $\widehat{\mathrm{MSE}}\left[\hat{\bar{y}}\right]_2$. Even setting aside such cases, both MSE estimators frequently produced negative estimates when applied at the domain level. For example, for the indirect estimates employing data with a lag buffer of $\delta = 1$ year, $\widehat{\mathrm{MSE}}\left[\hat{\bar{y}}\right]_2$ was negative in 41% of cases (domains × lags) while $\widehat{\mathrm{MSE}}\left[\hat{\bar{y}}\right]_1$ was negative in 71% of cases. As δ increased, the frequency of negative $\widehat{\mathrm{MSE}}\left[\hat{\bar{y}}\right]_1$ declined (though never fell below 50%), but the frequencies of negative $\widehat{\mathrm{MSE}}\left[\hat{\bar{y}}\right]_2$ increased to converge with those of $\widehat{\mathrm{MSE}}\left[\hat{\bar{y}}\right]_1$. **Figure 8** shows estimated relative MSE (%) for the indirect domain estimator with $\delta = 2$ year plotted against domain area (ha), computed individually for each 10 year domain × lag combination, using Equation (11). While variability declined with domain area, it is clear that both MSE estimators are too variable across domains and within domains across lags to be of operational utility at the domain level.

Furthermore, both MSE estimators were still negative when averaged over proximate domains. Specifically, across different temporal buffers δ, $\widehat{\mathrm{MSE}}\left[\hat{\bar{y}}\right]_2$ yielded negative estimates for 20–40% of groups and $\widehat{\mathrm{MSE}}\left[\hat{\bar{y}}\right]_1$ for 50–90% of groups. Squared bias

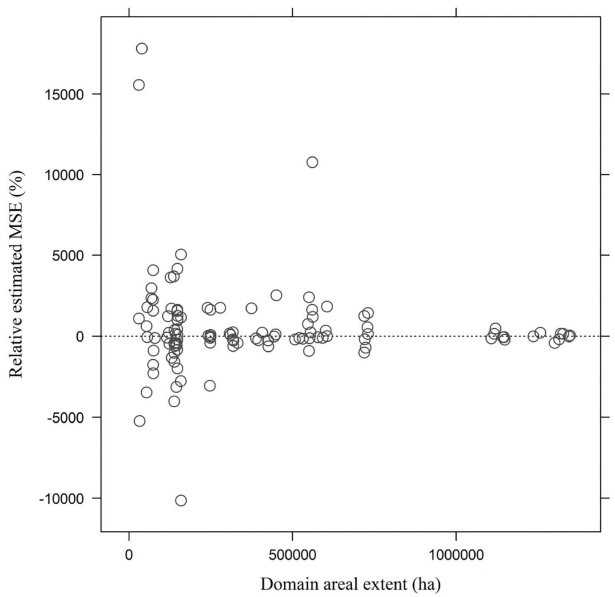


FIGURE 8 | Estimated relative MSE (%) for the indirect domain estimator with $\delta = 2$ plotted against domain area (ha), computed individually for each 10 year domain × lag combination using Equation (11).

as estimated by Equation (12), which subtracts the variance of the indirect estimator from a corresponding MSE estimate, was necessarily negative even more often than either MSE estimator taken alone.

DISCUSSION

The framework for indirect domain estimation we propose could be generalized to any probability sample of a target forest attribute (e.g., mean forest biomass density, total merchantable timber volume) distributed across spatiotemporal domains. Domains may span any time periods (for which requisite inventory data are available) and be comprised of contiguous or disjoint spatial polygons. It's worth remarking on the inherently complex nature of spatiotemporal domains comprised of burn perimeters intersecting a specific ownership category. Polygons are disjoint, often intersecting (reburns), and irregularly distributed in time and space according to neighborhood fire legacies.

When we expand domain delineations in space or time, the number of FIA P2 plot measurements will increase at a pace just below the nominal rate of approximately 1 measurement per 24,000 additional hectare-years (**Figure 2**). As we expand domains, however, they gradually lose administrative utility. For example, estimates of post-fire regeneration in areas burned over a reasonably narrow burn period length but extending over a vast geographic region (e.g., multiple states), or alternatively over a reasonably small geographic area but extending between 1984 and 2004 (20-year burn year window), would provide information of little utility to managers trying to optimize limited post-fire management resources for maximal regeneration impact.

Domain samples fluctuate around their anticipated sizes (given the nominal FIA sampling intensity and domain areas) owing in part to how the stratified random spatial distribution of plots intersects historic burn patterns. However, that the relationship between realized domain sample sizes and areas consistently falls short of its expectation must be due in large part to the fact that tree data are available only for FIA plots that are classified as partially forested. It may also be due in part to a tendency to fall short of annual plot remeasurement targets (see e.g., Roesch 2018). It is important to note that the consistent 1 observation per additional 24,000 ha^{-1} yr^{-1} burned area sample augmentation rate can only be expected to reliably emerge in years following the implementation of FIA's annualized inventory measurement protocols. This wasn't until 2001 at the earliest, 2011 in Wyoming, and with irregularities due to inconsistencies in funding in the interim (**Supplementary Figure 1**).

Our analysis of domain and augmented sample sizes and associated standard errors showed 10-year state-level domains to be the smallest spatiotemporal domains of administrative or management utility feasible for estimation of post-fire forest density using the domain estimators evaluated. As a general approach to estimation, direct FIA-based domain expansion estimation is unfeasible due to insufficiently small domain sample sizes and resultant high domain-level standard errors, even in 10-year domains. We note as well that we didn't account for the effects of retained plot size (e.g., only burned subplots) as implemented here on variance estimates. Hill et al. (2018) describe a methodology for incorporating differential plot sizes. Finally, though it wasn't an objective of this research, experimentation with other means of estimating the variance of domain estimates may be warranted (e.g., through the use of generalized variance functions as described by Wolter 2007, Chapter 7). A strong relationship between direct domain tree density estimates and their relative standard errors (**Figure 3**, right panel) was observed, as has been noted in other studies (e.g., Breidenbach et al. 2018).

Indirect estimators may offer an alternative. They are attractive in their potential to decrease domain-level standard errors. However, they rely on an implicit model that has the density of the attribute of interest changing slowly beyond the domain, at least relative to the variance of the attribute. We considered two strategies for borrowing data to augment domain samples for indirect estimation: borrowing in time (lag borrowing) and borrowing in space (space borrowing).

Under space borrowing, the rate of increase of the augmented sample size is dependent on the neighborhood fire legacy, the neighborhood land use patterns, and the overall sampling intensity. If many nearby forested hectares burned in the time range of interest, the augmented sample size will increase more quickly when data are drawn from a region only slightly expanded in space. Conversely, in areas with lower levels of nearby historic fire activity or lower levels of nearby forest land, one would need to expand further in space to obtain comparable increases in sample size. Yet borrowing extra-domain sample data in this way necessarily introduces bias to domain estimates. As plot observations from further away are selected for inclusion in the augmented domain sample, the biotic and abiotic environmental conditions of disparate forests may resemble those of the focal domain to a lesser extent. For example, borrowing in space can (and was observed to) draw on plot observations from distinct ecological conditions.

Another means of borrowing data that are proximate in space is to restrict the augmented sample to measurements (with appropriate postfire lag) from the same or similar ecological domains, regions or subsections (e.g., as delineated by Cleland et al. 1997). Nationwide availability of ecoregion designations of varied resolution would permit such restrictions. The capacity to augment the domain sample at a consistent rate, however, would still be governed by regional fire perimeter distributions in time and space. It would also then be impacted by regional landscape heterogeneity as exemplified by, for instance, varied ecoregions in mountainous terrain (with distinct forest and wildfire fuel type changes occurring over relatively short distances). An alternative approach wherein the augmented sample sizes could be fixed would be to borrow from the ideas underlying coarsened exact matching (see e.g., Van Deusen and Roesch 2013). That is, an initial spatial and/or ecological buffer could be evaluated and then, for domains still having an insufficient augmented sample size, the spatial buffer could be extended or the ecological classification coarsened. More generally, drawing data from

outside the domain of interest but from regions that share other characteristics (e.g., ecological subsection) has parallels in the ideas underlying post-stratification. Yet post-stratified estimation is most commonly implemented as a strictly direct estimation approach (e.g., Haakana et al. 2020) without drawing on data from strata that extend beyond the domain of interest.

The spatial buffering algorithm utilized here can also be related to nearest neighbor techniques (e.g., McRoberts 2012) in that both define a neighborhood from which to borrow data. However, nearest neighbor techniques select a fixed number of observations using a neighborhood defined in a broader auxiliary space (typically not restricted to or even dependent on geographic variables), while under space borrowing the number of observations selected into the augmented sample is a random function of neighborhood fire legacy. For domains where few additional observations are obtained under space borrowing even with large buffer distances, nearest neighbor techniques may need to reach very far in geographic space to obtain the specified fixed number of neighbors, with the potential to increase estimation bias.

Adoption of the temporal buffering algorithm allows for the use of plot observations from the same geographic extent as the domain of interest but measured at differing lengths of time-since-disturbance. Though data from additional plot locations falling within that extent are introduced, this method borrows only in time. Other SAE applications in forest inventory have pooled data from multiple years to generate domain estimates for domains with fixed spatial extents and (usually implicit) multi-year temporal extents (e.g., Breidenbach and Astrup, 2012; McRoberts, 2012; Hill et al., 2018). Here, we explicitly borrow sample observations with measurement years other than those denoted by the spatiotemporal domain parameters and target estimation lag. Spatiotemporal disturbance domains require a high degree of specificity in domain definition, and by extension in the definition of the temporal component of the target attribute. This specificity led to the determination that to include observations with measurement years other than those specified by the relevant disturbance lag is to operate in the realm of indirect estimation. Thus, the general estimation strategy employed by Breidenbach and Astrup (2012) that integrates data measured between 2005 and 2010 to estimate a periodic mean is distinct from our lag-borrowing indirect estimation strategy. With $\delta = 2$ year, the latter would draw on observations from 2005 to 2009 to indirectly estimate a target attribute in 2007, but on observations from 2006 to 2010 to indirectly estimate a target attribute in 2008.

Even in areas exhibiting highly unfavorable conditions for post-fire forest regeneration, some seeds will germinate, some seedlings will establish, and some patches of forest will eventually begin to regenerate over time. Thus, to include plots with measurement years earlier than specified by d and l in $\tilde{s}(d, l)$ is to include observations which may not capture the full extent of forest stand development in the focal domain, leading to negative bias. Conversely, to include plots with later measurement years is to include observations which may exaggerate the extent of true forest stand development in the focal domain, leading to positive bias.

As implemented in this study, lag borrowing augmented domain samples (**Figure 6**) and decreased relative domain standard errors (**Figure 7**) to a greater extent, and in a faster, more consistent manner, than space borrowing (**Figures 4**, **5**). The smaller increases in precision of the indirect estimator achieved via space borrowing relative to lag borrowing largely reflect instances where few additional plots were obtained by space borrowing (e.g., as in the case of the domain represented by blue points in **Figure 4**). This could also result from instances where plots from adjacent ecoregions with markedly different regeneration conditions were selected, adding to within-sample variability. Space borrowing has been shown to be effective in domains whose spatiotemporal neighborhoods yield more observations available for sample augmentation, for instance the estimation of an attribute over a single time period distributed across most or all adjacent forested area (e.g., Breidenbach and Astrup, 2012; Magnussen et al., 2014a).

As methods for borrowing increase in complexity, so do their associated sources, and likely magnitudes, of bias. For this reason we evaluated explicit space and lag borrowing only. Overall, lag borrowing exhibited greater magnitude and consistency of increases in both augmented samples and precision of estimates relative to space borrowing. These facts combine to suggest lag borrowing to be a superior borrowing strategy to space-borrowing for indirect expansion estimation of post-fire tree density in western US-wide spatiotemporal domains with respect to domain-level standard errors. That said, estimation of the bias of indirect domain estimators remains a challenge. An obstacle in formulating estimators of the MSE or squared bias of an indirect domain estimator from Equations (4) to (5) is the difficultly of reliably estimating the variance of an unbiased domain estimator—for the absence of a precise direct estimator is generally what motivates indirect estimation in the first place. The MSE estimators proposed by Rao and Molina (2015) and Gonzalez and Waksberg (1973), and squared bias estimator proposed by Marker (1995), can be negative and yield widely disparate MSE estimates for a single domain at lags separated by just one or several years, as occurred in our application. This resulted from subtraction of the unstable and often large estimated variance of the direct domain estimator.

The MSE estimator we proposed, which accounts for the covariance between direct and indirect domain estimates, constituted some improvement but was still unstable and frequently negative (**Figure 8**). It was also very high in some domains, and in fact is necessarily larger than the other estimator investigated. As suggested by Gonzalez and Waksberg (1973) and Rao and Molina (2015), we also averaged $\widehat{\mathrm{MSE}}\left[\hat{\bar{y}}\right]_1$ and $\widehat{\mathrm{MSE}}\left[\hat{\bar{y}}\right]_2$ over proximate domains to improve stability, but this yielded only marginal improvements.

Indirect FIA-based expansion estimation of post-fire tree regeneration in US state-level domains is probably most feasible in domains with burn year periods of 10 years, owing to small augmented sample sizes in many domains of shorter burn period lengths. By $\delta = 2$ yr, the vast majority of standard errors of indirect lag-borrowed estimates are substantially lower than their direct counterparts (**Figure 7**), suggesting $\delta = 2$ or 3 as a

potential starting point for operational domain estimation. This is with the understanding that we were unable to effectively characterize the bias of indirect estimates. Composite estimators (Rao and Molina, 2015) seek to balance the instability of an unbiased (or approximately unbiased) direct estimator with the bias of a more precise indirect estimator. Weights controlling the relative contributions of the component estimators are typically constructed using either domain sample sizes or their relative MSEs. Owing to our unreliable estimates of the MSE of the indirect estimator, we could not have constructed a composite estimator based on MSE. Though we could have devised weights using domain sample sizes, we did not expect the resultant composite estimates to be more precise than the indirect estimates based on lag borrowing alone, and in any case did not expect MSE estimation techniques to apply successfully to the composite estimator for the same reasons discussed above. These results point to the need for exploration of model-assisted or model-based SAE strategies that could draw on systematic associations (or effective post-stratifications) of postfire tree density as a function of auxiliary variables available across the population.

CONCLUSION

Direct FIA-based estimation of postfire tree density at particular times-since-disturbance is deemed unfeasible due to insufficiently small domain sample sizes. Indirect domain ratio estimators that borrow sample observations from outside a focal domain are alternatives to auxiliary-assisted methods and have the potential to consistently and rapidly augment decrease domain level standard errors. Borrowing in time proved to augment domain samples more consistently than borrowing in space. On the basis of relative standard errors alone, indirect estimation of postfire tree regeneration in 10-year state-level domains with $\delta = 2$ or 3 presents a promising alternative to direct estimation.

As indirect estimators necessarily add bias to domain estimates, reliable estimators of MSE are required. MSE estimators of indirect domain estimators have been proposed and evaluated in the literature, and we evaluate a new MSE estimator that accounts for the covariance between direct and synthetic domain estimates. However, none of the MSE estimators evaluated performed adequately.

Our results highlight the difficulties of estimating MSE and squared bias, and point to the need for further experimentation with methods for estimating MSE, including potentially modeling MSE using appropriate covariates. Alternatively, unbiased SAE techniques that preclude the need for bias estimation, and that leverage auxiliary data, warrant inquiry in this context.

DATA AVAILABILITY STATEMENT

True FIA plot center locations are confidential. Requests to access these datasets should be directed to https://www.fia.fs.fed.us/.

AUTHOR CONTRIBUTIONS

GG and DA made major contributions to research and analysis. Both authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The availability of geospatial data used in this publication was made possible, in part, by an agreement from the United States Department of Agriculture's Forest Service (USFS). This publication may not necessarily express the views or opinions of the USFS. Support for this research was also provided by the Inland Northwest Growth & Yield Cooperative and by the USFS Rocky Mountain Research Station (20-JV-11221636-110).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/ffgc.2021.761509/full#supplementary-material

REFERENCES

Baffetta, F., Fattorini, L., Franceschi, S., and Corona, P. (2009). Design-based approach to k-nearest neighbours technique for coupling field and remotely sensed data in forest surveys. *Remote Sens. Environ.* 113, 463–475. doi: 10.1016/j.rse.2008.06.014

Bechtold, W. A., and Patterson, P. L. (2005). *The Enhanced Forest Inventory and Analysis Program-National Sampling Design and Estimation Procedures*. General Technical Report GTR-SRS-80, USDA Forest Service, Southern Research Station.

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

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

Cleland, D. T., Avers, P. E., McNab, W. H., Jensen, M. E., Bailey, R. G., King, T., et al. (1997). National hierarchical framework of ecological units. *Ecosyst. Manage. Appl. Sustain. For. Wildlife Resour.* 20, 181–200.

Cordy, C. B. (1993). An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. *Statist. Probability Lett.* 18, 353–362.

Coulston, J. W., Green, P. C., Radtke, P. J., Prisley, S. P., Brooks, E. B., Thomas, V. A., et al. (2021). Enhancing the precision of broad-scale forestland removals estimates with small area estimation techniques. *Forestry* 94, 427–441. doi: 10.1093/forestry/cpaa045

Datta, G. S., Kubokawa, T., Molina, I., and Rao, J. N. K. (2011). Estimation of mean squared error of model-based small area estimators. *Test* 20, 367–388. doi: 10.1007/s11749-010-0206-2

Eidenshink, J., Schwind, B., Brewer, K., Zhu, Z., Quayle, B., and Howard, S. (2007). A project for monitoring trends in burn severity. *Fire Ecol.* 3, 3–21. doi: 10.4996/fireecology.0301003

Frank, B., and Monleon, V. J. (2021). Comparison of variance estimators for systematic environmental sample surveys: considerations for post-stratified estimation. *Forests* 12:772. doi: 10.3390/f12060772

Gonzalez, M. E., and Waksberg, J. (1973). "Estimation of the error of synthetic estimates," in *First Meeting of the International Association of Survey Statisticians* (Vienna), 18–25.

Grafström, A., Schnell, S., Saarela, S., Hubbell, S. P., and Condit, R. (2017). The continuous population approach to forest inventories and use of information in the design. *Environmetrics* 28:e2480. doi: 10.1002/env.2480

Gregoire, T. G. (1998). Design-based and model-based inference in survey sampling: appreciating the difference. *Can. J. For. Res.* 28, 1429–1447. doi: 10.1139/x98-166

Haakana, H., Heikkinen, J., Katila, M., and Kangas, A. (2020). Precision of exogenous post-stratification in small-area estimation based on a continuous national forest inventory. *Can. J. For. Res.* 50, 359–370. doi: 10.1139/cjfr-2019-0139

Higuera, P. E., and Abatzoglou, J. T. (2021). Record-setting climate enabled the extraordinary 2020 fire season in the western United States. *Glob. Chang Biol.* 27, 1–2. doi: 10.1111/gcb.15388

Hill, A., Mandallaz, D., and Langshausen, J. (2018). A double-sampling extension of the German national forest inventory for design-based small area estimation on forest district levels. *Remote Sens.* 10, 1052–1078. doi: 10.3390/rs10071052

Magnussen, S., Mandallaz, D., Breidenbach, J., Lanz, A., and Ginzler, C. (2014a). National forest inventories in the service of small area estimation of stem volume. *Can. J. For. Res.* 44, 1079–1090. doi: 10.1139/cjfr-2013-0448

Magnussen, S., Næsset, E., and Gobakken, T. (2014b). An estimator of variance for two-stage ratio regression estimators. *For. Sci.* 60, 663–676. doi: 10.5849/forsci.12-163

Marker, D. A. (1995). *Small Area Estimation: A Bayesian Perspective* (Ph.D. thesis). University of Michigan.

McRoberts, R. E. (2012). Estimating forest attribute parameters for small areas using nearest neighbors techniques. *For. Ecol. Manage.* 272, 3–12. doi: 10.1016/j.foreco.2011.06.039

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rao, J. N. K., and Molina, I. (2015). *Small Area Estimation, 2nd Edn*. Hoboken, NJ: Wiley.

Roesch, F. (2018). Composite estimators for growth derived from repeated plot measurements of positively-asymmetric interval lengths. *Forests* 9, 427. doi: 10.3390/f9070427

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model Assisted Survey Sampling*. New York, NY: Springer Science & Business Media.

Ståhl, G., Saarela, S., Schnell, S., Holm, S., Breidenbach, J., Healey, S. P., et al. (2016). Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation. *For. Ecosyst.* 3, 1–11. doi: 10.1186/s40663-016-0064-9

Stevens-Rumann, C. S., Kemp, K. B., Higuera, P. E., Harvey, B. J., Rother, M. T., Donato, D. C., et al. (2017). Evidence for declining forest resilience to wildfires under climate change. *Ecol. Lett.* 21, 243–252. doi: 10.1111/ele.12889

Van Deusen, P., and Roesch, F. (2013). Trends and projections from annual forest inventory plots and coarsened exact matching. *Math. Comput. For. Natural Resour. Sci.* 5, 126–134.

Williams, M. S. (2001). Comparison of estimation techniques for a forest inventory in which double sampling for stratification is used. *For. Sci.* 47, 563–576. doi: 10.1093/forestscience/47.4.563

Wolter, K. M. (2007). *Introduction to Variance Estimation, 2nd Edn*. Springer.

Conflict of Interest: The 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

APPENDIX

Appendix A: Mean & Variance of the Direct Estimator

Adopting the continuous population framework of Cordy (1993), we consider equal probability sampling designs that can be described by specifying a constant inclusion density function $\pi(\mathbf{x}, t) = \pi(t)$ for all possible measurements locations over a land surface in a given year t. For such designs, the direct domain estimator (2) can be formulated as a ratio of the two Horvitz-Thompson domain estimators

$$\hat{\tau}_\pi(d,l) = \sum_{k \in s(d,l)} \frac{y_k}{\pi(l)} = \frac{\sum_{k \in s(d,l)} y_k}{\pi(l)} \tag{A1a}$$

$$\hat{A}_\pi(d,l) = \sum_{k \in s(d,l)} \frac{1}{\pi(l)} = \frac{n(d,l)}{\pi(l)} \tag{A1b}$$

where $\pi(l)$ is the inclusion density function at l years following the defining disturbance event. Estimators (A1) are unbiased for the total of y over $\mathcal{A}(d,l)$ at year l from disturbance, and for the total area of $\mathcal{A}(d,l)$, respectively. Yet the nonlinear combination of these estimators is generally biased for $\lambda(d,l)$. The domain sample mean (2) can be described as "approximately unbiased" in the sense that its bias diminishes with increasing expected $n(d,l)$ (see Särndal et al., 2003, p. 185), though this is of limited utility in a small area estimation context where we anticipate small $n(d,l)$.

Cordy (1993) provides a number of general results concerning the bias and variance of estimators such as $\bar{y}(d,l)$. In particular, his results allow that if the conditional inclusion density function $\pi_n(\mathbf{x}, l)$ given $n(d,l)$ is positive for all measurement locations within $\mathcal{A}(d,l)$, then

$$\mathsf{E}\left[\, \bar{y}(d,l) \,\middle|\, n(d,l) \,\right] = \lambda(d,l) \tag{A2}$$

provided $n(d,l) > 0$. This conditional unbiasedness result holds for SRS because under that design

$$\pi_n(\mathbf{x}, l) = \frac{n(d,l)}{|\mathcal{A}(d,l)|} \tag{A3}$$

for all $\mathbf{x} \in \mathcal{A}(d,l)$. However, conditional unbiasedness does not extend to all equal probability designs. For example, conditional on the hexagonal tessellation employed by the FIA's unaligned systematic design it is possible to have $\pi_n(\mathbf{x}, l) = 0$ for some $\mathbf{x} \in \mathcal{A}(d,l)$ given $n(d,l)$. In particular, suppose $\mathcal{A}(d,l)$ spans one entire FIA phase 1 hexagon (see Bechtold and Patterson, 2005) slated for measurement in year l as well as portions of several other phase 1 hexagons; if $n(d,l) = 1$ then the conditional inclusion density function will be positive over the completely subsumed hexagon but must be 0 over the other intersected hexagons. This will generally result in bias. The above also assumes that $y_k = y(\mathbf{x}_k, l_k)$ is a point-measurement (or a measurement employing protocols suitably adjusted for boundary overlap) and that one can thus ignore any boundary overlap effects (see e.g., Gregoire, 1998).

The variance of $\bar{y}(d,l)$ for random $n(d,l)$ has no analytically tractable form as it is a function of the variability of both estimators in (A1). From Cordy (1993), under SRS the conditional variance of $\bar{y}(d,l)$ given $n(d,l)$ can be written in the familiar form

$$\mathsf{V}\left[\bar{y}(d,l) \middle| n(d,l)\right] = \frac{1}{n(d,l)\,|\mathcal{A}(d,l)|} \int_{\mathcal{A}(d,l)} \left[y(\mathbf{x}, l) - \lambda(d,l)\right]^2 \mathrm{d}\mathbf{x} = \frac{\sigma_y^2(d,l)}{n(d,l)} \tag{A4}$$

Furthermore, that variance can be (conditionally) unbiasedly estimated using

$$\widehat{\mathsf{V}}\left[\bar{y}(d,l)\right] = \frac{1}{n(d,l)\,[n(d,l)-1]} \sum_{k \in s(d,l)} \left[y_k - \bar{y}(d,l)\right]^2 = \frac{\hat{\sigma}_y^2(d,l)}{n(d,l)} \tag{A5}$$

For spatially structured designs such as the USFS FIA, the variance will be a function of more complex pairwise inclusion density functions (see Cordy, 1993). Moreover, it may not be possible to derive (conditionally) unbiased variance estimators because the pairwise inclusion density function can be 0 for sets of proximate locations. In such settings, estimator (A5) has been recommended as a conservative variance estimator in the sense that it is expected to overestimate variability in cases where the spatial design effectively reduces sampling error (e.g., Baffetta et al., 2009; see also Wolter, 2007, pp. 47–48). Alternatively, variance estimation strategies developed for systematic designs (e.g., Frank and Monleon, 2021) could be evaluated.

Appendix B: Mean, Variance, & MSE of the Indirect Estimator

Certain properties of the indirect domain estimator (3) follow directly from the results of **Appendix A**. These are extended below suppressing the parenthetical domain and lag dependence notation (d,l) unless necessary.

Under SRS the conditional expectation of $\hat{\tilde{y}}$ is a function of the distribution of y over the expanded spatiotemporal region $\tilde{\mathcal{A}} = \tilde{\mathcal{A}}(d,l)$, i.e.,

$$\mathsf{E}\left[\hat{\tilde{y}} \middle| \tilde{n}\right] = \frac{1}{|\tilde{\mathcal{A}}|} \int_{\tilde{\mathcal{A}}} y(\mathbf{x}, l)\, \mathrm{d}\mathbf{x} = \frac{|\mathcal{A}|}{|\tilde{\mathcal{A}}|}\, \lambda + \frac{|\tilde{\mathcal{A}}| - |\mathcal{A}|}{|\tilde{\mathcal{A}}|}\, \mathring{\lambda} = \tilde{\lambda}$$

where $\tilde{\lambda}$ is the density of y over $\tilde{\mathcal{A}}$ and $\mathring{\lambda}$ is the density of y over only the extra-domain region supplying additional data. The conditional bias of (3) as an estimator of λ will therefore be a function of the extent to which the density of y over the "small area" $\mathcal{A}$ differs from that over the "large area" $\tilde{\mathcal{A}}$. Additionally, under SRS the conditional variance of $\hat{\tilde{y}}$ can be written as

$$\mathsf{V}\left[\hat{\tilde{y}} \middle| \tilde{n}\right] = \frac{1}{\tilde{n}\,|\tilde{\mathcal{A}}|} \int_{\tilde{\mathcal{A}}} \left[y(\mathbf{x}, l) - \tilde{\lambda}\right]^2 \mathrm{d}\mathbf{x} = \frac{\tilde{\sigma}_y^2}{\tilde{n}}$$

where $\tilde{\sigma}_y^2$ is the variance in y over $\tilde{\mathcal{A}}$. This conditional variance can be unbiasedly estimated using

$$\widehat{\mathsf{V}}\left[\hat{\bar{y}}\right] = \frac{1}{\tilde{n}\left[\tilde{n}-1\right]}\sum_{k\in\tilde{s}}\left[y_k - \hat{\bar{y}}\right]^2 \tag{A6}$$

conditional on the realized sample size $\tilde{n}$. As for the direct sample mean (2) these results do not extend generally to other (equal or unequal probability) spatial designs, but equation (A6) can again be applied as a conservative estimator of variance.

To describe the MSE of $\hat{\bar{y}}$, it is useful to note that it can be broken down much like its expectation above

$$\hat{\bar{y}} = \frac{1}{\tilde{n}}\sum_{k\in\tilde{s}} y_k = \frac{1}{\tilde{n}}\left(\sum_{k\in s} y_k + \sum_{\substack{k\in\tilde{s}\\ k\notin s}} y_k\right) = \frac{n}{\tilde{n}}\,\bar{y} + \frac{\tilde{n}-n}{\tilde{n}}\,\ddot{\bar{y}} \tag{A7}$$

where $\ddot{\bar{y}} = \ddot{\bar{y}}(d,t)$ is the mean of the observations in $\tilde{s}$ but not in s (i.e., of the observations that have been borrowed from outside the domain of interest). Then, adopting the approach used by Rao and Molina (2015, p. 43), write the conditional MSE of the indirect estimator (3) given n as

$$\begin{aligned}\mathsf{MSE}\left[\hat{\bar{y}}\middle|n\right] &= \mathsf{E}\left[\left(\hat{\bar{y}}-\bar{y}\right)^2\middle|n\right] + \mathsf{E}\left[\left(\bar{y}-\lambda\right)^2\middle|n\right]\\ &\quad + 2\,\mathsf{E}\left[\left(\hat{\bar{y}}-\bar{y}\right)\left(\bar{y}-\lambda\right)\middle|n\right]\\ &= \mathsf{E}\left[\left(\hat{\bar{y}}-\bar{y}\right)^2\middle|n\right] + \mathsf{E}\left[\left(\bar{y}-\lambda\right)^2\middle|n\right]\\ &\quad - 2\,\mathsf{E}\left[\bar{y}\left(\bar{y}-\lambda\right)\middle|n\right] + 2\,\mathsf{E}\left[\hat{\bar{y}}\left(\bar{y}-\lambda\right)\middle|n\right]\end{aligned} \tag{A8}$$

The second and third terms on the right hand side of (A8) relate to the variability of $\bar{y}(d,t)$ while the last term connects to the association between $\bar{y}(d,t)$ and $\hat{\bar{y}}(d,t)$. Indeed, under SRS, (A8) can be simplified to

$$\mathsf{MSE}\left[\hat{\bar{y}}\middle|n\right] = \mathsf{E}\left[\left(\hat{\bar{y}}-\bar{y}\right)^2\middle|n\right] - \mathsf{V}\left[\bar{y}|n\right] + 2\,\mathsf{C}\left[\hat{\bar{y}},\bar{y}\middle|n\right] \tag{A9}$$

where $\mathsf{C}\left[\hat{\bar{y}},\bar{y}\middle|n\right]$ denotes (conditional) covariance. Further simplification is possible under SRS by focusing on the covariance term

$$\begin{aligned}\mathsf{C}\left[\hat{\bar{y}},\bar{y}\middle|n\right] &= \mathsf{E}\left[\hat{\bar{y}}\left(\bar{y}-\lambda\right)\middle|n\right]\\ &= \mathsf{E}\left\{\mathsf{E}\left[\hat{\bar{y}}\left(\bar{y}-\lambda\right)\middle|n,\tilde{n},s\right]\middle|n\right\}\\ &= \mathsf{E}\left\{\left(\bar{y}-\lambda\right)\mathsf{E}\left[\hat{\bar{y}}\middle|n,\tilde{n},s\right]\middle|n\right\}\end{aligned} \tag{A10}$$

Substituting (A7), the inner expectation of (A10) becomes

$$\mathsf{E}\left[\hat{\bar{y}}\middle|n,\tilde{n},s\right] = \frac{n}{\tilde{n}}\,\bar{y} + \frac{\tilde{n}-n}{\tilde{n}}\,\mathsf{E}\left[\ddot{\bar{y}}|n,\tilde{n},s\right] = \frac{n}{\tilde{n}}\,\bar{y} + \frac{\tilde{n}-n}{\tilde{n}}\,\mathring{\lambda}$$

Thus,

$$\begin{aligned}\mathsf{C}\left[\hat{\bar{y}},\bar{y}\middle|n\right] &= \mathsf{E}\left\{\left(\bar{y}-\lambda\right)\frac{n}{\tilde{n}}\,\bar{y} + \left(\bar{y}-\lambda\right)\frac{\tilde{n}-n}{\tilde{n}}\,\mathring{\lambda}\,\middle|n\right\}\\ &= \mathsf{E}\left\{\frac{n}{\tilde{n}}\middle|n\right\}\mathsf{E}\left\{\left(\bar{y}-\lambda\right)\bar{y}|n\right\} + \mathring{\lambda}\,\mathsf{E}\left\{\frac{\tilde{n}-n}{\tilde{n}}\left(\bar{y}-\lambda\right)\middle|n\right\}\\ &= \mathsf{E}\left\{\frac{n}{\tilde{n}}\middle|n\right\}\mathsf{V}\left[\bar{y}|n\right]\end{aligned}$$

Finally, substituting this last result into (A9) gives

$$\mathsf{MSE}\left[\hat{\bar{y}}\,\middle|n\right] = \mathsf{E}\left[\left(\hat{\bar{y}}-\bar{y}\right)^2\middle|n\right] - \mathsf{V}\left[\bar{y}|n\right]\left[1 - 2\,\mathsf{E}\left\{\frac{n}{\tilde{n}}\middle|n\right\}\right] \tag{A11}$$

Note that if $\tilde{n} = n$ so that no observations are borrowed and that therefore $\hat{\bar{y}} = \bar{y}$, then $\mathsf{MSE}\left[\hat{\bar{y}}\,\middle|n\right]$ collapses to simply $\mathsf{V}\left[\bar{y}|n\right]$, as it should. However, if data are drawn from a much larger area such that $\tilde{n} \gg n$ then $\mathsf{MSE}\left[\hat{\bar{y}}\,\middle|n\right]$ tends to $\mathsf{E}\left[\left(\hat{\bar{y}}-\bar{y}\right)^2\middle|n\right] - \mathsf{V}\left[\bar{y}|n\right]$. The latter expression is suggested as an approximation by Rao and Molina (2015, p. 44), but will be too small unless data are drawn from a substantially larger area than the domain of interest. Finally, note again that this expression applies in the case of SRS, but not more generally.

Expression (A11) suggests a simple sample-based estimator of the conditional MSE

$$\widehat{\mathsf{MSE}}\left[\hat{\bar{y}}\middle|n\right] = \left(\hat{\bar{y}}-\bar{y}\right)^2 - \frac{\hat{\sigma}_y^2}{n}\left[1 - 2\,\frac{n}{\tilde{n}}\right] \tag{A12}$$

This estimator differs from the framework suggested by Rao and Molina (2015, p. 44) only by the factor $\left[1-2\,\frac{n}{\tilde{n}}\right]$; this factor guarantees larger estimates of MSE, but still cannot guarantee non-negative estimates. We are unaware of any investigation of its sampling properties, however.

frontiers in Forests and Global Change

PERSPECTIVE
published: 06 December 2021
doi: 10.3389/ffgc.2021.763487

United States Forest Service Use of Forest Inventory Data: Examples and Needs for Small Area Estimation

Sarah S. Wiener[1], Renate Bush[2], Amy Nathanson[3], Kristen Pelz[4], Marin Palmer[5], Mara L. Alexander[1], David Anderson[6], Emrys Treasure[3], Joanne Baggs[3] and Ray Sheffield[7]*

[1] Ecosystem Management Coordination, USDA Forest Service, Washington, DC, United States, [2] Northern Region, USDA Forest Service, Missoula, MT, United States, [3] Southern Region, USDA Forest Service, Atlanta, GA, United States, [4] Rocky Mountain Research Station, USDA Forest Service, Santa Fe, NM, United States, [5] Pacific Northwest Region, USDA Forest Service, Portland, OR, United States, [6] Southwestern Region, USDA Forest Service, Albuquerque, NM, United States, [7] Retired, USDA Forest Service, Hendersonville, NC, United States

OPEN ACCESS

Edited by:
Philip Radtke, Virginia Tech, United States

Reviewed by:
Roque Rodríguez-Soalleiro, University of Santiago de Compostela, Spain
Matija Klopčič, University of Ljubljana, Slovenia

***Correspondence:**
Sarah S. Wiener
sarah.s.wiener@usda.gov

Specialty section:
This article was submitted to Forest Management, a section of the journal Frontiers in Forests and Global Change

Received: *24 August 2021*
Accepted: *09 November 2021*
Published: *06 December 2021*

Citation:
Wiener SS, Bush R, Nathanson A, Pelz K, Palmer M, Alexander ML, Anderson D, Treasure E, Baggs J and Sheffield R (2021) United States Forest Service Use of Forest Inventory Data: Examples and Needs for Small Area Estimation. Front. For. Glob. Change 4:763487. doi: 10.3389/ffgc.2021.763487

Forest Inventory and Analysis (FIA) data provides robust information for the United States Forest Service's (USFS) mid-to-broad-scale planning and assessments, but ecological challenges (i.e., climate change, wildfire) necessitate increasingly strategic information without significantly increasing field sampling. Small area estimation (SAE) techniques could provide more precision supported by a rapidly growing suite of landscape-scale datasets. We present three Regional case studies demonstrating current FIA uses, how SAE techniques could enhance existing uses, and steps FIA could take to enable SAE applications that are user-friendly, comprehensive, and statistically appropriate. The Northern Region uses FIA data for planning and assessments, but SAE techniques could provide more specificity to guide vegetation management activities. State and transition simulation models (STSM) are run with FIA data in the Southwestern Region to predict effects of treatments and disturbances, but SAE could support model validation and more precision to identify treatable areas. The Southern Region used FIA to identify existing longleaf pine stands and evaluate condition, but SAE techniques within FIA tools would streamline analyses. Each case study demonstrates a desire to have FIA data on non-forested conditions and non-tree variables. Additional tools to measure statistical confidence would help maximize utility. FIA's SAE techniques could add value to a widely used data set, if FIA can support key supplements to basic data and functionality.

Keywords: small area estimation (SAE), Forest Inventory and Analysis (FIA), United States Forest Service, forest planning, forest assessment, National Forest System, forest management

INTRODUCTION

The United States Forest Service's (USFS) National Forest System (NFS) manages 78 million hectares of National Forests and Grasslands. NFS is legally bound to a multiple-use mandate (i.e., timber, recreation, watersheds, and wildlife), which creates complex decision-making environments and diverse information needs. With a vast land base challenged by climate change

and increasing wildfire intensity, and a proportionally limited ability to actively manage forest area, collect vegetation data, and analyze and interpret data due to budget and staffing constraints, NFS has a critical need for strategic information that can support adaptive management at the scale of the challenge without greatly increasing data collection and analyses.

Small area estimation (SAE) is a statistical technique used to enhance data in a specific area (i.e., geographic, demographic) with data not confined to that area (Rao, 2003; Jiang and Rao, 2020). SAE borrows strength from larger areas and uses auxiliary information to establish relationships with the response. With National Forest Inventories (NFIs), SAE can integrate auxiliary data (i.e., remote sensing, climate layers, and landscape-scale geospatial data) with field-sampled data. For example, NFIs in Scandinavia were combined with satellite and other geospatial data to parameterize image data and perform pre-processing, enabling enhancement of various monitoring applications (Tomppo et al., 2008). Models improve with more highly correlated auxiliary information and response data, and with higher resolution auxiliary information. For more information on SAE, see Ghosh and Rao (1994); Rao (2003), Pfeffermann (2013); Jiang and Rao (2020). Given NFS' limited capacity for additional field sampled vegetation data and increased availability of landscape scale data, SAE using NFI data could support land management planning for NFS.

The Forest Inventory and Analysis (FIA) program (the NFI for NFS) is the most comprehensive and consistent national vegetation data set for the agency, delivering a unique set of field-measured data and accompanying analysis tools that provide baseline information and the ability to monitor current vegetation conditions through repeated measurement of permanent plots. FIA operates across all United States land the program defines as "forested" (generally, 10% tree canopy cover) (USFS, 2021b) and uses an annualized, repeated sampling system designed to make estimates of forested land vegetation conditions across multiple scales. FIA plots are on a semisystematic sampling grid. Locations are unbiased geographically, with approximately one plot per 2,428 hectares of forested land, and plot data are collected according to the FIA protocol (USFS, 2021b) in a largely nationally consistent way (Bechtold and Patterson, 2005). FIA forest-plot data are remeasured every 10 years in the western United States, and every 7 or 5 years in the eastern and southern United States (McRoberts et al., 2005). Data about trees and associated characteristics are collected on all inventoried plots (with some differences in tree data among the four FIA units (USFS, 2021c). Additional information about down-woody material, understory vegetation, and noxious weeds may be collected depending upon FIA unit.

Forest Inventory and Analysis data are useful for NFS to assess vegetation conditions at the national to Regional scale. NFS contains nine Regions that each manage approximately 9 to 14 million hectares (USFS, 2020). At this scale, with approximately 1 plot per 2,428 hectares of forested land, plot numbers are sufficient for estimates to have small errors, even when broken into multiple sub-categories (such as forest land area, with large trees present, by forest type). Most individual National Forests or Grasslands (hereafter referred to as Units) are at least 100 thousand hectares, with most western Units over 300 thousand hectares, and up to 1.7 million hectares (USFS, 2020). Statistical analysis suggest that estimates are unbiased when there are 10 forested plots per land ownership type, such as on NFS land with over 24 thousand forested hectares (Westfall et al., 2011). At the Unit scale, plot numbers are usually sufficient (Units with 100 thousand forested hectares should have about 40 plots) for small errors and confident estimates, particularly for uncomplicated queries (i.e., total forest land area, forest land area by major forest type), but errors increase for more complicated queries. If users can interpret and judge levels of uncertainty acceptable around estimates, FIA data are appropriate for a variety of mid- to broad-scale needs for Regional and Unit monitoring, Forest Plan revision, and assessments. Core FIA data using standard estimation procedures are sufficient for many information needs, though data users may require increased precision (more plots) for certain estimates and scales, where SAE could assist.

Forest Inventory and Analysis SAE techniques are under development and not used programmatically by NFS. However, opportunities exist to enhance NFS' ability to monitor ecosystems with SAE, particularly by integrating remote sensing data (Lister et al., 2020). SAE techniques would expand the utility of FIA information for NFS, and could in certain circumstances replace the need for adding FIA plots within a geographic area (known as intensification), by providing better estimates at smaller scales. Having reliable estimates with precision information, that are spatially and temporally appropriate for management questions, would help land managers understand current condition and monitor trends. SAE techniques would expand the ability of NFS to make informed decisions on where, for example, specific wildlife habitat is located, the condition of the habitat, and habitat changes through time. SAE could provide estimates based on NFS classifications or algorithms about specific small areas with smaller error than currently possible using FIA's plot data, which would support Regional and Unit-based monitoring and allow FIA data application with enhanced confidence to inform management.

Because NFS SAE techniques are under development, we provide three Regional case studies of FIA uses without SAE, which demonstrate varied data applications and analysis techniques. Case studies include descriptions of how SAE could improve these applications and how specific enhancements to FIA data could better support SAE from the perspective of NFS FIA users.

CASE STUDIES

Northern Region (R1): Using Forest Inventory and Analysis Data for Land Management Assessments and Biennial Monitoring

Forest Inventory and Analysis data are used for assessments, planning and implementation of management, and monitoring extensively in R1. To evaluate current vegetation condition, R1 developed a hierarchical existing vegetation classification system

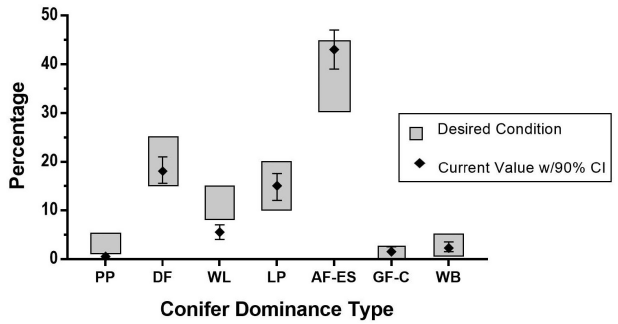


FIGURE 1 | Comparison of existing condition to desired conditions (rectangles) for ponderosa pine (PP), western larch (WL), and spruce-subalpine fir (AF-ES) Dominance Type Groups for the Flathead National Forest. Existing condition (diamond) and the 90% confidence interval (whiskers) of estimates of Dominance Types using Forest Inventory and Analysis (FIA) inventory data.

(R1 ExVeg Classification; Barber et al., 2011) to attribute lifeform, alliance, cover type, and dominance types groups (DTG) from FIA data. This system aligns with USFS technical guidance through the Existing Vegetation Classification, Mapping, and Inventory Technical Guide (Nelson et al., 2015). Applying the R1 ExVeg Classification algorithms to FIA data allows Units to derive estimates, with confidence intervals, of DTG distribution to understand vegetation composition across a Unit. Current condition can then be compared to natural ranges of variability to develop desired conditions for ecological integrity and guide vegetation management. **Figure 1** displays estimates of Dominance Types for the 970-thousand-hectare Flathead National Forest from the Unit's most recent Land Management Plan, compared to desired conditions. The Flathead National Forest seeks to increase ponderosa pine (PP) and western larch (WL) DTGs while decreasing spruce-subalpine fir. Since FIA plots are remeasured every 10-years in R1, Dominance Type algorithms will be applied longitudinally to monitor progress toward desired conditions.

R1 partnered with FIA to collect information across the entire FIA plot footprint, not just the "forested condition" portion to enable expanding Dominance Type classifications and algorithms for non-tree dominated systems. Having consistent sampling protocols across the entire plot allows estimates and confidence intervals to be derived regardless of the presence of trees. This allows R1 to use FIA data to inform assessments, analysis, and monitoring across all NFS land types managed by the Region.

Small area estimation techniques could enhance use of FIA data in R1 for assessments and planning activities by deriving more precise estimates of DTGs within the biophysical setting and geographic areas used for goals and objectives in Forest Plans. Estimates of DTGs could also be monitored at a finer geographic scale, allowing the Unit to better understand current condition, prioritize vegetation management, and monitor trends. Using SAE, these goals could be accomplished by relying more on remote sensing and other auxiliary data and less on costly field data collection. For SAE to be meaningful to NFS, Existing Vegetation Classification algorithms should be used in SAE techniques, and all data collected nationally by FIA (i.e., including non-forested condition and non-tree data) should be utilized in the estimates. This would allow more accurate estimates and monitoring of attributes derived from FIA data such as distribution of old growth, large-tree and snag densities, and wildlife species habitat models. Precise estimates for smaller geographic areas could alleviate the need for plot intensification but cannot entirely replace field data collection within project areas.

To enable monitoring trends within non-forested areas, all data that is consistently collected by FIA across the Unit should be available within FIA products and tools and utilized for SAE techniques, including non-tree centric protocols that support algorithms for non-tree dominated systems. This would allow Units to understand vegetation composition as it changes over time, and monitor the extent of sagebrush cover, fuel loadings, potential fire behavior, and tree encroachment onto non-forested areas.

Finally, for SAE to be useful to R1, we also desire information on when the reliability of the estimates deteriorates. NFS should work with FIA to explore which attributes can be estimated at which resolution.

Southwestern Region (R3): Using Forest Inventory and Analysis Data to Estimate State and Transition Model Parameters and Inform Vegetation Mapping

R3 has used FIA data for nearly two decades to inform forest planning decisions. Around 2005 R3 began to revise the Region's eleven Forest Plans due to concerns about Mexican Spotted Owl and Northern Goshawk habitat sustainability. To ease the analytical burden on national forest staff, be regionally consistent, and utilize the best available scientific information, R3 uses state and transition simulation models (STSM; Daniel et al., 2016) to assess future vegetation conditions under a range of management actions. STSM's classify a landscape into a set of distinct states. Probabilistic transitions describe the change from one state to another due to succession and disturbance, both human and natural. FIA data was a primary source to parameterize the STSMs. Parameters consist of a set of probabilities that describe the transition from one state to another for natural successional processes and a suite of disturbance regimes such as wildfire, insect and disease, silvicultural prescriptions, and prescribed burning.

Regionally consistent vegetation modeling processes require all models to start with the same initial vegetation conditions. In cooperation with the Oregon State Institute of Natural Resources, R3 completed a mid-scale vegetation database covering Arizona and New Mexico. Gradient nearest neighbor (GNN) techniques (Ohmann and Gregory, 2002) and random forest classification were used for attribute imputation. Forest attributes came from FIA plots. Additional processing of FIA plot data using the Forest Vegetation Simulator (FVS, a forest growth simulation model) (USFS, 2021d) to produce the stand-level outputs provided additional information for forested polygons. GNN techniques

used several auxiliary geospatial datasets to assign FIA plots to landscape location, including National Elevation Data, soils data, and texture metrics derived from National Agriculture Imagery Program data.

Forest Inventory and Analysis plots were stratified into states by potential vegetation type, size class, canopy cover percentage, and number of stories. After stratification FIA plots were used as the tree list inputs into FVS. FVS outputs were classified into states using the stratification criteria. The number of FIA plots that changed from one state to a different state in each time step divided by the number of plots in the initial state determined the transition probabilities, which help determine which management activities will steer the forest toward desired conditions. For a complete description of the analytical techniques consult Weisz et al. (2010) or Weisz and Vandendriesche (2012).

With the imminent completion of all eleven Forest Plans, the analytical framework developed using FIA data with FVS processing is being adapted to run landscape level vegetation management projects at the 40 thousand-hectare scale. Preliminary work is favorable for the continued use of FIA data at this project size.

Small area estimation could enhance these techniques with more precise estimates of delineations for identifying short-term treatable areas, particularly Northern Goshawk and Mexican Spotted Owl habitat. SAE shows promise in modeling wildlife habitat more precisely than regional models (Wilson et al., 2009), but more research is needed to support these applications. SAE could also support identifying locations and quantities for old growth forest and large trees, which are preferred by these two species. Finally, STSM validation could utilize SAE by examining effects of small treatment areas and small disturbances to determine if treatment effectiveness and direction of disturbance levels align with model output.

Providing data that is more readily accessible for automated analyses would facilitate SAE applications. About seventy distinct ecosystem types occur in R3 (USFS, 2014) ranging from semi-desert grasslands to alpine tundra. Having FIA data on these ecosystem types, and integrating these data with similar data collected by the Natural Resources Conservation Service and Bureau of Land Management, would support more comprehensive analyses. Providing those data in a format compatible with the Range Vegetation Simulator [RVS; Reeves (2016)], similar to the FVS ready data currently provided with FIA databases, would simplify processes.

Southern Region (R8): Evaluating Existing Longleaf Pine Ecosystem Condition With Forest Inventory and Analysis

Longleaf pine ecosystems in the Southern Region have declined to 3% of their original distribution (America's Longleaf Regional Working Group, 2009). There is high interest from the USFS and partners in maintaining and restoring these forest types due to their high biological diversity and importance as wildlife habitat. Understanding location and current condition of these ecosystems is vital to restoration efforts. Estimates of existing area of longleaf pine ecosystems derived from FIA frequently rely on composition criteria (forest type) that do not capture key characteristics of these ecosystems, particularly forest structure.

The Range-Wide Conservation Plan for Longleaf Pine (America's Longleaf Regional Working Group, 2009) established condition-based restoration goals for 2025. When the Plan was published, analysis techniques for estimating condition classes were not available. The 2009 estimate of 1.4 million existing hectares came from a combination of FIA data for non-NFS lands and local inventory data for NFS lands. The latter primarily came from the FSVeg (USFS, 2021a) database, which contains the agency's Common Stand Exam data. The split between "good condition/maintain" and "poor condition/restore" was based on professional judgment, informed by understanding local fire regimes with limited field sampling. Recently, NatureServe (a non-profit organization that assembles data on species and ecosystems) led an interagency effort to develop improved definitions of condition classes for longleaf pine ecosystems (NatureServe, 2016; Nordman et al., 2016). There are 13 Open Pine Metrics: 5 canopy, 4 midstory/shrub, and 4 ground layer. Each Metric has designated thresholds for each condition class (excellent, good, fair, or poor), which are combined to produce an overall condition score. A simplified version of the Open Pine Metrics was adopted in R8's strategic direction regarding longleaf restoration.

To advance our understanding of existing longleaf pine ecosystem conditions across R8, we applied the Open Pine Metrics to FIA data. We selected all FIA plots that contained at least one longleaf pine, and used the relevant FIA plot measurements (height, species, basal area, etc.) to assign a score. Note that the FIA protocols (USFS, 2021b) for R8 only collect sufficient data to score 7 of the 13 metrics. We also scored factors such as fire tolerance, that are not included in the FIA protocols.

Figure 2 shows preliminary results. This approach allowed us to assess longleaf extent and condition regardless of assigned forest type, and we estimated considerably more area occupied by longleaf pine ecosystems than previous estimates derived from FIA based on forest-type alone. Note that the current method is likely overestimating area in each condition class by including plots where moving toward longleaf pine-dominated systems is not desired. Also, because this analysis was conducted by a contractor outside of standard FIA analysis tools that provide statistical error information, and it was not part of the contract request, statistical confidence intervals were not part of this analysis.

Still, this preliminary analysis shows promising results in characterizing existing condition. Results have strategic value, are firmly rooted in current best available science, and use the most robust inventory data available (FIA). However, the challenges with calculating measures of statistical confidence using this methodology are a hindrance, especially when the estimates are calculated for smaller scales. If SAE techniques were integrated within existing FIA tools that include integrated calculations of statistical confidence, these types of analyses could be simplified, streamlined, and performed consistently across R8 and the agency.

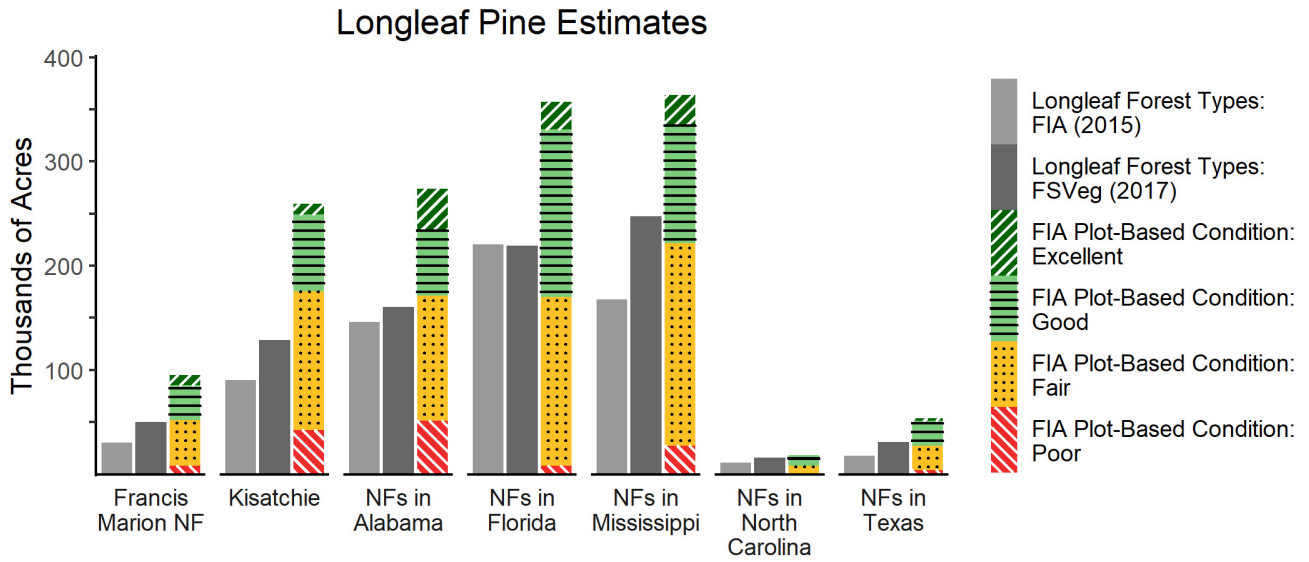


FIGURE 2 | Longleaf pine estimates for Southern Region (R8) National Forests. Longleaf extent based on forest type [FIA/ field sampled vegetation (FSVeg)] does not include any measure for condition and may miss areas that have a longleaf component but were not classified as a longleaf forest type. FIA plot-based condition estimates include any FIA plot that contained at least a single longleaf pine, and therefore may overestimate the extent of longleaf ecosystems in places where longleaf is a minor ecosystem component.

The addition of non-tree variables such as shrub, grass, forb, and invasive plant cover would enable a more accurate analysis with all 13 Open Pine Metrics. The ability to easily combine FIA data with local inventory data, including using metrics that span scales and inventory systems, could further enhance these analyses. Finally, enhanced functionality of FIA tools would simplify similar analyses – R8 had to hire a contractor with specialized skills to implement the longleaf condition assessment outside of standard FIA analysis tools.

RECOMMENDATIONS AND CONCLUSION

These three case studies demonstrate how SAE techniques could enhance and expand existing applications of FIA data for NFS users to meet planning and management information needs. SAE using FIA data, coupled with auxiliary data such as remote sensing, would improve the ability to monitor key ecosystem components spatially while providing consistent confidence intervals to accompany estimates. More precise, comprehensive, and consistent vegetation information will support more strategic decision making by providing land managers information on current condition and trends over time. This enables tactically targeting areas for management actions, restoration strategies, and more intensive monitoring. In the face of climate change, understanding the impact of management activities is imperative to practicing adaptive management, and SAE with FIA data can improve our understanding without greatly increasing costly field data collection.

For FIA to most effectively support SAE techniques for NFS needs, baseline FIA data should comprehensively and consistently support the assessment of diverse forest and non-forest ecosystems managed by the agency and its multiple-use mandate. SAE techniques could ultimately reduce some of the need for field-sampled vegetation to meet information needs of NFS, but some initial expansions in the variables and locations of FIA data collected would best support widespread use of SAE. The data expansions proposed below would enable SAE across all NFS lands, supporting a multitude of information needs with improved consistency and scientific integrity.

Specifically, NFS desires information collected across the entire FIA plot, and not only on those portions that meet FIA's definition of forested. This would allow monitoring of vegetation conditions across their entire land base. Without data from non-forest areas, it is difficult to disentangle FIA "forested" land definition changes from actual changes in tree densities and ecosystem shifts, such as those that may be occurring due to climate change. This is particularly important in the Western United States where non-forest land cover is common inside NFS boundaries. Standard FIA protocols for the "All Condition Inventory" (ACI) are available, and are collected on all plots with "non-forest" condition on certain NFS lands, including in Regions 1, 4, 6, and 10 (i.e., USFS, 2011). The ability to use "ACI" data should be available to all NFS Regions and available for analysis in the NFS analysis tools, allowing NFS classifications (i.e., wildlife habitat models and existing vegetation classifications) to be applied, stored, and used in estimations for all NFS land. This functionality would enable SAE applications within existing workflows and reduce training and workload required for NFS staff to apply SAE techniques.

Finally, NFS will desire information on the scale at which FIA-derived estimates become unreliable (and some estimates will be

more robust than others given inherent variability in the attribute and modeling techniques). Reliability of SAE will vary depending upon how common or rare the attribute of interest is, and this potential limitation should be considered prior to reporting these estimates (Moisen et al., 2004). Guidance and assistance are also needed to integrate finer-scale spatial datasets in SAE products. Ultimately, FIA's SAE techniques will not replace site-specific stand exam data, but will help NFS be more targeted in selecting sites for field reconnaissance and collection of site-specific information, further expanding the uses of FIA data.

Small area estimation techniques could broaden the applicability of a data set that is widely used by the NFS, and with certain additions and enhancements to FIA data and tools, NFS users can be more precise, accurate, consistent, and comprehensive in their analytical capabilities to inform good forest management across a complex 78-million-hectare land base.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.fia.fs.fed.us/tools-data/.

AUTHOR CONTRIBUTIONS

SW coordinated the manuscript and was in charge of overall edits and writing of abstract and conclusion. RB wrote the R1 case study, contributed to the framing and introduction of the manuscript, and overall edits. AN was the lead on the R8 case study, developed **Figure 2**, and contributed to overall framing and edits. KP contributed to manuscript framing, introduction, and overall edits, and wrote the background information on FIA. MP contributed to framing of the manuscript and provided written contributions to introduction and background. MA contributed to framing of the manuscript and overall edits. ET contributed to overall framing and the R8 case study. JB and RS contributed to the R8 case study. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The authors would like to thank Carl Nordman of NatureServe and Rickie White, formerly of NatureServe but currently at Ellerbe Creek Watershed Association, for contributions to the R8 case study.

REFERENCES

America's Longleaf Regional Working Group (2009). *Range-Wide Conservation Plan for Longleaf Pine*. Available online at: https://americaslongleaf.org/resources/conservation-plan/ (accessed January 15, 2021).

Barber, J., Berglund, D., and Bush, R. (2011). *The Region 1 Existing Vegetation Classification System and its Relationship to Inventory Data and the Region 1 Existing Vegetation Map Products. Region 1 Vegetation Classification, Mapping, Inventory, and Analysis Report # 11-10*. Missoula, MT: USDA Forest Service.

Bechtold, W. A., and Patterson, P. L. (2005). *The Enhanced Forest Inventory and Analysis Program—National Sampling Design and Estimation Procedures. General Technical Report (GTR). SRS-80*. Ashville, NC: U.S. Department of Agriculture.

Daniel, C. J., Frid, L., Sleeter, B., and Fortin, M. J. (2016). State-and-transition simulation models: a framework for forecasting landscape change. *Methods Ecol. Evol.* 7, 1413–1413. doi: 10.1111/2041-210X.12597

Ghosh, M., and Rao, J. (1994). Small area estimation: an appraisal. *Stat. Sci.* 9, 55–76. doi: 10.1214/ss/1177010647

Jiang, J., and Rao, J. (2020). Robust small area estimation: an overview. *Annu. Rev. Stat. Appl.* 7, 337–360. doi: 10.1146/annurev-statistics-031219-041212

Lister, A., Andersen, H., Frescino, T., Gatziolis, D., Healey, S., Heath, L., et al. (2020). Use of remote sensing data to improve the efficiency of national forest inventories: a case study from the united states national forest inventory. *Forests* 11:12. doi: 10.3390/f11121364

McRoberts, R., Bechtold, W., Patterson, P., Scott, C., and Reams, G. (2005). The enhanced forest inventory and analysis program of the USDA forest service: historical perspective and announcement of statistical documentation. *J. For.* 103, 304–308.

Moisen, G. G., Blackard, J. A., and Finco, M. (2004). "Small area estimation in forests affected by wildfire in the interior West," in *Proceedings of the 10th Forest Service Remote Sensing Applications; Remote Sensing for Field Users*, ed. J. D. Greer (Salt Lake City, UT: American Society of Photogrammetry and Remote Sensing).

NatureServe (2016). *NatureServe Explorer [web application]*. Arlington, TX: NatureServe.

Nelson, M. L., Brewer, C. K., and Solem, S. J. (eds) (2015). *Existing Vegetation Classification, Mapping, and Inventory Technical Guide, Version 2.0. Gen. Tech. Rep. WO–90*. Washington, DC: U.S. Department of Agriculture.

Nordman, C., White, R., Wilson, R., Ware, C., Rideout, C., Pyne, M., et al. (2016). *Rapid Assessment Metrics to Enhance Wildlife Habitat and Biodiversity within Southern Open Pine Ecosystems, Version 1.0*. Washington, DC: U.S. Fish and Wildlife Service and NatureServe.

Ohmann, J. L., and Gregory, M. J. (2002). Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, U.S.A. *Can. J. For.* 32, 725–741. doi: 10.1139/x02-011

Pfeffermann, D. (2013). New important developments in small area estimation. *Stat. Sci.* 28, 40–68. doi: 10.1214/12-STS395

Rao, J. (2003). Some new developments in small area estimation. *J. Iran. Stat. Soc.* 2, 145–169. doi: 10.1002/0471722189

Reeves, M. C. (2016). *Development of the Rangeland Vegetation Simulator: A Module of the Forest Vegetation Simulator. A Report to the Joint Fire Sciences Program. JFSP 14-S-01-01, 12-1-02-15, 12-1-02-15*. Missoula, MT: Forestry Sciences Lab, 129.

Tomppo, E., Olsson, H., Ståhl, G., Nilsson, M., Hagner, O., and Katila, M. (2008). Combining national forest inventory field plots and remote sensing data for forest databases. *Remote Sens. Environ.* 112, 1982–1999. doi: 10.1016/j.rse.2007.03.032

USFS (2011). *Interior West Forest Inventory and Analysis Region 4 All Condition Inventory Supplemental Field Guide*. Available online at: https://www.fs.fed.us/rm/ogden/data-collection/pdf/r4_acp_manual_5162011_wo_markup.pdf (accessed January 15, 2021).

USFS (2020). *Land Area Report*. Available online at: https://www.fs.fed.us/land/staff/lar/LAR2020/FY2020_LAR_Book.pdf (accessed October 4, 2021).

USFS (2021b). *Forest Inventory and Analysis National Core Field Guide, Version 9.1*. Available online at: https://www.fia.fs.fed.us/library/field-guides-methods-proc/docs/2021/core_ver9-1_9_2021_final.pdf (accessed October 5, 2021).

USFS (2021c). *Forest Inventory and Analysis National Program*. Available online at: https://www.fia.fs.fed.us/about/about_us/index.php (accessed October 5, 2021).

USFS (2021d). *Forest Vegetation Simulator*. Available online at: https://www.fs.fed.us/fvs/ (accessed July 1, 2021).

USFS (2021a). *Field Sampled Vegetation*. Available online at: https://www.fs.fed.us/nrm/fsveg/ (accessed August 1, 2021).

USFS (2014). *Ecological Response Units of the Southwestern United States. USDA Forest Service Technical Report*. Available online at: https://www.fs.usda.gov/Internet/FSE_DOCUMENTS/fseprd609789.pdf (accessed November 15, 2018).

Weisz, R., Triepke, J., Vandendriesche, D., Manthei, M., Youtz, J., Simon, J., et al. (2010). "Evaluating the ecological sustainability of a pinyon-juniper grassland ecosystem in northern Arizona," in *Proceedings of the 2009 National Silviculture Workshop Integrated Management of Carbon Sequestration and Biomass Utilization Opportunities in a Changing Climate*, eds T. B. Jain, R. T. Graham, and J. Sandquist (Boise, ID: U.S. Department of Agriculture), 321–336.

Weisz, R., and Vandendriesche, D. (2012). "Use of the forest vegetation simulator to quantify disturbance activities in state and transition models," in *Proceedings of the First Landscape State-and-Transition Simulation Modeling Conference*, eds B. K. Kerns, A. J. Shlisky, and C. J. Daniel (Portland, OR: U.S. Department of Agriculture), 143–160.

Westfall, J. A., Patterson, P. L., and Coulston, J. W. (2011). Post-stratifed estimation: within-strata and total sample size recommendations. *Can. J. For. Res.* 41, 1130–1139. doi: 10.1139/x11-031

Wilson, D., Stoddard, M., Betts, M., and Puettmann, K. (2009). Bayesian small area models for assessing wildlife conservation risk in patchy populations. *Conserv. Biol.* 23, 982–991. doi: 10.1111/j.1523-1739.2008.01160.x

Conflict of Interest: The 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.

frontiers in Forests and Global Change

ORIGINAL RESEARCH
published: 15 December 2021
doi: 10.3389/ffgc.2021.764495

A Comparison of Model-Assisted Estimators, With and Without Data-Driven Transformations of Auxiliary Variables, With Application to Forest Inventory

Magnus Ekström[1,2]* *and Mats Nilsson*[1]

[1] Department of Forest Resource Management, Swedish University of Agricultural Sciences, Umeå, Sweden, [2] Department of Statistics, Umeå School of Business, Economics and Statistics, Umeå University, Umeå, Sweden

OPEN ACCESS

Edited by:
Barry Wilson, Northern Research Station, United States Forest Service, United States Department of Agriculture (USDA), United States

Reviewed by:
Michael Goerndt, Missouri State University, United States
Jacob Strunk, Pacific Northwest Research Station, United States Forest Service, United States Department of Agriculture (USDA), United States

***Correspondence:**
Magnus Ekström
Magnus.Ekstrom@slu.se

Specialty section:
This article was submitted to Forest Management, a section of the journal Frontiers in Forests and Global Change

Received: *25 August 2021*
Accepted: *19 November 2021*
Published: *15 December 2021*

Citation:
Ekström M and Nilsson M (2021) A Comparison of Model-Assisted Estimators, With and Without Data-Driven Transformations of Auxiliary Variables, With Application to Forest Inventory. Front. For. Glob. Change 4:764495. doi: 10.3389/ffgc.2021.764495

Forest information is requested at many levels and for many purposes. Sampling-based national forest inventories (NFIs) can provide reliable estimates on national and regional levels. By combining expensive field plot data with different sources of remotely sensed information, from airplanes and/or satellite platforms, the precision in estimators of forest variables can be improved. This paper focuses on the design-based model-assisted approach to using NFI data together with remotely sensed data to estimate forest variables for small areas, where the variables studied are total growing stock volume, volume of Norway spruce (*Picea abies*), and volume of broad-leaved trees. Remote sensing variables may be highly correlated with one another and some may have poor predictive ability for target forest variables, and therefore model selection and/or coefficient shrinkage may be appropriate to improve the efficiency of model-assisted estimators of forest variables. For this purpose, one can use modern shrinkage estimators based on lasso, ridge, and elastic net regression methods. In a simulation study using real NFI data, Sentinel 2 remote-sensing data, and a national airborne laser scanning (ALS) campaign, we show that shrinkage estimators offer advantages over the (weighted) ordinary least-squares (OLS) estimator in a model-assisted setting. For example, for a sample size n of about 900 and with 72 auxiliary variables, the RMSE was up to 41% larger when based on OLS. We propose a data-driven method for finding suitable transformations of auxiliary variables, and show that it can improve estimators of forest variables. For example, when estimating volume of Norway spruce, using a smaller expert selection of auxiliary variables, transformations reduced the RMSE by up to 10%. The overall best results in terms of RMSE were obtained using shrinkage estimators and a larger set of 72 auxiliary variables. However, for this larger set of variables, the use of transformations yielded at most small improvements of RMSE, and at worst large increases of RMSE, except in combination with ridge and elastic net regression.

Keywords: model-assisted estimation, generalized regression estimators, data-driven transformations, lasso, ridge, elastic net, forest inventory, remote sensing

INTRODUCTION

Information about forests is needed for many purposes and at various geographical levels. Large area sampling-based national forest inventories (NFIs) provide reliable estimates of mean values or totals on a national and regional level (Tomppo et al., 2011; Fridman et al., 2014). These estimates are used, for example, to form national forest policies, sustainability assessment, and reporting to international conventions. However, terrestrial inventory systems such as NFIs are typically designed to provide reliable estimates on a national and regional scale and may not provide sufficiently precise estimates for small areas without including auxiliary information, for example remote sensing data (McRoberts et al., 2014).

The availability of airborne laser scanning (ALS) data, and spectral data from Sentinel 2 and Landsat 8 satellites that are freely available, offers new possibilities for NFIs to produce more precise statistical estimates than by using field data alone. In order to utilize the full potential of auxiliary remote sensing data for statistical estimates, comprehensive remote sensing data can be combined with sample-based field measurements utilizing sampling theory (Gregoire et al., 2011). An important category of sample-based estimators that can be used for this purpose are known as design-based model-assisted estimators (Särndal et al., 1992). Such estimators use models and auxiliary data to improve the efficiency, while maintaining design-based properties of asymptotic design-unbiasedness and consistency (Breidt and Opsomer, 2016). Thus, model-assisted estimators are asymptotically design-unbiased irrespective of whether the assigned model is correct or not, where design-unbiasedness means that the estimator is unbiased over repeated sampling of field data. In contrast, model-based estimators, which do not utilize the sampling design for the inference, do not share these desirable properties (Kangas et al., 2016; Ståhl et al., 2016). When models are correctly assigned, model-based estimators can be very efficient, but model misspecifications easily result in severely biased estimators (Chambers et al., 2006).

The range of prediction techniques that can be used in a model-assisted estimator has dramatically increased during the last couple of decades. The main reason for this is the rapid development in the field of statistical learning and its very close cousin machine learning (Hastie et al., 2009, 2015; Berk, 2016). Breidt and Opsomer (2016) provide a review of such techniques in a model-assisted context. With a machine learning or statistical learning perspective, model-assisted methods are judged on their ability to produce precise estimates rather than on their ability to build interpretable models (McConville et al., 2020).

The model-assisted framework has gained an increasing popularity in forest inventory, and various prediction techniques have been utilized within this framework. Breidt et al. (2005) considered penalized spline regression together with auxiliary information such as GIS data. Opsomer et al. (2007) applied generalized additive models (GAMs), using three sources of auxiliary data, digital elevation models, Landsat TM imagery, and spatial coordinates. Baffetta et al. (2009, 2010) developed an estimator using *k*-nearest neighbor regression, and used Landsat 7 ETM+ imagery as auxiliary data. Chirici et al. (2016) compared the performance of *k*-nearest neighbor regression with linear regression, using auxiliary ALS based metrics. Kangas et al. (2016) considered three different predictions techniques, linear regression (where no transformations were carried out to linearize the relationship), GAM regression, and kernel regression, and used ALS data as auxiliary data. Moser et al. (2017) used non-linear regression and auxiliary ALS data, and explored variable selection techniques based on genetic algorithms and random forests. McConville et al. (2017) considered various lasso regression methods, using auxiliary variables from a national land cover database and Landsat 5 TM imagery, and comparisons were made with other predictions techniques such as linear regression and ridge regression. Further studies on lasso regression and its close cousins ridge regression and elastic net regression were made in McConville et al. (2020), using auxiliary data from Landsat imagery, forest maps, and a digital elevation model, and comparisons were made with standard prediction techniques, including linear regression (for continuous target variables) and logistic regression (for categorical target variables).

Remote sensing data or data that originates from remotely sensed data are used as auxiliary data in many forest inventory applications. This often means that the auxiliary data are known for the entire finite population under consideration, and that the number of potential auxiliary variables is large. As in Moser et al. (2017), methods for variable selection can be used for selecting a "best" set of auxiliary variables. Ridge, lasso, and elastic net regression shrink coefficient estimates toward zero, relative to least-squares estimates in a standard multiple linear regression. In the case of lasso and elastic net, coefficient estimates can be forced to be exactly zero. Consequently, these methods can also perform variable selection.

In this paper, we consider ridge, lasso, and elastic net regression in a model-assisted framework. Since the relationship between the target variable y and an auxiliary variable x can be non-linear, transformations of x may be needed. The key step is the identification of an appropriate transformation. In many applications, the form of transformation is suggested by prior experience. Unfortunately, in many cases, prior knowledge or theory may not suggest a suitable transformation to be used. In such situations, it would be convenient to determine the transformation adaptively, using a data-driven method for selecting appropriate transformations. This is especially useful when the number of auxiliary variables is large. For this reason, we suggest and investigate the performance of a data-driven method for finding suitable transformations in a model-assisted framework, where the method used is based on fractional polynomials (Royston and Altman, 1994).

The objective of this study was to evaluate ridge, lasso, and elastic net regression for prediction of volume per hectare of total growing stock, Norway spruce (*Picea abies*), and broad-leaved trees in a model-assisted setting, with or without data-driven transformations of auxiliary variables. The evaluation includes comparisons with the most well-known model-assisted estimator, the generalized regression estimator based on a multiple regression model, and is based on Monte Carlo simulations using real data, from the Swedish NFI, Sentinel-2,

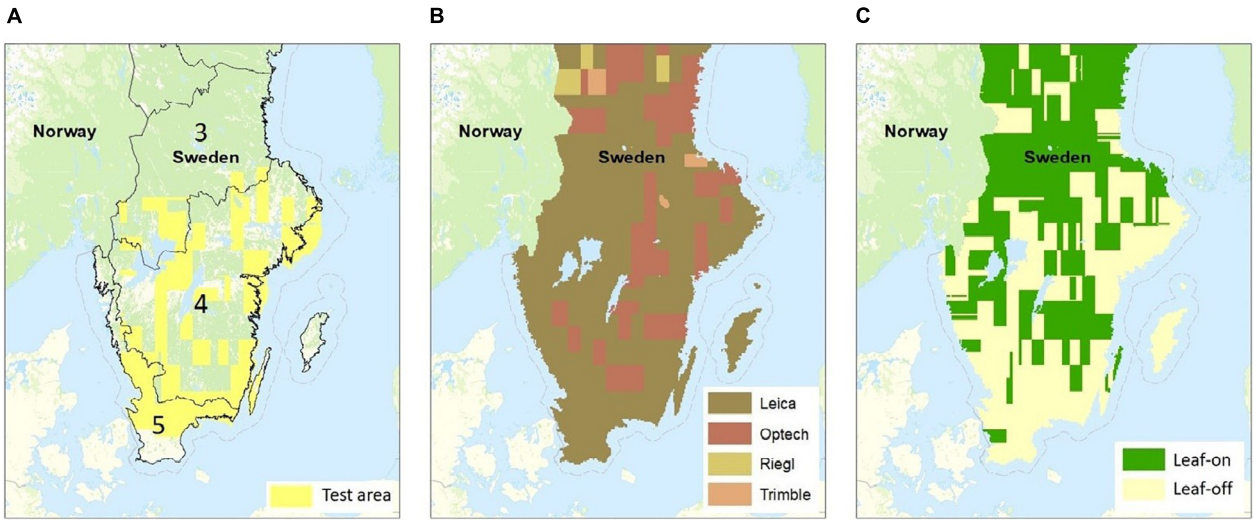


FIGURE 1 | Test area **(A)**, scanner brand **(B)**, and scanning season **(C)**. Three of the strata used in the Swedish NFI are shown in **(A)** (strata 3, 4, and 5). These strata roughly correspond to the vegetation zones: (3) southern-middle boreal; (4) mainly hemiboreal; and (5) temperate. Copyright Lantmäteriet.

and a national laser scanning campaign. Also, an expert's *a priori* selection of a smaller set of auxiliary variables is compared to using a full set of variables. The influence of outliers is discussed.

MATERIALS AND METHODS

Data

Test Area

In this study, we used a combination of data from a national ALS campaign, Sentinel 2, and the Swedish NFI to estimate volume per hectare of total growing stock, Norway spruce, and broad-leaved trees. Our test area is in southern Sweden and covers an area of approximately 6.0 million ha for which Sentinel 2 images and Leica ALS data registered during leaf-off conditions were available (**Figure 1**). The test area was restricted to areas mapped as land in the Swedish National Land Cover Database (NMD; Naturvårdsverket, 2020), except buildings (class 51 in NMD). Coniferous forest dominates the landscape within the test area, and the proportion of tree species are 28, 47, and 25% for Scots pine (*Pinus sylvestris*), Norway spruce (*Picea abies*), and broad-leaved trees, respectively, according to the Swedish NFI.

National Forest Inventory Data

The Swedish NFI provides information about forests for regional, national and international policy, planning, and reporting (Fridman et al., 2014). It has been operating since 1923 and at present more than 200 variables are recorded. The NFI covers all forests in Sweden (55–69°N) and the design includes both geographical stratification and clustering of sample plots into square-formed tracts with a side length that varies from 300 to 1,800 m among regions. There are two independent samples, one permanent and one temporary, where trees are measured on concentric sample plots with different radii depending on tree diameter at breast height (Fridman et al., 2014). On both temporary and permanent plots, trees with a diameter less than 4 cm are measured on two 1 m radius plots, and trees with a diameter between 4 and 10 cm are measured on a 3.5 m radius plot (**Figure 2**). If the diameter is 10 cm or more, the trees are measured on plots with 7 m or 10 m radius for temporary and permanent plots, respectively. Sample plots located on boundaries between forest stands or different land use classes are split and each part is described separately.

The NFI began positioning sample plots using GPS receivers in 1996. As of 2021, Garmin GPSMAP 64 receivers are used for the positioning that give a horizontal positional accuracy of approximately 5–10 m.

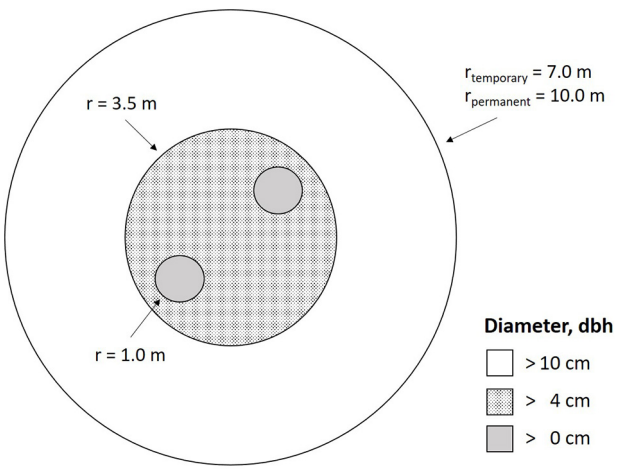


FIGURE 2 | Sample plot design used in the Swedish NFI.

In this study, we used NFI data from 2012 to 2016. Split plots were merged and volume per ha of total growing stock, Norway spruce, and broad-leaved trees were calculated for the merged plots (the rest of the growing stock volume was mainly Scots pine). In total, there were 9008 NFI plots within the test area, located in three different geographic strata (**Table 1**).

Airborne Laser Scanning Data

The first national ALS campaign in Sweden started in 2009 and ended in 2019. During the campaign, the National Mapping Agency (Lantmäteriet) collected data from flying heights between 1,700 and 2,300 m and with a point density of 0.5–1.0 pulses/m^2. A maximum scanning angle of 20° from nadir with a 20% overlap between adjacent scanning strips was used. For practical reasons, the campaign was divided into 397 blocks with a normal size of 25 km by 50 km. A block was always scanned using one scanner, but the scanner used varied between blocks. In total, 13 different scanners from Leica, Optech, Riegl and Trimble were used. As mentioned above, the study was restricted to areas where ALS data had been acquired with Leica scanners during leaf-off conditions (**Figure 1A**). All blocks within the test area were laser scanned between 2009 and 2013.

A national DEM (2 m × 2 m grid cell size), derived from the national ALS dataset by the National Mapping Agency, was used to calculate height above ground (normalized height) for all returns. A set of ALS metrics were calculated for each NFI plot using CloudMetrics (McGaughey, 2020) and used together with Sentinel 2 spectral data as auxiliary variables (**Table 2**).

Satellite Data

A mosaic of Sentinel-2 data from 2015 to 2017 with top-of-the-atmosphere (TOA) reflectance from bands 4, 5, 7, 8, 8a, 11, and 12 were used. About 95% of the test area was covered by images registered on May 27 and July 6, 2017 (**Table 3**). Additional images from 2015 to 2016 were used to cover the remaining parts of the test area, resulting in an almost cloud free mosaic. All image bands were resampled to 12.5 × 12.5 m pixel size and spectral data from all seven bands were extracted for the NFI plots using nearest neighbor interpolation. Sentinel-2 data were missing for 208 of the 9008 NFI plots due to clouds or cloud shadows. For these plots, spectral values were imputed based on all ALS metrics (**Table 2**), the sum of all daily mean temperature values exceeding 5° C° (Tsum), altitude, and plot coordinates (x and y) using the knnImputation function (k = 3) in the R package DMwR (Torgo, 2010).

TABLE 1 | Mean volume per hectare of total growing stock, Norway spruce, and broad-leaved trees, and number of plots by stratum.

Stratum	Volume (m^3/ha)			No. plots
	All species	Norway spruce	Broad-leaved trees	
3	131 (143)	63 (115)	20 (41)	819
4	114 (140)	52 (98)	24 (61)	5,692
5	110 (148)	51 (114)	45 (96)	2,497
Total	114 (142)	52 (105)	29 (71)	9,008

Standard deviations are given within parentheses.

TABLE 2 | Auxiliary variables used in the study.

Variable	Description
x, y	Plot coordinates in SWEREF 99 TM
Altitude	Height above sea level (m)
TSUM	Sum of all daily mean temperature values exceeding 5 C
N	Total number of laser returns
N_{150}	Total number of laser returns above 1.5 m
N_{mean}	Total number of laser returns above mean
N_{mode}	Total number of laser returns above mode
N_{First}	Total number of first laser returns
$N_{First,150}$	Total number of first laser returns above 1.5 m
$N_{First,mean}$	Total number of first laser returns above mean
$N_{First,mode}$	Total number of first laser returns above mode
$ReturnCount_i$	Number of first, second, …, fifth laser returns above 1.5 m
Min, Max, Mean, Mode	Min, max, mean and mode for all laser returns above 1.5 m
Stddev[a], CV, IQ, Skewness, Kurtosis	Standard deviation, coefficient of variation (CV), interquartile distance, skewness and kurtosis for all laser returns above 1.5 m
P_i	The *i*th height percentile for laser returns above 1.5 m, *i*= 1, 5, 10, 20, …, 90[a], 95[b], 99
CRR	Canopy relief ratio [(Mean–Min)/(Max–Min)]
Q_{Mean}, C_{Mean}	Quadratic mean and cubic mean for all laser returns above 1.5 m
Prop[b]	Proportion of all laser returns above 1.5 m
$Prop_{Mean}$	Proportion of all laser returns above mean
$Prop_{Mode}$	Proportion of all laser returns above mode
$Prop_{First}$	Proportion of first laser returns above 1.5 m
$Prop_{First,Mean}$	Proportion of first laser returns above mean
$Prop_{First,Mode}$	Proportion of first laser returns above mode
$Prop_{All}$	Number of returns above 1.5 m/number of first returns * 100
$Prop_{All,Mean}$	Number of returns above mean/number of first returns * 100
$Prop_{All,Mode}$	Number of returns above mode/number of first returns * 100
AAD	Average of the absolute deviations of laser returns from the overall mean.
MAD_{Median}	Median of the absolute deviations of laser returns from the overall median
MAD_{Mode}	Median of the absolute deviations of laser returns from the overall mode
L_1, L_2, L_3, L_4	L-moments (Hosking, 1990)
L_{CV}, $L_{skewness}$, $L_{kurtosis}$	L-moment ratios corresponding to coefficient of variation, skewness, and kurtosis
$P_{90}Vr$[a]	The 90th height percentile * Prop. of all returns above 1.5 m
$Band_i$[b]	Sentinel 2, band *i*, *i*= 4, 5, 7, 8, 8a, 11, and 12

[a] *Included in the expert's selection of auxiliary variables for estimation of volume of all tree species.*
[b] *Included in the expert's selection of auxiliary variables for estimation volume of Spruce and volume of broad-leaved trees.*

Final Auxiliary Data

Three different datasets were defined from the variables in **Table 2**. The first dataset consisted of all 72 variables in the table

TABLE 3 | Registration dates for Sentinel-2 images used in the study and the area covered at each registration date.

Registration date	Area cover in image mosaic, ha
August 19, 2015	20,600
June 14, 2016	167,600
July 21, 2016	23,800
May 23, 2017	37,200
May 27, 2017	3,947,100
July 6, 2017	1,736,500
August 11, 2017	22,500

and will be referred to as "all available auxiliary variables." The two other datasets were subsets of the variables in **Table 2**, and will be referred to as "expert's selections of auxiliary variables." The first subset was used to estimate total growing stock volume and included 90th (P_{90}) ALS height percentile for all laser returns above 1.5 m, proportion of all laser returns above 1.5 m multiplied by P_{90} (P_{90}Prop), and standard deviation for all laser returns above 1.5 m (Stddev). These variables were chosen because they previously were used to predict the total growing stock volume in the production of a nationwide raster database of forest variables using data from the first national ALS campaign (Nilsson et al., 2017). The second subset was used to estimate volume for Norway spruce and broad-leaved trees and included 95th height percentile for all laser returns above 1.5 m (P_{95}), the proportion of all laser returns above 1.5 m (Prop), and Sentinel-2 bands 4, 5, 7, 8, 8a, 11, and 12. The metrics were selected based on experiences from an ongoing project with the aim to predict standing volume by tree species from a combination of ALS metrics and Sentinel 2 data.

A correlation matrix was calculated for the 72 auxiliary variables in **Table 2**, containing 2556 unique correlation coefficients. The absolute values of these were larger than 0.5 in 1209 cases. In 212 cases they were larger than 0.9, and in 39 cases larger than 0.99. The largest absolute correlation coefficient between growing stock volume and an auxiliary variable was 0.75. For volume of Norway spruce and volume of broad-leaved trees, the corresponding values were 0.55 and 0.35, respectively.

Methods

To construct estimators of forest variables, the area of interest was tessellated into a finite number of population units, labeled by $\{1, 2, ..., N\}$, where the set was denoted by U. In our setting, a square tessellation was used, given by the 12.5 × 12.5 m raster cells in the wall-to-wall auxiliary data. The objective was to estimate the population mean, $\overline{Y} = N^{-1} \sum_{i \in U} y_i$, where y_i denotes value of the target forest variable for the ith unit.

A sample s of units is selected with a view to obtain information about the whole population. In large-area surveys like NFIs and vegetation monitoring programs, samples are usually taken using complex probability sampling designs that include, for example, geographical stratification (Ekström et al., 2018). In these designs, each population unit i typically has a non-zero probability π_i of getting included in the sample.

Design-based estimators incorporate sample design characteristics into their formulae, typically to achieve desirable properties such as unbiasedness. The Horvitz and Thompson (1952) estimator (HT) of the population mean, $\overline{Y}$, incorporates design information through inverse-probability weighting,

$$\widehat{\overline{Y}} = \frac{1}{N} \sum_{i \in s} \frac{y_i}{\pi_i}. \tag{1}$$

The HT is a design-unbiased estimator, which means that the mean of the estimator, taken over all possible samples under the sampling design, is equal to $\overline{Y}$. The estimator of the variance of $\widehat{\overline{Y}}$ in (1), suggested by Horvitz and Thompson (1952), is

$$\hat{V} = \frac{1}{N^2} \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}, \tag{2}$$

where π_{ij} is the probability that both units i and j are included in the sample s, and $\pi_{ii} = \pi_i$ for all i.

Model-Assisted Estimators

One possible approach to improving the efficiency of estimators is to incorporate auxiliary information, and model-assisted estimation is a form of design-based estimation that incorporates both design information (through the inclusion probabilities π_i) and auxiliary information (through a model). Many super-population models for this purpose can be written in the form

$$y_i = \mu(\boldsymbol{x}_i) + \epsilon_i, \tag{3}$$

with random, zero-mean ϵ_i, and a vector of auxiliary variables for unit i, $\boldsymbol{x}_i = (1, x_{i1}, ..., x_{ip})$. The predictor function $\mu(\cdot)$ is typically unknown, but can be estimated using the sample data. Denoting the estimated predictor by $\widehat{\mu}(\cdot)$, a general class of model-assisted estimators of the population mean, known as generalized regression estimators (GREG), can be defined as

$$\widehat{\overline{Y}} = \frac{1}{N} \sum_{i \in U} \widehat{\mu}(\boldsymbol{x}_i) + \frac{1}{N} \sum_{i \in s} \frac{y_i - \widehat{\mu}(\boldsymbol{x}_i)}{\pi_i}. \tag{4}$$

It should be noted that the estimator (4) depends on the sampling design, the form of the model, and the method used for estimating the predictor function $\mu(\cdot)$. The estimator (4) consists of two parts, the mean of the predicted values over the population and the design bias adjustment consisting of inverse probability-weighted "residuals" $(y_i - \widehat{\mu}(\boldsymbol{x}_i))$. This adjustment term protects against model misspecification, and makes the estimator approximately design-unbiased for many commonly used prediction methods (see, e.g., Breidt and Opsomer (2016) and the references therein).

To estimate the variance for (4) we use a common variance estimator approach based on (2) but replacing the "raw" y_i values with the "residuals" $(y_i - \widehat{\mu}(\boldsymbol{x}_i))$ (cf. Breidt and Opsomer, 2016). Provided that the residuals have smaller variation than the raw values, we can expect GREG to have a smaller variance than HT.

Under a multiple linear regression model with $\mu(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\beta}$, the parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ can be estimated using

weighted least-squares. This approach gives the predictor $\widehat{\mu}(\boldsymbol{x}) = \boldsymbol{x}^T\widehat{\boldsymbol{\beta}}$, where

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i\in s} \frac{\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2}{\pi_i} = \left(\sum_{i\in s} \frac{\boldsymbol{x}_i\boldsymbol{x}_i^T}{\pi_i}\right)^{-1} \sum_{i\in s} \frac{\boldsymbol{x}_i y_i}{\pi_i},$$

where arg min means the value of $\boldsymbol{\beta}$ which minimizes the sum of design-weighted squared residuals. With $\widehat{\mu}(\boldsymbol{x}_i) = \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}$ plugged into (4), we refer to (4) as the regression estimator (REG).

For our analyses, $\widehat{\boldsymbol{\beta}}$ is computed using the glm function in R (R Core Team, 2020). If some auxiliary variables are perfectly or nearly perfectly collinear, the glm function automatically excludes at least one of them and sets the corresponding coefficients to NA (not available). For this reason, we investigate the following two variants for handling this problem:

(i) calculate pairwise correlations among the variables in the sample and, among each pair of variables correlated above a given threshold, exclude the variable least correlated with the target variable;
(ii) if a coefficient is NA, then simply set it to 0.

If, for example, the second variant is used, we refer to (4) as REG^{ii}. A benefit of the first variant is that it decreases the danger of multicollinearity, but as argued in Vaughan and Berry (2005), multicollinearity is "not quite as damning" when linear modeling is used for prediction rather than explanation. That is, in case of (severe) multicollinearity, coefficient estimates and their standard errors can become (very) sensitive to small changes in the model, but this usually has little effect on the prediction capability of the model. However, if the fitted model is used to predict values for new data, and the pattern of multicollinearity in the new data differs from that in the data that was fitted, this may introduce large errors in the predictions (Chatterjee et al., 2012).

Another possibility is to estimate the parameter vector $\boldsymbol{\beta}$ using penalized weighted least squares. Elastic net regression (Zou and Hastie, 2005; McConville et al., 2020), introduced as compromise between lasso and ridge regression, is an approach that uses a penalty. Here, the parameter vector is estimated by

$$\widehat{\boldsymbol{\beta}}_\alpha = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i\in s} \frac{\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2}{\pi_i} + \lambda \sum_{j=1}^{p} \{(1-\alpha)\beta_j^2 + \alpha|\beta_j|\} \right\}, \tag{5}$$

where $0 \le \alpha \le 1$. When $\alpha = 0$, elastic net regression becomes ridge regression, and when $\alpha = 1$ it becomes lasso regression. Ridge regression tends to give similar coefficient values to highly correlated auxiliary variables, whereas lasso regression tend to give quite different coefficient values to highly correlated variables. Unlike ridge regression, lasso regression performs variable selection by forcing some of the coefficient estimates to be exactly equal to zero (this happens if the "tuning parameter" λ is sufficiently large). Elastic net regression, with α equal to a value between 0 and 1, shrinks together the coefficients of correlated auxiliary variables like ridge, and performs variable selection like the lasso (Zou and Hastie, 2005). Thus, the α value in (5) is the "mixing proportion" that toggles between a pure lasso penalty (when $\alpha = 1$) and a pure ridge penalty ($\alpha = 0$). The parameter λ controls the total amount of penalization. Both penalties shrink the coefficient estimates toward zero, relative to the usual (weighted) least-squares estimates, and the more so the larger λ is. As λ increases, the shrinkage of the coefficient estimates reduces the variance of the predictions, at the expense of an increase in bias (James et al., 2021). Selecting a good value for λ is therefore critical for finding a good balance between variance and bias, and cross-validation is commonly used for this purpose.

With the estimator function $\widehat{\mu}(\boldsymbol{x}_i)$ set to the generalized penalized estimator $\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_\alpha$, we refer to (4) as RIDGE, ELNET, and LASSO, for $\alpha = 0$, 0.5, and 1, respectively. These three are available through the R package mase (McConville et al., 2018), which uses cross-validation to choose the tuning parameter λ. If there are issues with multicollinearity, McConville et al. (2020) recommend using RIDGE or ELNET rather than REG or LASSO.

In our study and for a given set of auxiliary variables, the parameter vector $\boldsymbol{\beta}$ is estimated using all data from a sample s. In **Supplementary Material**, results are presented also for the case where outliers in the sample s are removed before $\boldsymbol{\beta}$ is estimated. The identified outliers are those where field measured tree height and the 95th height percentile in the ALS data deviate more than 7 m.

Data-Driven Choices of Transformations

A model with $\mu(\boldsymbol{x}_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ assumes a linear relationship between the expected value of the target variable y_i in (3) and each auxiliary variable (when the other auxiliary variables are held fixed). If linearity fails to hold, it is sometimes possible to transform the auxiliary variables in the model to improve the linearity. Examples of a non-linear transformation of variable x_{ij} are the square root or the reciprocal of x_{ij}. Suitable transformations can be found through studies of residual plots, but this is tedious work when the number of variables is large. For this reason, we investigate the performance of a data-driven method for finding suitable transformations. The method is based on fractional polynomials (FPs; Royston and Altman, 1994). FP is an approach that uses a function selection procedure to check whether a non-linear function fits the data significantly better than a linear function. We use the level of significance 5% for the function selection. To reduce the computational burden, the function selection is done for one auxiliary variable at a time.

The class of FP functions is an extension of power transformations of a variable, and in this study the attention is restricted to FPs of the first degree. That is, the powers are selected from the collection $\{-3, -2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, where 0 denotes the log transformation, using the sample data and the fp and mfp functions in the R package mfp (Ambler and Banner, 2015). FPs are defined only for positive auxiliary variables, but real data may contain non-positive observations. Therefore, at population level, if non-positive values are encountered (or the range of values of the auxiliary variables is unreasonably large), the auxiliary variables are shifted (and rescaled). The method for doing this is adopted from the mfp algorithm (Sauerbrei et al., 2006; Sabanés Bové and Held, 2011).

In our study, outliers in the sample data are not used in the selection procedure of transformations. Again, the identified

outliers are those where field measured tree height and the 95th height percentile in the ALS data deviate more than 7 meters. (In **Supplementary Material**, results are presented also for the case where transformations are selected based on all sample data).

Evaluation of the Estimators

The performances of estimators were compared using Monte Carlo simulations. The population units were defined by the 9008 pixels that we matched with the corresponding plots given in **Table 1**. Three strata were defined according to **Table 1**, and Monte Carlo simulations were implemented with a stratified simple random sampling design. With this design, a simple random sample without replacement is drawn from each strata, the drawings being made independently in different strata. In comparison with the Swedish NFI, the main difference is that we ignored that plots are grouped into tracts. The number of sampled units in each stratum was proportional to the size of the stratum. Two sample sizes were considered in the simulations, $n = 901$ and $n = 2703$. In the former case, the sample sizes in the three NFI strata within the study area (**Figure 1A**) were 82, 569, and 250, and in the latter case, 246, 1708, and 749, respectively. For each forest variable to be estimated and for each estimator considered, we used the same set of samples of size $n = 901$ or $n = 2703$. In total, $m = 10000$ samples of each sample size were drawn.

The estimators of the population mean were evaluated with respect to root mean square error (RMSE), standard deviation (SD; also commonly referred to as the standard error), and bias, obtained with the $m = 10000$ repeated samples under the aforementioned stratified simple random sampling design. With $\widehat{\overline{Y}}$ denoting an estimator of a population mean $\overline{Y}$, and $\widehat{\overline{Y}}_i$ denoting an estimate based on the ith sample, these quantities were computed as

$$\widehat{\text{bias}}(\widehat{\overline{Y}}) = \frac{1}{m}\sum_{i=1}^{m}\widehat{\overline{Y}}_i - \text{true value},$$

$$\widehat{\text{SD}}(\widehat{\overline{Y}}) = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(\widehat{\overline{Y}}_i - \frac{1}{m}\sum_{j=1}^{m}\widehat{\overline{Y}}_j\right)^2},$$

and

$$\widehat{\text{RMSE}}(\widehat{\overline{Y}}) = \sqrt{\widehat{\text{SD}}(\widehat{\overline{Y}})^2 + \widehat{\text{bias}}(\widehat{\overline{Y}})^2}.$$

For the ease of comparisons across variables, all values of bias, SD, and RMSE are presented as percentages of $\overline{Y}$. That is, as

$$\widehat{\text{bias}}_{\%} = 100\frac{\widehat{\text{bias}}(\widehat{\overline{Y}})}{\overline{Y}}, \widehat{\text{SD}}_{\%} = 100\frac{\widehat{\text{SD}}(\widehat{\overline{Y}})}{\overline{Y}},$$

$$\text{and} \quad \widehat{\text{RMSE}}_{\%} = 100\frac{\widehat{\text{RMSE}}(\widehat{\overline{Y}})}{\overline{Y}}.$$

Likewise, let $\hat{V}_i$ denote an estimate of the variance of $\widehat{\overline{Y}}$ based on the ith sample. For example, in the case of the HT estimator, $\hat{V}_i$ is computed using formula (2). Then

$$\widehat{\text{SD}}_{\%,\,i} = 100\frac{\sqrt{\hat{V}_i}}{\overline{Y}}$$

is the value of an estimated standard deviation, using data from the ith sample, and presented as a percentage of the corresponding population mean. Let

$$ave\left(\widehat{\text{SD}}_{\%,i}\right) = \frac{1}{m}\sum_{i=1}^{m}\widehat{\text{SD}}_{\%,i},$$

where "*ave*" denotes average. If $ave\left(\widehat{\text{SD}}_{\%,i}\right)$ is approximately equal to $\widehat{\text{SD}}_{\%}$, then this suggests that the estimator of the standard deviation of $\widehat{\overline{Y}}$ is nearly unbiased.

For comparing the RMSE of one estimator (with auxiliary variables in their original scale) to the RMSE of another estimator (with power transformed auxiliary variables), the basic bootstrap confidence interval (e.g., Davison and Hinkley, 1997) for their difference is applied. Let $\widehat{\overline{Y}}_{1,i}$ and $\widehat{\overline{Y}}_{2,i}$ denote the two estimates based on sample i, where the first is based on auxiliary variables in the original scale while the other uses power transformed auxiliary variables. A bootstrap sample $\left\{(\widehat{\overline{Y}}^*_{1,i}, \widehat{\overline{Y}}^*_{2,i})\right\}_{i=1}^{m}$ is taken as a random sample with replacement from $\left\{(\widehat{\overline{Y}}_{1,i}, \widehat{\overline{Y}}_{2,i})\right\}_{i=1}^{m}$. Based on the bootstrap sample, bootstrap replicates of the two estimated RMSEs are computed. Based on $R = 9999$ such bootstrap replicates, a basic bootstrap 95% confidence interval for the difference of the two RMSEs is computed using the boot.ci function in the R package boot (Davison and Hinkley, 1997). In these computations, all RMSEs are expressed as percentages of the corresponding population means. A 95% confidence interval that does not cover zero means that the use of power transformed auxiliary variables significantly changes the efficiency of the estimator at the 5% significance level. If the interval contains only positive values, the conclusion is that the transformations significantly improves the efficiency of the estimator at the 5% level. Thus, as in, for example, Samuels et al. (2012), if we find significant evidence for a change, our conclusion can be directional. Some authors prefer not to draw a directional conclusion in these cases (Samuels et al., 2012).

RESULTS

The results for HT are presented in **Table 4**, i.e., the results for the case where no auxiliary data were used in the estimation. Since the HT estimator is unbiased, as expected, the values of (estimated) bias in **Table 4** were close to zero. In addition, and also as expected, the values of $ave\left(\widehat{\text{SD}}_{\%,\,i}\right)$ were all close to the corresponding values of $\widehat{\text{SD}}_{\%}$, suggesting that the estimator of the standard deviation of $\widehat{\overline{Y}}$ [i.e., the square root of the variance estimator (2)] is nearly unbiased.

When comparing the RMSEs in **Table 4** with the RMSEs in **Table 5** for the various model-assisted estimators based on an expert selection of auxiliary variables, notice that the use

TABLE 4 | Monte Carlo results for the Horvitz and Thompson estimator (HT).

Forest variable	$\widehat{\text{bias}}_{\%}$	$\widehat{\text{SD}}_{\%}$	*ave* $\left(\widehat{\text{SD}}_{\%,\,i}\right)$	$\widehat{\text{RMSE}}_{\%}$
(a) $n = 2703$				
Volume (m^3/ha) of total growing stock	0.020	2.017	2.002	2.017
Volume (m^3/ha) of Norway spruce	0.011	3.223	3.206	3.223
Volume (m^3/ha) of broad-leaved trees	0.037	3.904	3.863	3.904
(b) $n = 901$				
Volume (m^3/ha) of total growing stock	0.097	3.958	3.935	3.959
Volume (m^3/ha) of Norway spruce	0.118	6.311	6.300	6.313
Volume (m^3/ha) of broad-leaved trees	0.143	7.615	7.577	7.618

Estimated values of bias, SD, and RMSE ($\widehat{\text{bias}}_{\%}$, $\widehat{\text{SD}}_{\%}$, and $\widehat{\text{RMSE}}_{\%}$) are given as percentages of the corresponding population mean, and are based on m = 10000 stratified samples of size n = 2703 or 901 from the population. For each sample, an estimate of standard deviation of the HT was computed, and ave $\left(\widehat{\text{SD}}_{\%,\,i}\right)$ is the average of these estimates.

of assisting models and auxiliary information improved the efficiency of estimation. For volume of total growing stock, the reduction in RMSE was larger than 40% for each model-assisted estimator used and for both sample sizes considered. Moreover, the confidence intervals in **Table 5** show that the use of data-driven choices of transformations of auxiliary variables significantly improved the RMSEs of the estimators. However, the improvements were quite small, except for Norway spruce, with reductions of RMSE by 7.7–10.0%. The performances of REG, LASSO, RIDGE, and ELNET were very similar.

The results when all 72 available auxiliary variables in **Table 2** were used are shown in **Table 6**. For REG^{i} and the larger sample size, results are presented for the case where we excluded auxiliary variables with correlations above thresholds $\pm$ 0.90 and $\pm$ 0.95. When we tried $\pm$ 0.99 as threshold, then for many of the samples not all model coefficients could be estimated. For many samples of the smaller size ($n = 901$), this was the case even if the threshold was as low as $\pm$ 0.70. Therefore, no results for REG^{i} were presented for the smaller sample size.

For the larger sample size ($n = 2703$), the estimators based on auxiliary data in their original scale in **Table 6** had lower RMSEs than the corresponding estimators based on the smaller selection of auxiliary variables in **Table 5**. For example, for Norway spruce the RMSEs were about 15% lower and for broad-leaved trees about 7% lower, except for RIDGE where the gain was somewhat smaller. For the smaller sample size ($n = 901$) and LASSO, RIDGE, and ELNET, the corresponding reductions of RMSEs were 11% or larger for Norway spruce. For total growing stock and broad-leaved trees, the reduction was only 2 and 4%, respectively, for RIDGE, and even smaller than that for LASSO and ELNET. For the smaller sample size, REG^{ii} based on all the 72 auxiliary variables had RMSEs 22–34% *larger* than when using REG and a small expert selection of variables. For volume of broad-leaved trees, its performance was worse than the Horvitz-Thompson estimator.

The results for the larger sample size in **Table 6** show that the estimators based on all available auxiliary variables in their original scale had about the same performance in terms of RMSE. The corresponding results for the smaller sample size show that LASSO, RIDGE, and ELNET were very close in terms of RMSE, and that they performed much better than REG^{ii}. More precisely, the latter estimator had RMSEs 34–41% larger than those for LASSO, RIDGE, and ELNET.

When for example estimating total growing stock volume (both sample sizes) or volume of Norway spruce (the larger sample size), the confidence intervals in **Table 6** show that the use of data-driven choices of transformations of auxiliary variables significantly improved the RMSEs of LASSO, RIDGE, and ELNET. Although there were significant improvements when using transformations, the improvements in **Table 6** were never larger than 5%. When estimating volume of broad-leaved trees using a large number of auxiliary variables, the data-driven method for selecting transformations did not perform well. For REG^{ii} and LASSO, the use of transformations sometimes resulted in extreme and unreasonable estimates of volume of broad-leaved trees, which in turn resulted in very large values of RMSE. This was also the case for the REG^{ii} estimator of total growing stock and volume of Norway spruce when using the smaller sample size. In comparison, RIDGE was quite robust against poor choices of transformations, and to a lesser degree, ELNET.

In **Tables 5**, **6**, each value of *ave* $\left(\widehat{\text{SD}}_{\%,\,i}\right)$ is smaller than the corresponding value of $\widehat{\text{SD}}_{\%}$. This implies that the estimated standard deviations, $\widehat{\text{SD}}_{\%,\,i}$, $i = 1, \ldots, n$, were somewhat too small, on average, which is quite typical in model-assisted estimation (cf. Kangas et al., 2016). As suggested by simulation results in McConville et al. (2020), it is better to estimate standard deviations (or variances) of model-assisted estimators by using a bootstrap method, especially as the number of explanatory variables grows. However, because of the additional computational burden generated by bootstrapping, we did not use this estimator in our study.

In summary for the larger sample size, when estimating total growing stock volume or volume of Norway spruce, the best results in terms of RMSE were obtained when using all available auxiliary variables. Here, for LASSO, RIDGE, and ELNET, the use of data-driven choices of transformations significantly improved the RMSEs, but the improvements were small. For volume of broad-leaved trees, LASSO, ELNET, and REG^{ii} based on all available auxiliary variables in their original scale produced the best results, and were slightly better than the corresponding REG^{i} (with threshold $\pm$ 0.95) and RIDGE estimators. Finally, the use of data-driven choices of transformations was most successful when estimating volume of Norway spruce, using an expert selection of auxiliary variables. Here, the transformations reduced the RMSEs by up to 10%.

In summary for the smaller sample size, when estimating total growing stock volume or volume of Norway spruce, LASSO, RIDGE, and ELNET, with or without the use of data-driven choices of transformations, performed the best and were close in terms of RMSE. For volume of broad-leaved trees, LASSO, RIDGE, and ELNET with auxiliary variables in their original scale showed the best results. For all target variables, REG^{ii} based on all available auxiliary variables in their original scale had 34–41% higher RMSEs than the corresponding LASSO, RIDGE, and ELNET estimators, and

TABLE 5 | Monte Carlo results for REG, LASSO, RIDGE, and ELNET, when based on an expert selection of auxiliary variables.

	Auxiliary variables in original scale				Power transformed auxiliary variables					
Estimator	$\widehat{\text{bias}}_{\%}$	$\widehat{\text{SD}}_{\%}$	$ave\left(\widehat{\text{SD}}_{\%,\,i}\right)$	$\widehat{\text{RMSE}}_{\%}$	$\widehat{\text{bias}}_{\%}$	$\widehat{\text{SD}}_{\%}$	$ave\left(\widehat{\text{SD}}_{\%,\,i}\right)$	$\widehat{\text{RMSE}}_{\%}$	**LCL**	**UCL**
(a) Volume (m^3/ha) of total growing stock; $n = 2703$										
REG	−0.005	1.167	1.155	1.167	0.000	1.160	1.148	1.160	0.004	0.010
LASSO	−0.005	1.167	1.155	1.167	−0.001	1.160	1.148	1.160	0.004	0.010
RIDGE	−0.002	1.186	1.173	1.186	0.001	1.177	1.165	1.177	0.006	0.012
ELNET	−0.005	1.167	1.155	1.167	0.000	1.160	1.148	1.160	0.004	0.010
(b) Volume (m^3/ha) of Norway spruce; $n = 2703$										
REG	0.009	2.637	2.621	2.637	0.086	2.375	2.359	2.376	0.242	0.277
LASSO	0.009	2.637	2.621	2.637	0.075	2.376	2.362	2.377	0.243	0.277
RIDGE	−0.010	2.644	2.630	2.644	0.029	2.381	2.374	2.381	0.248	0.279
ELNET	0.009	2.637	2.621	2.637	0.074	2.376	2.362	2.377	0.243	0.277
(c) Volume (m^3/ha) of broad-leaved trees; $n = 2703$										
REG	−0.016	3.515	3.462	3.515	−0.092	3.444	3.389	3.445	0.053	0.086
LASSO	−0.013	3.514	3.463	3.514	−0.099	3.444	3.391	3.446	0.053	0.084
RIDGE	0.000	3.515	3.465	3.515	−0.050	3.448	3.396	3.449	0.052	0.081
ELNET	−0.013	3.513	3.463	3.513	−0.098	3.444	3.391	3.445	0.053	0.084
(d) Volume (m^3/ha) of total growing stock; $n = 901$										
REG	0.034	2.281	2.262	2.281	0.046	2.275	2.252	2.275	0.000	0.012
LASSO	0.034	2.284	2.263	2.284	0.047	2.276	2.252	2.277	0.001	0.013
RIDGE	0.049	2.319	2.300	2.319	0.056	2.307	2.286	2.307	0.007	0.017
ELNET	0.035	2.283	2.263	2.283	0.048	2.275	2.252	2.276	0.002	0.013
(e) Volume (m^3/ha) of Norway spruce; $n = 901$										
REG	0.098	5.151	5.133	5.152	0.476	4.732	4.610	4.755	0.356	0.436
LASSO	0.092	5.153	5.136	5.153	0.385	4.729	4.630	4.745	0.372	0.445
RIDGE	0.047	5.154	5.154	5.155	0.269	4.692	4.652	4.700	0.424	0.485
ELNET	0.093	5.152	5.136	5.153	0.382	4.726	4.629	4.742	0.375	0.447
(f) Volume (m^3/ha) of broad-leaved trees; $n = 901$										
REG	0.032	6.924	6.766	6.924	−0.231	6.812	6.627	6.816	0.074	0.143
LASSO	0.050	6.909	6.772	6.909	−0.190	6.811	6.642	6.814	0.063	0.129
RIDGE	0.077	6.905	6.778	6.905	−0.121	6.807	6.662	6.809	0.068	0.126
ELNET	0.049	6.908	6.772	6.909	−0.192	6.806	6.642	6.808	0.068	0.133

Estimated values of bias, SD, and RMSE ($\widehat{\text{bias}}_{\%}$, $\widehat{\text{SD}}_{\%}$, and $\widehat{\text{RMSE}}_{\%}$) are given as percentages of the corresponding population mean, and are based on m = 10000 stratified samples of size n = 2703 or 901 from the population. For each sample, an estimate of SD was computed, and ave $\left(\widehat{\text{SD}}_{\%,\,i}\right)$ is the average of these estimates. The values of LCL and UCL denote the lower and upper confidence limits of the 95% confidence interval for the difference in RMSE between the estimators based on auxiliary variables in original scale and power transformed auxiliary variables, respectively. If the interval contains only positive values, it suggests that the use of power transformed auxiliary variables improves the efficiency of the estimator.

performed worse in terms of RMSE than using REG and an expert selection of variables. For REG^i it was often not possible to estimate the model coefficients. Data-driven choices of transformations reduced the RMSEs by about 8% for Norway spruce when using an expert selection of auxiliary variables. For all other cases, the transformations resulted in at best minor improvements of RMSE, and at worst very large increases of RMSE. Of the estimators considered, RIDGE, and to a lesser extent, ELNET, were found robust against poor choices of transformations.

Remark: In our population, 18% of the units (raster cells) had a height difference larger than 7 m between the field measured tree height and the 95th height percentile in the ALS data. We may consider these units as outliers, and we may ask ourselves: (*i*) Is it better to perform data-driven choices of transformations of auxiliary variables with these outliers present in the sample? (*ii*) Is it better to estimate the parameter vector $\boldsymbol{\beta}$ (after possible transformations of auxiliary variables) with these outliers present in the sample? In order to find out, we performed Monte Carlo simulations for each of the four possible combinations of answers

TABLE 6 | Monte Carlo results for REG^{i}, REG^{ii}, LASSO, RIDGE, and ELNET, when based on all auxiliary variables.

		Auxiliary variables in original scale				Power transformed auxiliary variables					
Estimator	**Threshold**	$\widehat{\text{bias}}_{\%}$	$\widehat{\text{SD}}_{\%}$	$ave\left(\widehat{\text{SD}}_{\%,\,i}\right)$	$\widehat{\text{RMSE}}_{\%}$	$\widehat{\text{bias}}_{\%}$	$\widehat{\text{SD}}_{\%}$	$ave\left(\widehat{\text{SD}}_{\%,\,i}\right)$	$\widehat{\text{RMSE}}_{\%}$	**LCL**	**UCL**
(a) Volume (m^3/ha) of total growing stock; $n = 2703$											
REG^{i}	± 0.95	−0.033	1.149	1.126	1.150	0.023	1.152	1.096	1.153	−0.030	0.043
REG^{i}	± 0.90	−0.026	1.153	1.134	1.154	0.002	1.183	1.119	1.183	−0.062	0.018
REG^{ii}		0.025	1.148	1.094	1.148	0.069	1.191	1.072	1.193	−0.103	0.053
LASSO		−0.028	1.143	1.112	1.144	0.023	1.116	1.089	1.116	0.021	0.033
RIDGE		−0.038	1.151	1.133	1.152	0.007	1.118	1.100	1.118	0.029	0.039
ELNET		−0.027	1.143	1.112	1.143	0.023	1.116	1.089	1.116	0.020	0.033
(b) Volume (m^3/ha) of Norway spruce; $n = 2703$											
REG^{i}	± 0.95	−0.045	2.236	2.215	2.236	0.156	2.241	2.162	2.246	−0.044	0.034
REG^{i}	± 0.90	−0.072	2.335	2.312	2.337	0.083	2.290	2.222	2.291	0.003	0.099
REG^{ii}		−0.007	2.229	2.153	2.229	0.148	2.297	2.074	2.301	−0.205	0.135
LASSO		−0.013	2.227	2.177	2.227	0.193	2.166	2.090	2.174	0.035	0.071
RIDGE		−0.067	2.316	2.294	2.317	0.123	2.214	2.186	2.218	0.085	0.114
ELNET		−.014	2.227	2.177	2.227	0.195	2.167	2.091	2.176	0.033	0.068
(c) Volume (m^3/ha) of broad-leaved trees; $n = 2703$											
REG^{i}	± 0.95	−0.107	3.315	3.171	3.317	8.338	25.026	3.191	26.378	−25.20	−20.75
REG^{i}	± 0.90	−0.074	3.387	3.262	3.388	9.292	25.019	3.242	26.688	−25.39	−21.07
REG^{ii}		−0.082	3.248	3.017	3.249	9.279	36.230	3.002	37.399	−36.37	−31.88
LASSO		−0.107	3.248	3.067	3.250	1.385	10.75	3.078	10.839	−8.715	−6.466
RIDGE		−0.047	3.306	3.189	3.306	−0.056	3.324	3.206	3.324	−0.036	−0.001
ELNET		−0.104	3.250	3.067	3.251	0.508	3.968	3.083	4.001	−0.833	−0.665
(d) Volume (m^3/ha) of total growing stock; $n = 901$											
REG^{ii}	0.035	3.061	2.055	3.062	4.015	195.081	2.011	195.123	−307.1	−100.8	
LASSO	−0.065	2.285	2.169	2.286	0.111	2.215	2.118	2.218	0.054	0.082	
RIDGE	−0.069	2.272	2.195	2.273	0.084	2.209	2.134	2.211	0.052	0.073	
ELNET	−0.070	2.278	2.170	2.279	0.108	2.212	2.119	2.214	0.051	0.078	
(e) Volume (m^3/ha) of Norway spruce; $n = 901$											
REG^{ii}	−0.057	6.309	4.074	6.309	1.571	30.957	3.91	30.997	−29.70	−19.65	
LASSO	−0.190	4.457	4.237	4.461	0.734	4.450	4.081	4.510	−0.099	0.005	
RIDGE	−0.170	4.548	4.439	4.551	0.547	4.417	4.224	4.451	0.066	0.135	
ELNET	−0.214	4.461	4.232	4.466	0.731	4.440	4.084	4.500	−0.079	0.011	
(f) Volume (m^3/ha) of broad-leaved trees; $n = 901$											
REG^{ii}	−0.508	9.136	5.594	9.150	119.412	5273.475	5.567	5274.827	−8776	−2318	
LASSO	−0.397	6.696	5.977	6.707	1.501	19.866	5.994	19.923	−15.27	−11.18	
RIDGE	−0.206	6.604	6.164	6.607	−0.285	6.667	6.233	6.673	−0.102	−0.029	
ELNET	−0.378	6.706	5.978	6.716	−0.068	7.500	6.011	7.500	−0.931	−0.629	

Estimated values of bias, SD, and RMSE ($\widehat{\text{bias}}_{\%}$, $\widehat{\text{SD}}_{\%}$, and $\widehat{\text{RMSE}}_{\%}$) are given as percentages of the corresponding population mean, and are based on m = 10000 stratified samples of size n = 2703 or 901 from the population. For each sample, an estimate of SD was computed, and ave ($\widehat{SD}_{\%,\,i}$) is the average of these estimates. The values of LCL and UCL denote the lower and upper confidence limits of the 95% confidence interval for the difference in RMSE between the estimators based on auxiliary variables in original scale and power transformed auxiliary variables, respectively. If the interval contains only positive values, it suggests that the use of power transformed auxiliary variables improves the efficiency of the estimator. In ***Table 6****, no results are presented for the REG^{i} estimator when n = 901. The reason is that for many of the samples, not all model coefficients could be estimated (not even if the threshold was as low as ± 0.70).*

to questions *i* and *ii* (No-No, Yes-No, Yes-Yes, or No-Yes), and for both the sample sizes, $n = 2703$ and $n = 901$. The case No-Yes is presented in **Tables 5**, **6**. Results for all other possible cases are given in **Supplementary Material**. For each sample size and in terms of RMSE, it turned out that it was generally better to remove the outliers in the sample prior to performing data-driven choices of transformations, but to estimate the parameter vector $\boldsymbol{\beta}$ without removing the outliers in the sample of auxiliary variable

data values (where variables may have been transformed before the estimation is performed). For auxiliary variables in their original scale, the following increases of RMSEs were obtained if outliers were removed before the parameter vector $\boldsymbol{\beta}$ was estimated: (a) 0.5–1.7% when the models were based on an expert selection of auxiliary variables; (b) 0.9–6.0% when using all available auxiliary variables and $n = 2703$; and (c) 1.4–68% when using all available auxiliary variables and $n = 901$. In (c), the increases of RMSEs were in the range 35–68% for REG^{ii}, but less than 9% for LASSO, RIDGE, and ELNET.

DISCUSSION

In this paper, we have compared the performances of the Horvitz-Thompson estimator and several model-assisted estimators, using Monte Carlo simulations and real data, from the Swedish NFI, Sentinel-2, and a national laser scanning campaign. The model-assisted estimators were based either on modern prediction techniques (lasso, ridge, and elastic net regression), or on a traditional working model of multiple regression.

When based on an expert selection of a rather small set of auxiliary variables, the performances of the model-assisted estimators were quite similar in terms of RMSE. Our proposed data-driven method for finding suitable transformations of auxiliary variables was shown to improve the efficiency of these estimators. For Norway spruce, improvements by up to 10% were obtained. Rather than using an expert selection of a smaller set of auxiliary variables, it can be tempting to use auxiliary information contained in a larger set of variables. In such cases, a standard use of REG often fails due to (near) collinearity, and some auxiliary variables may need to be excluded before the estimate can be computed. We considered two different approaches of excluding "problematic" auxiliary variables, and the variant of the REG estimator that excluded as few variables as possible (the REG^{ii} estimator) provided the best results (with a few exceptions). The simulations showed that the efficiency in terms of RMSE improved when using the large set of auxiliary variables for LASSO, RIDGE, and ELNET, but that this was not necessarily the case for REG estimators. When estimating, for example, total growing stock volume (for both sample sizes considered) or volume of Norway spruce (for the larger sample size), the data-driven method for selecting transformations of auxiliary variables further improved the efficiency of LASSO, RIDGE, and ELNET. Although these improvements were statistically significant at the 5% level, they were all small.

When estimating total growing stock volume or volume of Norway spruce, LASSO, RIDGE, and ELNET based on the large set of auxiliary variables were the best in terms of RMSE. For the smaller sample size, they performed much better than the corresponding REG^{ii} estimator. For volume of broad-leaved trees, LASSO, RIDGE, and ELNET based on the large set of auxiliary variables in their original scale showed the best performance. Here, for the smaller sample size, they performed much better than REG^{ii}, which in this case had an RMSE even larger than the Horvitz-Thompson estimator.

The suggested data-driven choices of transformations performed the best when estimating volume of Norway spruce, using an expert selection of auxiliary variables, where they reduced the RMSEs by 7–10%. Although the transformations resulted in statistically significant reductions of RMSE in many other cases, too, these improvements cannot be regarded as practically significant. In addition, for the smaller sample size, the data-driven choices of transformations sometimes resulted in huge increases of RMSE, in particular when combined with REG^{ii}, and to a lesser degree with LASSO. In comparison, RIDGE (and to some extent also ELNET) was found to be quite robust against poor choices of transformations. Thus, the data-driven method for selecting transformations has not been proven promising enough to be recommended for the type of applications considered in this paper, except possibly in combination with RIDGE and ELNET.

Cook's distance is a commonly used metric to indicate the influence of a data point when performing a multiple regression analysis. In an attempt to make the data-driven method more robust and in an additional simulation study not presented here, we disallowed transformations that caused an excessive increase in Cook's distance. This improved the performance of the estimators of volume of broad-leaved trees, but it was still found that for broad-leaved trees it is better to use auxiliary variables in their original scale.

In our proposed data-driven method for finding suitable transformations, the transformation selection was done for one auxiliary variable at a time. To improve the method, and the efficiency of the resulting model-assisted estimators, one can use multivariable fractional polynomials, which simultaneously determine a functional form for continuous auxiliary variables and delete uninfluential auxiliary variables (Sauerbrei et al., 2006; Sauerbrei and Royston, 2017). For our simulation study, however, the additional computational burden of using multivariable fractional polynomials was considered too high. Another topic for further studies is the inclusion of interaction terms in the models. Except for one interaction term in the model for total growing stock volume based on an expert selection of auxiliary variables, only main effects were included in our models. Potentially, many interactions can be used. To avoid overfitting, and not only for models with interactions, a possibility is to use an information criterion, such as the Akaike information criterion (Akaike, 1974).

Although the methods might be further improved, our results indicate that model-assisted methods like LASSO, RIDGE, and ELNET could be used by the Swedish NFI to provide reliable estimates for smaller areas than possible using field data alone. Today, counties are the smallest unit for which the NFI present reliable estimates. The smallest area for which reliable results can be presented depends in large part on how the model-assisted estimators perform when using smaller sample sizes than the ones used in this study ($n < 901$). Thus, it remains to be investigated how small areas can be to produce reliable estimates of different forest variables with a sufficiently low RMSE.

A relatively large proportion of the units (raster cells) in our population (18%) had a difference between P_{95} and field measured tree height that was greater than 7 m. These units

were considered as outliers. Many of them were units that were clear felled after the field survey, but before the laser scanning took place. The large proportion of outliers could also be a consequence of using merged split-plots for which the linkage with laser data is more sensitive to plot location errors compared to un-split plots. In the Monte Carlo study, it was found better to perform the data-driven choices of transformations *without* using these outliers in a sample, but to estimate model parameters *with* the outliers in a sample of auxiliary variable data values (where variables may have been transformed before the estimation is done). In addition to these outliers, there were additional units in the population with an unusual relationship between field data and laser metrics. This could be, for example, due to thinning cuttings, wind-thrown trees, and other changes. It was noticed that the proportion of such units was higher for plots with a high proportion of broad-leaved trees. To some extent, this can be an effect of using laser data acquired during leaf-off conditions, which gives lower laser density metrics for broad-leaved forests than using data acquired during leaf-on conditions (White et al., 2013). Although the number of such units was relatively low, they might have a large influence on the selection of transformations, and may explain why the use of data-driven choices of transformations was not successful when estimating volume of broad-leaved trees.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article are available from the Dryad Digital Repository: doi: 10.5061/dryad.s4mw6m97k.

AUTHOR CONTRIBUTIONS

ME conceived the study, was in charge of overall direction and planning, and carried out the Monte Carlo simulations. MN retrieved all data and contributed to the analysis with expertise in remote sensing. ME wrote the first draft of the manuscript, except for section "Materials and Methods," written by MN. Both authors contributed to manuscript revision, read, and approved the submitted version. Both authors involved the participatory research process.

FUNDING

This research was financially supported by a research grant from the Swedish National Space Board.

ACKNOWLEDGMENTS

We acknowledge the Swedish National Forest Inventory for providing field data. We thank Håkan Olsson, Anton Grafström, guest associate editor BW, and two referees for their comments on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/ffgc.2021.764495/full#supplementary-material

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Ambler, G., and Banner, A. (2015). *MFP: Multivariable Fractional Polynomials. R package version 1.5.2.*

Baffetta, F., Corona, P., and Fattorini, L. (2010). Design-based diagnostics for k-nn estimators of forest resources. *Can. J. For. Res.* 41, 59–72. doi: 10.1139/X10-157

Baffetta, F., Fattorini, L., Franceschi, S., and Corona, P. (2009). Design-based approach to k-nearest neighbours technique for coupling field and remotely sensed data in forest surveys. *Remote Sens. Environ.* 113, 463–475. doi: 10.1016/j.rse.2008.06.014

Berk, R. A. (2016). *Statistical Learning from a Regression Perspective*, 2nd Edn. Cham: Springer International Publishing. doi: 10.1007/978-3-319-44048-4

Breidt, F. J., and Opsomer, J. D. (2016). Model-assisted survey estimation with modern prediction techniques. *Stat. Sci.* 32, 190–205. doi: 10.1214/16-STS589

Breidt, F. J., Claeskens, G., and Opsomer, J. D. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika* 92, 831–846. doi: 10.1093/biomet/92.4.831

Chambers, R., van den Brakel, J., Hedlin, D., Lehtonen, R., and Zhang, L.-C. (2006). Future challenges of small area estimation. *Stat. Transit.* 7, 759–769.

Chatterjee, S., Hadi, A. S., and Price, B. (2012). *Regression Analysis by Example*, 5th Edn. Hoboken, NJ: Wiley.

Chirici, G., McRoberts, R. E., Fattorini, L., Mura, M., and Marchetti, M. (2016). Comparing echo-based and canopy height model-based metrics for enhancing estimation of forest aboveground biomass in a model-assisted framework. *Remote Sens. Environ.* 174, 1–9. doi: 10.1016/j.rse.2015.11.010

Davison, A. C., and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: University Press. doi: 10.1017/CBO9780511802843

Ekström, M., Esseen, P.-A., Westerlund, B., Grafström, A., Jonsson, B. G., and Ståhl, G. (2018). Logistic regression for clustered data from environmental monitoring programs. *Ecol. Informatics* 43, 165–173. doi: 10.1016/j.ecoinf.2017.10.006

Fridman, J., Holm, S., Nilsson, M., Nilsson, P., Ringvall, A. H., and Ståhl, G. (2014). Adapting national forest inventories to changing requirements – the case of the Swedish National Forest Inventory at the turn of the 20th century. *Silva Fenn.* 48:1095. doi: 10.14214/sf.1095

Gregoire, T., Ståhl, G., Næsset, E., Gobakken, T., Nelson, R., and Holm, S. (2011). Model-assisted estimation of biomass in a LiDAR sample survey in Hedmark county, Norway. *Can. J. For. Res.* 41, 83–95. doi: 10.1139/X10-195

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd Edn. New York, NY: Springer. doi: 10.1007/978-0-387-84858-7

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton, FL: CRC Press. doi: 10.1201/b18401

Horvitz, D. G., and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* 47, 663–685. doi: 10.1080/01621459.1952.10483446

Hosking, J. R. M. (1990). L-moments: analysis and estimation of distributions using linear combinations of order statistics. *J. R. Stat. Soc. Ser. B* 52, 105–124. doi: 10.1111/j.2517-6161.1990.tb01775.x

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning*, 2nd Edn. New York, NY: Springer. doi: 10.1007/978-1-0716-1418-1

Kangas, A., Myllymäki, M., Gobakken, T., and Næsset, E. (2016). Model-assisted forest inventory with parametric, semiparametric, and nonparametric models. *Can. J. For. Res.* 46, 855–868. doi: 10.1139/cjfr-2015-0504

McConville, K. S., Breidt, F. J., Lee, T. C. M., and Moisen, G. G. (2017). Model-assisted survey regression estimation with the lasso. *J. Surv. Stat. Methodol.* 5, 131–158. doi: 10.1093/jssam/smw041

McConville, K. S., Moisen, G. G., and Frescino, T. S. (2020). A tutorial on model-assisted estimation with application to forest inventory. *Forests* 11:244. doi: 10.3390/f11020244

McConville, K., Tang, B., Zhu, G., Cheung, S., and Li, S. (2018). *Mase: Model-Assisted Survey Estimation. R package version 0.1.2.*

McGaughey, R. J. (2020). *FUSION/LDV: Software For LIDAR Data Analysis and Visualization.* Available online at: http://forsys.cfr.washington.edu/fusion/ (accessed January 18, 2021).

McRoberts, R. E., Liknes, G., and Domke, G. M. (2014). Using a remote sensing-based, percent tree cover map to enhance forest inventory estimation. *For. Ecol. Manag.* 331, 12–18. doi: 10.1016/j.foreco.2014.07.025

Moser, P., Vibrans, A. C., McRoberts, R. E., Næsset, E., Gobakken, T., Chirici, G., et al. (2017). Methods for variable selection in LiDAR-assisted forest inventories. *Forestry* 90, 112–124. doi: 10.1093/forestry/cpw041

Naturvårdsverket (2020). *Nationella Marktäckedata 2018, Basskikt – Produktbeskrivning. Utgåva 2.2, Naturvårdsverket*. Available online at: http://gpt.vic-metria.nu/data/land/NMD/NMD_Produktbeskrivning_NMD2018Basskikt_v2_2.pdf (accessed November 26, 2021).

Nilsson, M., Nordkvist, K., Jonzén, J., Lindgren, N., Axensten, P., Wallerman, J., et al. (2017). A nationwide forest attribute map of Sweden predicted using airborne laser scanning data and field data from the National Forest Inventory. *Remote Sens. Environ.* 194, 447–454. doi: 10.1016/j.rse.2016.10.022

Opsomer, J. D., Breidt, F. J., Moisen, G. G., and Kauermann, G. (2007). Model-assisted estimation of forest resources with generalized additive models (with discussion). *J. Am. Stat. Assoc.* 102, 400–416. doi: 10.1198/016214506000001491

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Royston, P., and Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *J. R. Stat. Soc. Ser. C* 43, 429–467. doi: 10.2307/2986270

Sabanés Bové, D., and Held, L. (2011). Bayesian fractional polynomials. *Stat. Comput.* 21, 309–324. doi: 10.1007/s11222-010-9170-7

Samuels, M. L., Witmer, J. A., and Schaffner, A. A. (2012). *Statistics for the Life Sciences*, 4th Edn. Boston, FL: Prentice Hall.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York, NY: Springer. doi: 10.1007/978-1-4612-4378-6

Sauerbrei, W., and Royston, P. (2017). "The multivariable fractional polynomial approach, with thoughts about opportunities and challenges in big data," in *Big Data Clustering: Data Preprocessing, Variable Selection, And Dimension Reduction.* WIAS Report 29, ed. H.-J. Mucha (Berlin: Weierstraß-Institut für Angewandte Analysis und Stochastik), 36–54.

Sauerbrei, W., Meier-Hirmer, C., Benner, A., and Royston, P. (2006). Multivariable regression model building by using fractional polynomials: description of SAS, STATA and R programs. *Comput. Stat. Data Anal.* 50, 3464–3485. doi: 10.1016/j.csda.2005.07.015

Ståhl, G., Saarela, S., Schnell, S., Holm, S., Breidenbach, J., Healey, S. P., et al. (2016). Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation. *For. Ecosyst.* 3:5.

Tomppo, E., Heikkinen, J., Henttonen, H. M., Ihalainen, A., Katila, M., Mäkelä, H., et al. (2011). "Designing and conducting a forest inventory – case: 9th national forest inventory of finland," in *Managing Forest Ecosystems*, Vol. 22, ed. K. von Gadow (Dordrecht: Springer). doi: 10.1007/978-94-007-1652-0

Torgo, L. (2010). *Data Mining With R: Learning With Case Studies*. Boca Raton, FL: Chapman and Hall/CRC. doi: 10.1201/b10328

Vaughan, T. S., and Berry, K. E. (2005). Using Monte Carlo techniques to demonstrate the meaning and implications of multicollinearity. *J. Stat. Educ.* 13, 1–9. doi: 10.1080/10691898.2005.11910640

White, J. C., Wulder, M. A., Vastaranta, M., Coops, N. C., Pitt, D., and Woods, M. (2013). The utility of image-based point clouds for forest inventory: a comparison with airborne laser scanning. *Forests* 4, 518–536.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Conflict of Interest: The 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.

frontiers in Forests and Global Change

ORIGINAL RESEARCH
published: 15 December 2021
doi: 10.3389/ffgc.2021.745840

Small-Area Estimation for the USDA Forest Service, National Woodland Owner Survey: Creating a Fine-Scale Land Cover and Ownership Layer to Support County-Level Population Estimates

Vance Harris[1], Jesse Caputo[1,2], Andrew Finley[3], Brett J. Butler[1,2]*, Forrest Bowlick[1] and Paul Catanzaro[1]*

[1] Department of Environmental Conservation, Family Forest Research Center, University of Massachusetts Amherst, Amherst, MA, United States, [2] United States Department of Agriculture, Northern Research Station, Forest Service, Amherst, MA, United States, [3] Department of Forestry, College of Agriculture and Natural Resources, Michigan State University, East Lansing, MI, United States

OPEN ACCESS

Edited by:
Philip Radtke, Virginia Tech, United States

Reviewed by:
Steve Prisley, National Council for Air and Stream Improvement, Inc., (NCASI), United States
Garret Dettmann, Virginia Tech, United States

***Correspondence:**
Vance Harris
vharris@umass.edu
Brett J. Butler
brett.butler2@usda.gov

Specialty section:
This article was submitted to Forest Management, a section of the journal Frontiers in Forests and Global Change

Received: *22 July 2021*
Accepted: *16 November 2021*
Published: *15 December 2021*

Citation:
Harris V, Caputo J, Finley A, Butler BJ, Bowlick F and Catanzaro P (2021) Small-Area Estimation for the USDA Forest Service, National Woodland Owner Survey: Creating a Fine-Scale Land Cover and Ownership Layer to Support County-Level Population Estimates. Front. For. Glob. Change 4:745840. doi: 10.3389/ffgc.2021.745840

Small area estimation is a powerful modeling technique in which ancillary data can be utilized to "borrow" additional information, effectively increasing sample sizes in small spatial, temporal, or categorical domains. Though more commonly applied to biophysical variables within the study of forest inventory analyses, small area estimation can also be implemented in the context of understanding social values, behaviors, and trends among types of forest landowners within small domains. Here, we demonstrate a method for deriving a continuous fine-scale land cover and ownership layer for the state of Delaware, United States, and an application of that ancillary layer to facilitate small-area estimation of several variables from the USDA Forest Service's National Woodland Owner Survey. Utilizing a proprietary parcel layer alongside the National Land Cover Database, we constructed a continuous layer with 10-meter resolution depicting land cover and land ownership classes. We found that the National Woodland Owner Survey state-level estimations of total acreage and total ownerships by ownership class were generally within one standard error of the population values calculated from the raster layer, which supported the direct calculation of several population-level summary variables at the county levels. Subsequently, we compare design-based and model-based methods of predicting commercial harvesting by family forest ownerships in Delaware in which forest ownership acreage, taken from the parcel map, was utilized to inform the model-based approach. Results show general agreement between the two modes, indicating that a small area estimation approach can be utilized successfully in this context and shows promise for other variables, especially if additional variables, e.g., United States Census Bureau data, are also incorporated.

Keywords: private forest land, family forest ownerships, commercial forest harvesting, small area estimation, model-based estimations

INTRODUCTION

Of the approximate 816 million acres of forestland in the United States, private land ownership accounts for an estimated 56% (Butler et al., 2021). Therefore, understanding private land ownership attitudes and behaviors is fundamental to successfully cultivating socially positive land stewardship practices (Kumer and Štrumbelj, 2017; Mozgeris et al., 2017; Sotirov et al., 2019; Butler et al., 2021). Private forest landowners consist of "forest industry companies, other businesses or corporations, partnerships, tribes, families, and individuals" (Butler and Leatherberry, 2004) according to the National Woodland Owner Survey (NWOS). However, dynamic heterogeneity in attitudes and behaviors, both within and between private ownership classes, require robust datasets and appropriate models to accurately differentiate trends in ownership typologies (Kumer and Štrumbelj, 2017; Sotirov et al., 2019). Efforts to conduct high resolution ownership analyses have historically been thwarted due to a lack of sufficient data (Sotirov et al., 2019). Low sampling sizes generated from the results of these surveys cause problems such as low statistical power, inflated effect estimations and poor replicability. Low sample sizes also require that population estimates be calculated within relatively large spatial domains (nationwide, regional or state-level) in order to ensure sufficient levels of precision.

Traditional mechanisms for understanding forest ownership behavior are through social surveys *via* mail, phone, or the internet (Kumer and Štrumbelj, 2017; Sotirov et al., 2019; Butler et al., 2021). In the United States, the NWOS, a product of the United States Department of Agriculture, Forest Service Forest Inventory and Analysis (FIA) Program, is the official survey aimed at increasing understanding regarding private forest owners (Butler et al., 2021). The target sample size for the NWOS is 250 responses per geographic unit, based on a target coefficient of variation of 5% (Butler and Caputo, 2021). In practice, however, the NWOS reporting protocol allows for published estimates for geographic regions with at least 100 responses. This lower target is always met at the regional level and for most states (Butler et al., 2021), but rarely at the sub-state or county level. This level of analysis is not always sufficient to make programmatic or policy decisions, such as for forestry assistance programs, at the county or sub-state scale.

To compensate for the low sample size at the sub-state level and to allow for accurate, precise estimation of NWOS attributes at finer scales, this pilot effort focuses on the development of a parcel-level land cover and ownership layer for use in small area estimation. Small area estimation (SAE) refers generally to approaches for making population-level estimates within small domains for which sample sizes are deemed inadequate to produce estimates of acceptable precision using traditional design-based techniques. This umbrella term refers to a number of methods that rely on ancillary data sources in order to "borrow" additional information, increasing the effective sample size – and consequently, the precision of the estimates – for the selected domain. In particular, model-based small area estimation techniques can deliver rich inference – especially when compared to traditional design-based approaches. Design-based and model-based modes of inference have long been contrasted in survey research (Little, 2004). Design-based inference automatically accounts for the survey design but has limited ability to leverage ancillary information and deliver precise estimates for small sample sizes (i.e., small area estimates). On the other hand, model-based inference must explicitly consider the design and data jointly, but can use ancillary information and borrow from the rich modeling literature to deliver robust inference for small samples sizes. Within the model-based realm, Bayesian methods provide additional flexibility in model specification and inference (Ghosh and Meeden, 1997; Rao, 2011; Chen et al., 2017). Small area estimation is an increasingly important tool for forest inventory analyses (Breidenbach et al., 2020). To date, however, most efforts have focused on estimation of biophysical variables (Breidenbach and Astrup, 2012; Goerndt et al., 2019; Green et al., 2020). An equal need, however, exists for precise estimates of ownership attributes within small domains, especially small (i.e., sub-state) spatial domains.

As part of the NWOS efforts, forest ownership spatial products have been periodically released (Hewes et al., 2014; Sass et al., 2020). Updates to these map layers have incorporated newer information and increasing resolution of forest ownership categories. These spatial products have used a Thiessen polygon approach based on FIA plot and ancillary data to produce wall-to-wall coverage of forest ownership across the conterminous United States (Butler et al., 2014). This technique is acceptable for strategical level analyses and visualization of broad ownership patterns, but it cannot be used for tactical level analyses or any applications where a high level of spatial precision is required. Although spatial layers such as these are potentially rich sources of ancillary data for small area estimation efforts, the current suite of NWOS-derived spatial products do not have the needed level of accuracy or resolution for this purpose.

There are two primary goals in this pilot study: (1) to produce a spatial layer that accurately depicts land cover and ownership classes that are compatible with FIA land classes at a fine resolution (i.e., at the parcel scale), and which result in summary statistics compatible with FIA and NWOS results, and (2) demonstrate the utility of the ownership layer as an input in model-based estimation to produce small domain (e.g., county-level) estimates of the proportion of ownerships engaging in commercial harvesting as good as or better than those produced using the standard NWOS methodology. For the focus of this initial pilot, we decided for operational efficiency to focus on a state that is small and has complete DMP coverage. For that reason, we confined our analysis to the state of Delaware, United States. Although Delaware is not state in which forests and forestry are traditionally seen as important, more than 50% of the state is forested and more than 40% of family forest landowners have harvested sawlogs, firewood, or other forest products (Butler et al., 2021).

MATERIALS AND METHODS

Land Cover and Ownership Layer

To date, NWOS results have primarily been reported for family forest ownerships (FFOs) – families, trusts, individuals, estates,

and family partnerships which own forestland (Butler et al., 2021). Ownerships are defined as groups of one or more owners that jointly own one or more forested parcels. FFOs are defined as ownerships owning at least one acre of forest, with forest defined as forested "...land that has at least 10 percent crown cover by live tally trees of any size or has had at least 10 percent canopy cover of live tally species in the past, based on the presence of stumps, snags, or other evidence." (USDA Forest Service, 2016). For comparison's sake, the primary emphasis of this study will also be on FFOs. Summary results include estimates of number of ownerships, acreage, and size of holdings. **Figure 1** illustrates the approach taken to generate these estimates and is further described in this section. However, an additional goal of this study is to generate these estimates for all ownership classes in addition to FFOs, including public, corporate, and other private ownerships. This is possible due to the application of a secondary dataset developed by Digital Map Products Lightbox™ (DMP) [Digital Map Products (DMP), 2021]. This proprietary data includes ownership information at the parcel scale across the United States, including parcel boundaries and owner name and address information. This data is aggregated from individual sources at the state, county, and local levels and represents a standardized continuous vector layer for analysis. Although several state agencies publish parcel-level landcover maps, no non-proprietary layers exist nationwide. Within the scope of this work, both address and name data can be utilized to classify individual parcels to the FIA ownership classes. For this study, we used data for the state of Delaware, nominally current as of 2020.

To calculate the number of unique ownerships within the DMP data, multiple instances of the same ownership associated with two or more parcels needed to be identified. Ownerships were identified and matched utilizing name and address data. Names were transformed into a standardized format to reduce effects such as misspellings, additional/missing name elements and similar erroneous influences. To account for differences in reporting practices, consideration for the type of owners was needed. Individual/family ownerships were isolated and processed differently than those of other legal ownership entities. This was primarily done to enhance the individual/family ownership matches by incorporating home address data, whereas another legal entity (such as a corporate owner) might have multiple mailing addresses associated with local/regional offices of the same company and were therefore matched on name alone. In order to prevent parents and children with similar names living at the same address from being erroneously identified as the same individual, generational suffixes (e.g., Jr, II, etc.) were used to split otherwise very similar names and ensure identification of multiple unique ownerships.

The primary tool used to match records by ownerships names was utilizing a reference table populated with every record and a phonetic code associated with the ownership names. The phonetic codes were generated *via* the Python package DoubleMetaphone (Philips et al., 2007). Utilizing this package, the text-based name data were converted to their phonetic spelling and reduced to their key phonetic elements.[1] Each record's code was written into a reference table, in which all records with the same phonetic spelling were linked. In the case of the individual/family owners, the dataset was first grouped into records with the same home address. From there the names were coded into internal reference tables to match names only within the subgrouping. Once unique IDs were generated, ownership data can be associated with all their relevant parcels. In this manner, for example, total size of forest holdings could be calculated at the ownership level.

The next phase of the analysis was to determine the ownership classes according to the FIA ownership typology (Family, Corporate, Other Private, and Public and Tribal). Both logical classification criteria and a machine learning model (hereto referred as the classification model) were implemented to build upon previously established FIA manual classification methodologies. As a training set, we used a portion of the sample from the 2018 iteration of the NWOS (Butler et al., 2021). All records (n = 8,862) used originally came from the same commercial vendor (DMP lightbox) as the data for the current study and were therefore in a very similar format. All records had been manually classified using the FIA classification. This training set was also preprocessed through the same name standardization methods mentioned earlier, in order to preserve consistency.

The logical classification stage was used to affix ownership classes if known conditions were met, thereby capitalizing on known elements within the ownership names. Primarily these conditions included the presence of keywords or language associated with the FIA ownership classes (e.g., "Revocable trust" or "Living trust" for family owners; "Authority" or "Maintenance" for corporate, etc.). Additionally, keyword searches were implemented in a ordered manner in which searches were given different levels of priority. If records were classified in one search, then they would not be classified in subsequent searches. Delaware has no tribal reservations, and the tribal category was therefore omitted from this analysis. After the logical classification, the remaining unclassified records were then isolated and passed through a classification model.

The classification model was fit to the training set, and implemented using Python's scikit-learn package (Pedregosa et al., 2011) and similarly utilized individual elements in the ownership names as opposed to the name in its entirety. The metric Term Frequency-Inverse Document Frequency (TF-IDF) was utilized to score every name element in the training set to determine every word's association with all other words in the dataset. Identifying impactful (i.e., highly associated) words allows for more robust data to be utilized in the training of the ML model. A random forest classification model was selected for its computational speed and its predictive accuracy.

Before passing the unclassified dataset through the resulting model, this data needed to undergo the same TF-IDF computations as the training set. The full, unclassified dataset was substantially larger than the training set and contained more unique words, therefore, only the words which appeared in the training set were selected as independent variables. The

[1]For example, the name VANCE HARRIS ALLEN JR would hypothetically be split into component words and associated with the following phonetic standardizations: ALNRSRRFNS, ALNRSFNS, ALNJRRFNS, HRSRRFNS, and ALNRSRR.

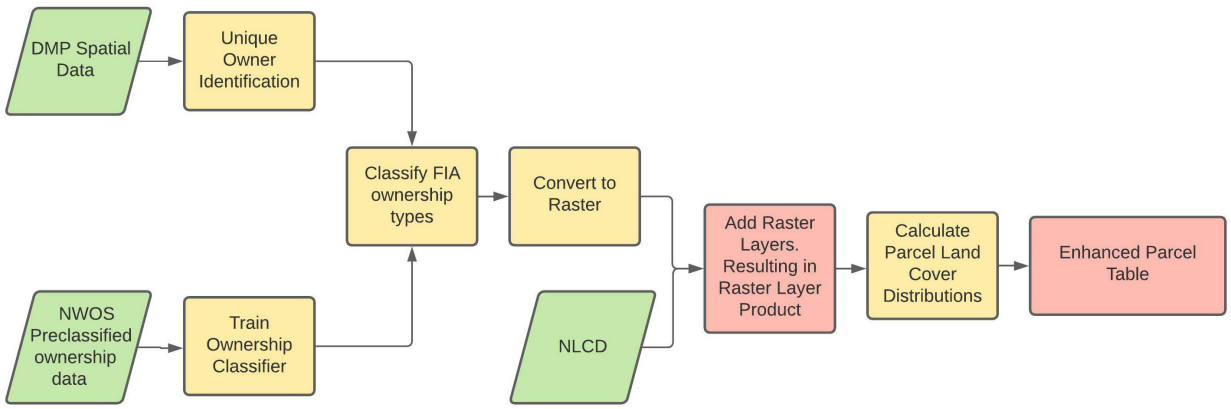


FIGURE 1 | Logic flow chart illustrating the process of creating a land cover and ownership spatial layer, from NLCD and a parcel ownership polygon layer.

unclassified records were then passed through the classification model and FIA ownership class assignments were then available for all records. In these procedures, we attempted to replicate as closely as possible the guidelines that are used for deduplication and classification of the NWOS sample (Butler et al., 2021).

The next phase was to determine the land cover types associated with each parcel. The National Land Cover Database (NLCD) (Dewitz, 2019) 30-meter land cover raster data was utilized here. In order to efficiently join the two data types, the vector-based DMP parcels were converted to a raster with a 10-meter resolution based on the unique Parcel ID field ranging from 1 to 423051. 10 meters allowed for the joining of the two rasters without dropping the smaller parcels. The NLCD data was resampled to a 10-meter resolution. The original 15 land cover classes were reclassified to a numeric label ranging from 100,000,000 (Open Water) to 900,000,000 (Mixed Forest). The larger integer classification allows the Parcel ID values to be appended and preserve both pieces of information in an aspatial format (i.e., parcel[420000] + Deciduous Forest[700000000] = 700420000).

Once achieved, the unique Parcel IDs were utilized to aggregate all pixels and their associated land cover types to each parcel. Total occurrences of each NLCD land cover class within a parcel were divided by the total parcel pixels to determine the percent coverage of each land cover class. We also merged the NLCD classes into a simplified three-class typology, to correspond to the FIA land use classification. Of the 15 original NLCD classes, four (Deciduous Forest, Evergreen Forest, Mixed Forest, and Woody Wetlands) are representative of the FIA-defined "forest" class, one corresponded to "open water", and the remaining 10 corresponded to "non-forest". Although, strictly speaking, FIA is measured and reported in terms of land *use* instead of land *cover*, these four NLCD classes have been determined to correspond adequately to the FIA definition of forest use (Nelson et al., 2020).

The final products include a raster layer in which both ownership and land cover are encoded at the level of the individual pixel, and an enhanced parcel table in which each row represents a single parcel, with parcels assigned ownership IDs allowing for aggregation across ownerships. This last product includes the calculated values for total acreage and acreage by land cover type. By aggregating across this table, we can directly calculate several population-level summary variables at the state or county levels, including the number of forest acres, ownerships, and parcels. In this paper, we focused on forest ownerships owning one or more acres – and in particular on family forest ownerships owning one or more acres. This corresponds to one of the primary strata/domains used for reporting NWOS results (Butler et al., 2021).

Custom National Woodland Owner Survey Estimates

Using the standard NWOS methodology for deriving population-level estimates (Butler and Caputo, 2021; Butler et al., 2021), we estimated statewide and county-level estimates of the total number of FFOs, total FFO acreage, mean size-of-holdings, and total acreage owned by ownerships who have undertaken commercial harvest in the past 5 years (all at the 1+ acre domain). These estimates provide a benchmark comparison to summary variables calculated either directly or indirectly (i.e., through estimation) from the parcel table. It is important to note that there is a small temporal scale mismatch between the DMP data (which are nominally current as of 2020) and the NWOS data (which were collected in 2017/2018), but this magnitude of this mismatch is assumed to be negligible relative to the rate of change inherent to land ownership.

Small Area Estimation

To illustrate the utility of model-based inference, we first developed state- and county-level estimates of a simulated, non-specific response variable, derived for each of the population units (i.e., ownerships) enumerated in the parcel map. We limited our population to FFOs within Delaware holding at least one acre of forest land. The simulated response variable is binary and non-specific, and can be thought of as representing a hypothetical binary attribute of interest (such as, for example, do landowners have forest management plans? Or do landowners hunt on their land?). We help inform these estimates using the county in

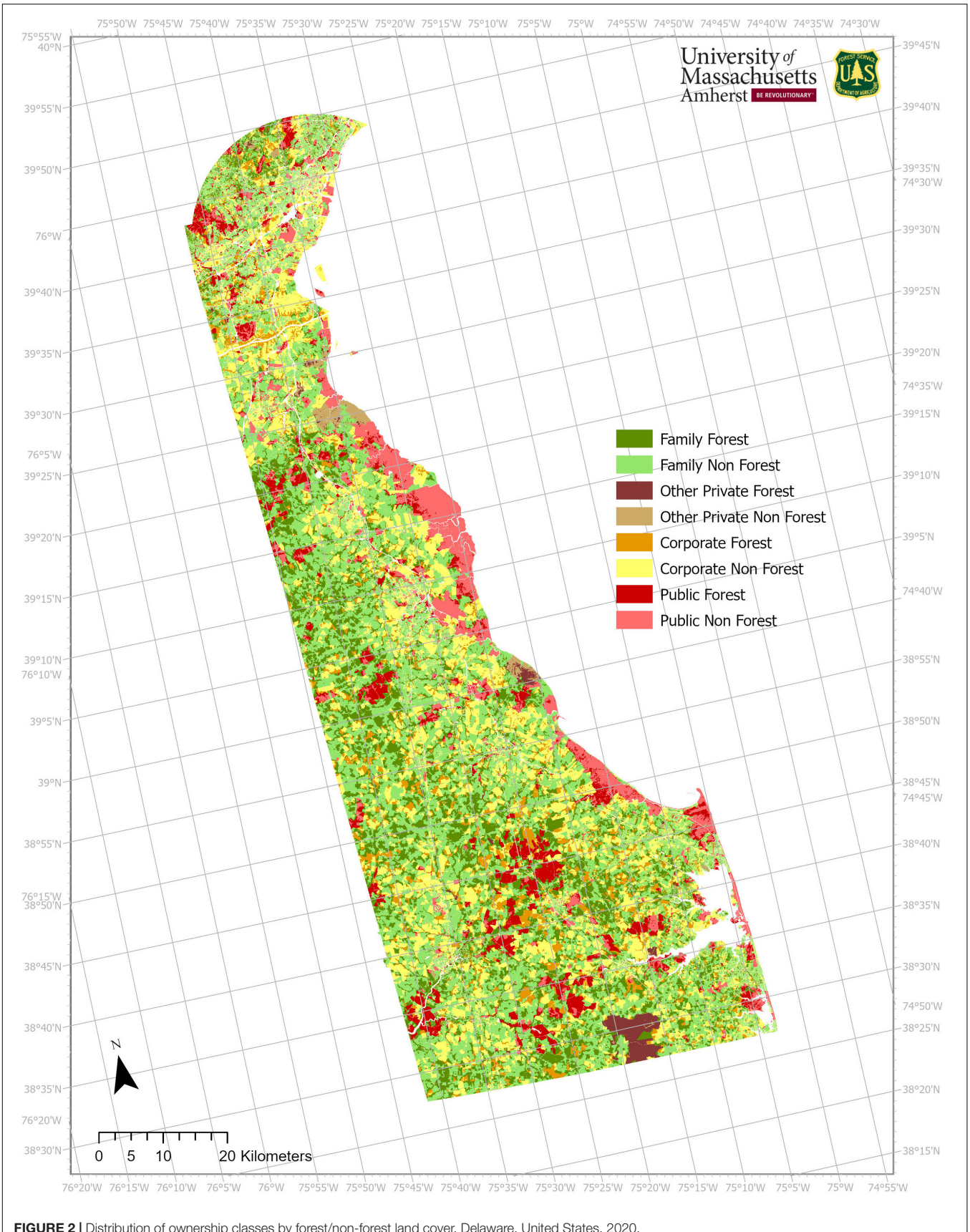


FIGURE 2 | Distribution of ownership classes by forest/non-forest land cover. Delaware, United States, 2020.

TABLE 1 | Total number of ownerships, total acreage, mean size of forest holdings, and mean number of forested parcels, by Ownership Class, statewide and by counties, as derived from a state-wide ownership and land use map.

State	Ownership class	Thousand acres	Thousand ownerships	Mean forested size-of-holdings	Mean number of parcels
DE	Corporate	94.64	3.12	32.54	3.50
DE	Family	192.06	17.68	10.87	1.41
DE	Other private	13.71	0.15	97.39	2.14
DE	Public	73.96	0.24	314.42	9.30
County					
Kent	Corporate	21.90	0.83	46.17	4.17
Kent	Family	62.75	4.83	14.34	1.68
Kent	Other private	1.71	0.03	57.40	2.18
Kent	Public	15.24	0.08	658.28	15.68
New castle	Corporate	20.29	1.24	27.69	3.72
New castle	Family	23.21	3.83	6.87	1.59
New castle	Other private	1.34	0.08	142.14	2.78
New castle	Public	17.77	0.09	552.29	17.25
Sussex	Corporate	52.45	1.63	38.27	4.28
Sussex	Family	106.10	9.53	11.73	1.52
Sussex	Other private	10.66	0.07	175.33	2.71
Sussex	Public	40.95	0.11	550.72	13.13

Delaware, United States, 2020.

which an ownership is situated as well as an ownership's size-of-holdings, defined as the total acreage of forested land owned by the ownership within the state (as derived from the parcel layer). We first undertook a simple simulation exercise, in which a simple random sample of ownerships was taken from the parcel layer and the simulated response variable was simulated for each. This simulation exercise allows us to compare model-based and design-based estimation methods against a known population, in order to demonstrate their utility and justify their subsequent use in making small-area estimates using NWOS data

Model-based inference about a finite population based on a probability sample can be viewed as a prediction problem. Parameters in the posited model are estimated using sample data and subsequently used to predict the response for the unobserved population units (i.e., those not included in the sample). Following Bayesian methods, we estimated the posterior distribution of model parameters and posterior predictive distribution for unobserved population units. The survey design used to select sampling units determines if and how the design is acknowledged in the posited model. Here, we assume a stratified simple random sampling (SRS) design that allows design components to be ignored in the modeling (see, e.g., Gelman et al., 2013).

Given the binary response, availability of covariates from the parcel map, and focus on county-level estimates, a natural model would be a logistic regression with county specific intercept and regression coefficients. We modeled the non-specific, binary response variable y_{ij}, where i and j index ownership and county, respectively, using a Bernoulli distribution and logit link function as

$$logit\left(p_{ij}\right) = \beta_{0j} + x_{ij}\beta_j,$$

$$y_{ij} \sim Bern\left(p_{ij}\right)$$

where, p_{ij} is the probability of the non-specific response variable being true for an ownership, x_{ij} is the total size-of-forest-holdings owned by that ownership, β_{0j} is the county-specific intercept, and β_j is the regression coefficient. To pool sample information, we modeled the normally distributed coefficients as

$$\beta_{0j} \sim N\left(\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0^2\right)$$

$$\beta_j \sim N(\boldsymbol{\mu}_x, \boldsymbol{\sigma}_x^2)$$

with means μ and variances σ^2. To complete the Bayesian specification, we assigned noninformative prior distributions to all model parameters. We refer to this as the *full model*. For comparison, we also consider a *sub model* that includes

TABLE 2 | Simulated data parameters (True) and candidate models' posterior distribution median and lower and upper 95% credible intervals in parentheses.

Parameter	True	Sub model	Full model
β_{01}	−1.31	0.83 (0.22, 1.51)	−1.24 (−2.44, 0.06)
β_{02}	−0.91	0.07 (−0.67, 0.83)	−1.09 (−2.13, 0.07)
β_{03}	−1.42	−0.97 (−1.45, −0.52)	−1.68 (−2.35, −1.07)
μ_0	−1	−0.03 (−7.67, 7.30)	−1.38 (−6.26, 3.73)
$\sigma^2{}_0$	0.5	2.00 (0.53, 26.75)	0.94 (0.05, 19.20)
β_{x1}	0.78		0.64 (0.22, 1.21)
β_{x2}	0.40		0.31 (0.05, 0.71)
β_{x3}	0.05		0.07 (0.03, 0.11)
μ_x	0.3		0.33 (−2.18, 3.04)
$\sigma^2{}_x$	0.3		0.62 (0.12, 10.10)
WAIC		209.14	178.4
P_w		3.07	5.66

The last two rows hold model WAIC goodness of fit and complexity penalty term. Subscript terms 1, 2, and 3 correspond to Kent, New Castle, and Sussex Counties, respectively.

TABLE 3 | County and state-wide proportions of family forest landowners in Delaware, United States, associated with a simulated, non-specific attribute.

			Model-based	
Parameter	**True**	**Design-based**	**Sub model**	**Full model**
Kent	0.78	0.70 (0.56, 0.84)	0.69 (0.55, 0.82)	0.74 (0.64, 0.84)
New castle	0.60	0.52 (0.32, 0.72)	0.52 (0.33, 0.70)	0.52 (0.36, 0.69)
Sussex	0.30	0.27 (0.18, 0.36)	0.28 (0.19, 0.37)	0.28 (0.20, 0.36)
Statewide	0.49	0.44 (0.37, 0.51)	0.44 (0.37, 0.51)	0.45 (0.39, 0.51)

The table compares the true (i.e., simulated) values with design-based and model-based estimates. Design-based estimates are proportions with 95% confidence interval, model-based estimates are proportions with 95% credible interval.

only the county-varying intercept without the forest ownership acreage covariate.

The simulation exercise compares design- and model-based inference for data similar to that collected in the NWOS. The study generated a realization of y_{ij}'s for all population units using the full model and parameter values provided in **Table 2**. From this population we drew a simple random sample of size $n = 167$ (which is the sample size of the most recent NWOS survey of Delaware). The number of sampling units within each county was proportional to the number of population units in that county. Given this sample, county and state estimates were generated using the sub model, full model, and design-based stratified estimator. The design-based estimator for proportions, given a stratified simple random sample, is defined in Lohr (1999). The sub and full models were compared using the widely applicable information criterion (*WAIC*; Watanabe, 2010). This criterion favors models with better fit to observed data while penalizing models by their effective number of parameters (P_w). Models with lower values WAIC have better fit to the observed data and should yield better out-of-sample prediction, see Gelman et al. (2013), Vehtari et al. (2017), or Green et al. (2020) for more details.

After assessing the results of the simulation exercise and affirming the suitability of model-based estimation in this case, we fit additional models (equivalents of both the full and sub models) to the actual 2018 NWOS sample for Delaware (Butler et al., 2021) in place of the simulated sample. The observational units of the NWOS are individual ownerships and survey responses apply to all parcels owned by each ownership. Survey responses include measurements of ownerships' size-of-holdings as well as the measured response variable, occurrence of commercial harvest. This variable is binary, coded one if harvest occurred on an ownership's holdings within the past 5 years and 0 otherwise. We then estimated the proportion of family forest ownerships (1+ acres) who undertook commercial harvests, by using these models to predict harvest occurrence for each of the unobserved parcels in the complete parcel layer. We then compared these estimates to estimates produced using the standard NWOS methodology and the same raw data. These estimates were used as a point-of-comparison in place of the SRS that was used for that purpose in the simulation exercise, as the NWOS sample does not use an SRS design (Butler et al., 2021). All estimation was done using R (R Core Team, 2019).

RESULTS AND DISCUSSION

Land Use and Ownership Layer

Per the intent of the first goal of this study, **Figure 2** represents the layer for the state of Delaware United States that accurately

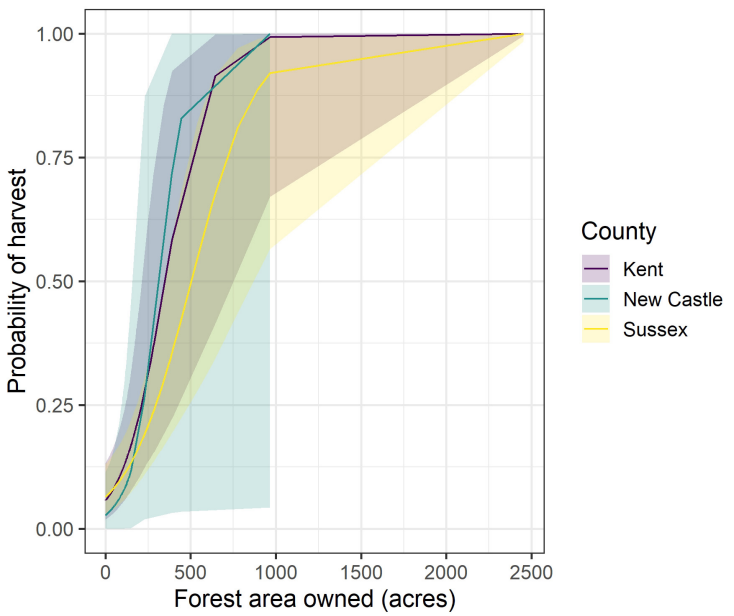


FIGURE 3 | Predicted probability of commercial forest harvest (previous 5 years) by family forest ownerships, by county and size-of-forest-holdings in Delaware, United States. Lines represent posterior median and bands show 95% credible intervals.

depicts land cover and ownership classes that are compatible with FIA land classes at a 10-meter resolution.

A cross-validation approach was utilized to assess the Ownership Classification model accuracy. The training dataset was split into a fitting set (80%, n = 7090) and a testing set (20%, n = 1772). The model was trained using the fitting set and was applied to the testing set in order to generate a predictive classification. The testing set's actual classification was compared to the predicted classification and yielded a measure of weighted accuracy of approximately 96%. This error metric only accounts for the specific variability introduced in this ownership classification phase and does not address the accuracy of the entire study as a whole.

At a 10-meter resolution, the final layer had a total of 48.6 million pixels. Forested pixels make up 32.3% of the total area, with family forest being the largest share – 17.1% of the total pixels. The other 67.7% consists of non-forest including developed land, agriculture, barren land, shrubland and the like. Public roads and open water are represented by null values. Most non-forest land is family (36.0%) or corporate (22.2%), followed by public (8.4%) and other private[2] (1.1%).

Among forest ownerships with one or more acres, the predominant type of ownership is family, accounting for 192.1 thousand acres (**Table 1**). The published NWOS estimate of this value (198 thousand acres, SE = 13; Butler et al., 2021) is within one standard error of the "true" value represented by the parcel table. Likewise, the published estimates of the acreages owned by public and other private entities are within one standard error of the true population values. The largest discrepancy is with corporate ownerships; the published estimate of 68 thousand acres (SE = 13) is more than two standard errors less than the true population value, 94.6 thousand acres. Delaware's family forests are owned by 17.7 thousand unique family forest ownerships. Family forest owners own on average a mean of 1.4 parcels and a mean of 10.9 acres of forest land. Standard estimates of the total number of ownerships and mean size-of-forest-holdings calculated using the published NWOS methodology (18.4 thousand total ownerships, SE = 2.9; mean size-of-forest-holdings = 10.7 acres, SE = 2.0) are within one standard error of the population values.

The greatest proportion of FFO acreage and ownerships is found in Sussex County, followed by Kent and New Castle Counties (**Table 1**). Total FFO acreage ranges from 23.2 to 10.6 thousand acres, and in all cases the standard estimates are within one standard error of the true population values. Total numbers of FFOs range from 3.8 to 9.5 thousand ownerships. In Kent and New Castle Counties, the estimates are within one standard error of the population totals. In Sussex County, they are not. The mean size-of-forest-holdings ranged from 6.9 acres in New Castle County to 14.3 acres in Kent County. In all but Sussex County, the standard estimates are within one standard error of the population level. Given the smaller sample sizes, it is not surprising that the county-level standard estimates are less accurate relative to the population levels as compared with the state-wide estimates. All counties have a sample size of less than 100, a standard adopted in NWOS reporting as an indicator of reliability (Butler et al., 2021).

[2]Other private consists of conservation organizations, NPOs, community groups, and unincorporated private entities.

Small Area Estimation

The true parameter values used to generate the simulated population along with their associated estimates from the sub and full models are provided in **Table 2**. Given the full model is the model used to generate the population data, it is reassuring that all posterior 95% credible intervals capture their respective *True* parameter values. Further, the WAIC given in the second to last row in **Table 2** correctly identifies the full model as the most plausible for the given data.

The true values and estimates for the proportion of ownerships associated with the simulated response variable are given in **Table 3**. The values in this table show negligible differences among design-based and model-based estimates. Also, while the interpretation of the design-based confidence interval and Bayesian model-based credible intervals is fundamentally

FIGURE 4 | Predicted probability of commercial forest harvest by family forest ownerships across Delaware, United States. **(A)** Shows the probability of commercial harvest in the previous 5 years. **(B)** Shows the width of the 95% credible interval. Null values are represented in gray and correspond to public ownerships, corporate ownerships, other non-family private ownerships, and family ownerships owning less than one acre of forest. Values at the scale of the individual pixel reflect harvest probabilities across the entire parcel, regardless of whether the specific pixel is forest or non-forest.

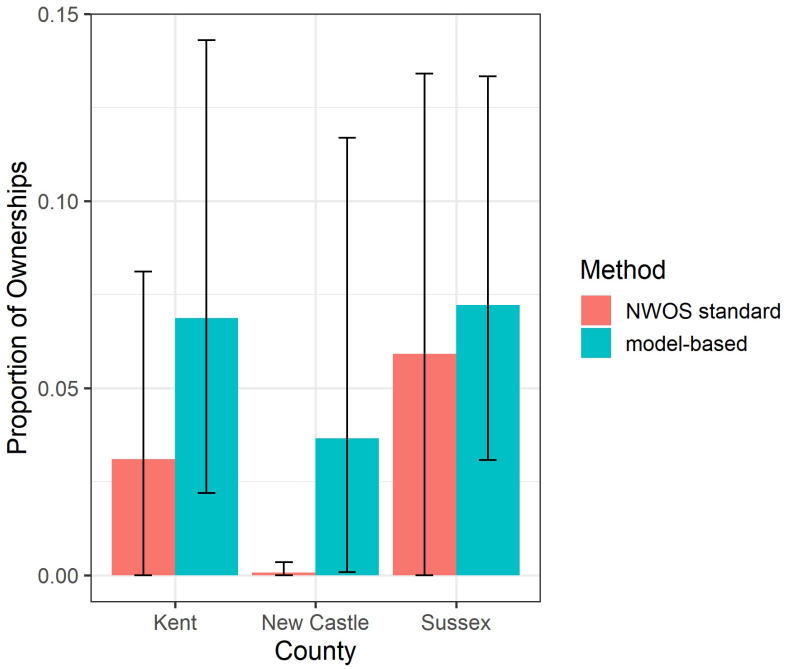


FIGURE 5 | Proportion of family forest ownerships (1+ ac) who have conducted commercial harvests in the past 5 years, by county. Delaware, United States, 2018–2020. Comparison of NWOS standard estimates and model-based estimates. Error bars are 95% confidence intervals for the NWOS estimates and 95% credible intervals for the model-based estimates.

different, they both reflect approximately the same level of uncertainty. Theoretically, given some conditions and parameter prior specifications, the sub model can be shown to replicate the design-based estimate (Ghosh and Meeden, 1997), and it is therefore not surprising their parameter estimates are so similar. Compared with the design-based and sub model, the full model yields slightly narrower credible interval estimates, reflecting the additional information provided by the covariate.

Based on the results of the simulation exercise, we feel confident that the model-based estimators are sufficiently accurate, precise, and unbiased for use in making estimates of NWOS attributes at the county-level. Consequently, we adopted the full model as the preferred model for making estimates of the measured response variable, commercial harvest. While population parameter estimates at various levels are our main interest, the model-based approach does offer additional insights into the relationship between the response and covariates. For example, **Figure 3** summarizes the county-level relationship between commercial harvest variable and the number of acres owned by an ownership as predicted by the full model. In all three counties, the probability of commercial harvest sharply increases with the total amount of forested land owned by an ownership. This probability approaches 1.0 at about 1,000 acres in both Kent and New Castle Counties, but not until ∼2500 acres in Sussex County. Such information can be useful when designing further survey instruments and guiding outreach/policy efforts.

In addition to county- and state-level estimates, the models provide individual population unit level posterior predictive distributions which can be summarized and mapped at the level of the individual parcel or pixel. As can be seen in **Figure 4**, commercial harvest is relatively improbable across the bulk of family forest land (median = 7.5%), with probability increased on larger parcels. Only a few of the largest parcels are both highly probable and highly certain (i.e., they have a narrow 95% credible interval) to conduct commercial harvest. Overall, the width of the intervals for individual parcels ranges from 0.4 to 96.7%. Information such as this is potentially valuable for targeting regions (or even individual parcels) for programs and interventions.

At the county-level, the full model estimates of the proportion of family forest ownerships with one or more acres having had commercial harvest in the previous 5 years range from 3.7% in New Castle County to 7.2% in Sussex County (**Figure 5**). In all cases, the error bars of the NWOS standard estimates (95% confidence intervals) overlapped with those of the model-based estimates (95% credible intervals). The complete code and output for both the full and sub commercial harvest models is available in **Supplement 1**.

CONCLUSION AND FUTURE DIRECTIONS

Using parcel level ownership data in tandem with the National Land Cover Database was more than sufficient for the creation of a continuous land cover and ownership class surface across the state of Delaware, United States. The models created were successful in the classification of ownership classes based

on ownership name data with an accuracy estimate of 96%. Furthermore, the final layer resulted in aggregated values of acreage and ownerships that agreed strongly with the published NWOS estimates. Future iterations of this work are going to include optimizing and expanding the application across the United States with a more rigorous focus on quantifying the error of uncertainty at each stage of the process. For ancillary products such as the NLCD, published error rates, ranging from 71 to 97% (Dewitz, 2019), can be utilized, but the uncertainties and inconsistencies in the underlying parcel data and errors in our classification models resist easy quantification. The goal is to obtain a continuous layer coverage, as well as county and sub-state scale estimates for a wide suite of NWOS attributes for every state accompanied with reflective error metrics for each spatial extent.

Future efforts will also need to address some operational hurdles that were not fully resolved within the context of this pilot study. Firstly, Delaware lacks any tribal reservations. However, other states will include that ownership class. Plans are to incorporate identification of tribal ownerships with a spatial overlay of tribal land boundaries at the end of the classification stage. This will render the tribal ownership class with the highest priority, thereby ensuring its preservation within the analysis. Additionally, Delaware had continuous parcel coverage in the DMP Lightbox's^{TM} dataset. Other states will have gaps in coverages, which will impact distribution estimates of ownerships. Faulty estimates will degrade the agreement with the NWOS estimates, and therefore will need to be addressed. Finally, quantifying the null values for the public road network and open water is necessary in order to make the raster product truly continuous.

The results of our simulation exercise, comparing estimates of a simulated variable against a known population, demonstrated that model-assisted estimation using our land cover and ownership layer as a primary input had the potential to produce precise, unbiased estimates. Using the same approach with the actual NWOS data for the state of Delaware, we estimated that 3.7 to 7.2% of ownerships conducted commercial harvest in the past 5 years – estimates that agreed closely with those made using the standard NWOS methodology. This supports the claim that an SAE approach to estimating ownership attributes at sub-state scales is appropriate. In order to increase precision and reduce error estimates, future efforts will likely rely on more optimized models with additional predictor variables. Additional ancillary datasets, such as the Census data, would likely be useful in this regard. This additional data will "lend" even more strength (and consequently precision) to the estimates that are produced.

Ultimately, a small area estimation approach to modeling the social attributes of forest landowners is both feasible and productive. The insights and understanding it can provide within small spatial domains offer many opportunities for research as well as to aid in the efficient implementation of forest management and landowner assistance programs at small scale.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: Raw data used in this analysis are confidential (e.g., NWOS) and/or proprietary (e.g., DMP) and are therefore not publicly available. Requests to access these datasets should be directed to JC, jessecaputo@umass.edu.

AUTHOR CONTRIBUTIONS

VH, JC, and AF did the analyses and wrote the manuscript. VH performed the initial analysis that generated the raster and table outputs to be utilized in JC's and AF's analyses. BB, FB, and PC contributed to the conceptualization and guided the project. All authors contributed to the article and approved the submitted version.

FUNDING

Funding for this research was provided by the USDA Forest Service, State and Private Forestry and the Northern Research Station (Grant No. 20-CS-11242305-116).

ACKNOWLEDGMENTS

The findings and conclusions in this publication are those of the author(s) and should not be construed to represent any official USDA or United States Government determination or policy. We thank Emma Sass for her contributions to the initial discussions that inspired this work. Additionally, we are grateful to Delaware's family forest owners for responding to the National Woodland Owner Survey.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/ffgc.2021.745840/full#supplementary-material

REFERENCES

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

Breidenbach, J., Granhus, A., Hylen, G., Eriksen, R., and Astrup, R. (2020). A century of National Forest Inventory in Norway – informing past, present, and future decisions. *For. Ecosyst.* 7:46. doi: 10.1186/s40663-020-00261-0

Butler, B. J., and Caputo, J. (2021). *Weighting for the USDA Forest Service, National Woodland Owner Survey. Gen. Tech. Rep. NRS-198*. Madison, WI: U.S. Department of Agriculture, Forest Service, Northern Research Station. 24. doi: 10.2737/NRS-GTR-198

Butler, B. J., Butler, S. M., Caputo, J., Dias, J., Robillard, A., and Sass, E. M. (2021). *Family Forest Ownerships of the United States, 2018: Results From the USDA Forest Service, National Woodland Owner Survey. Gen. Tech. Rep. NRS-199*. Madison, WI: USDA Forest Service, Northern Research Station, 52.

Butler, B. J., Hewes, J. H., Liknes, G. C., Nelson, M. D., and Snyder, S. A. (2014). A comparison of techniques for generating forest ownership spatial products. *Appl. Geogr.* 46, 21–34. doi: 10.1016/j.apgeog.2013.09.020

Butler, J. B., and Leatherberry, E. C. (2004). Leatherberry, America's family forest owners. *J. For.* 102, 4–14. doi: 10.1093/jof/102.7.4

Chen, Q., Elliott, M. R., Haziza, D., Yang, Y., Ghosh, M., Little, R., et al. (2017). Approaches to improving survey-weighted estimates. *Statist. Sci.* 32, 227–248.

Dewitz, J. (2019). *National Land Cover Database (NLCD) 2016 Products*. Reston, VA: U.S. Geological Survey, doi: 10.5066/P96HHBIE

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. London: Chapman and Hall/CRC.

Digital Map Products [DMP] (2021). *Unpublished data; Land Parcel and Ownership Layer*. Available online at: https://www.digmap.com/

Ghosh, M., and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. London: Chapman Hall/CRC Press.

Goerndt, M. E., Wilson, B. T., and Aguilar, F. X. (2019). Comparison of small area estimation methods applied to biopower feedstock supply in the Northern U.S. Region. *Biom. Bioener.* 121, 64–77. doi: 10.1016/j.biombioe.2018.12.008

Green, E., Finley, A., and Strawderman, W. (2020). *Introduction to Bayesian Methods in Ecology and Natural Resources*. Cham: Springer International Publishing.

Hewes, J. H., Butler, B. J., Liknes, G. C., Nelson, M. D., and Snyder, S. A. (2014). *Public and Private Forest Ownership in the Conterminous United States: Distribution of Six Ownership Types – Geospatial Database. RDS-2014-0002*. Newtown Square, PA: U.S. Department of Agriculture, Forest Service, Northern Research Station. doi: 10.2737/RDS-2014-0002

Kumer, P., and Štrumbelj, E. (2017). Clustering-based typology and analysis of private small-scale forest owners in slovenia. *For. Policy Econom.* 80, 116–124. doi: 10.1016/j.forpol.2017.03.014

Little, R. (2004). To model or not to model? Competing modes of inference for finite population sampling inference for finite population sampling. *J. Am. Statist. Associat.* 99, 546–556. doi: 10.1198/016214504000000467

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury Press.

Mozgeris, G., Vilis, B., Andrius, S., Marius, K., and Michailas, P. (2017). Owner mapping for forest scenario modelling — a lithuanian case study. *For. Policy Econom.* 85, 235–244. doi: 10.1016/j.forpol.2016.02.002

Nelson, M. D., Riitters, K. H., Coulston, J. W., Domke, G. M., Greenfield, E. J., Langner, L. L., et al. (2020). *Defining the United States Land Base: A Technical Document Supporting the USDA Forest Service 2020 RPA Assessment. Gen. Tech. Rep. NRS-191*. Madison, WI: U.S. Department of Agriculture, Forest Service, Northern Research Station, 70. doi: 10.2737/nrs-gtr-191

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Philips, L., Collins, A., Somerville, M., Barran, R., Dornself, M., Metrot, S., et al. (2007). *"Metaphone.py." Github, Python 3.* Available online at: gist.github.com/nsh87/cba8824ba720181e820f (accessed December 15, 2020).

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rao, J. N. K. (2011). Impact of frequentist and Bayesian methods on survey sampling practice: a selective appraisal. *Statist. Sci.* 26, 240–256.

Sass, E. M., Butler, B. J., and Markowski-Lindsay, M. (2020). *Forest Ownership in the Conterminous United States, 2017: Geospatial Dataset*. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, doi: 10.2737/RDS-2020-0044

Sotirov, M., Ola, S., and Ola, E. L. (2019). Forest owner behavioral models, policy changes, and forest management. an agent-based framework for studying the provision of forest ecosystem goods and services at the landscape level. *For. Policy Econom.* 103, 79–89. doi: 10.1016/j.forpol.2017.10.015

USDA Forest Service (2016). *Forest Inventory and Analysis Glossary*. Available online at: www.nrs.fs.fed.us/fia/data-tools/state-reports/glossary/default.asp (accessed March 4, 2021).

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statist. Comput.* 27, 1413–1432.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* 11, 3571–3594.

Conflict of Interest: The 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Forests and Global Change

ORIGINAL RESEARCH
published: 15 December 2021
doi: 10.3389/ffgc.2021.752911

Hierarchical Bayesian Small Area Estimation Using Weakly Informative Priors in Ecologically Homogeneous Areas of the Interior Western Forests

Grayson W. White[1], *Kelly S. McConville*[2]*, *Gretchen G. Moisen*[3] *and Tracey S. Frescino*[3]

[1] RedCastle Resources, Inc., Salt Lake City, UT, United States, [2] Mathematics Department, Reed College, Portland, OR, United States, [3] Rocky Mountain Research Station, U.S. Department of Agriculture (USDA) Forest Service, Ogden, UT, United States

The U.S. Forest Inventory and Analysis Program (FIA) collects inventory data on and computes estimates for many forest attributes to monitor the status and trends of the nation's forests. Increasingly, FIA needs to produce estimates in small geographic and temporal regions. In this application, we implement area level hierarchical Bayesian (HB) small area estimators of several forest attributes for ecosubsections in the Interior West of the US. We use a remotely-sensed auxiliary variable, percent tree canopy cover, to predict response variables derived from ground-collected data such as basal area, biomass, tree count, and volume. We implement four area level HB estimators that borrow strength across ecological provinces and sections and consider prior information on the between-area variation of the response variables. We compare the performance of these HB estimators to the area level empirical best linear unbiased prediction (EBLUP) estimator and to the industry-standard post-stratified (PS) direct estimator. Results suggest that when borrowing strength to areas which are believed to be homogeneous (such as the ecosection level) and a weakly informative prior distribution is placed on the between-area variation parameter, we can reduce variance substantially compared the analogous EBLUP estimator and the PS estimator. Explorations of bias introduced with the HB estimators through comparison with the PS estimator indicates little to no addition of bias. These results illustrate the applicability and benefit of performing small area estimation of forest attributes in a HB framework, as they allow for more precise inference at the ecosubsection level.

Keywords: forest inventory, empirical best linear unbiased prediction, remote sensing, post-stratification, indirect estimation, probabilistic graphical model, weakly informative priors, ecoregion

OPEN ACCESS

Edited by:
Annika Kangas, Natural Resources Institute Finland (Luke), Finland

Reviewed by:
Göran Ståhl, Swedish University of Agricultural Sciences, Sweden
Andrew Finley, Michigan State University, United States

***Correspondence:**
Kelly S. McConville
mcconville@reed.edu

Specialty section:
This article was submitted to Forest Management, a section of the journal Frontiers in Forests and Global Change

Received: *03 August 2021*
Accepted: *25 October 2021*
Published: *15 December 2021*

Citation:
White GW, McConville KS, Moisen GG and Frescino TS (2021) Hierarchical Bayesian Small Area Estimation Using Weakly Informative Priors in Ecologically Homogeneous Areas of the Interior Western Forests. Front. For. Glob. Change 4:752911. doi: 10.3389/ffgc.2021.752911

1. INTRODUCTION

The USDA Forest Service Forest Inventory and Analysis Program (FIA) collects a sample of inventory data nationwide to monitor status and trends in forested ecosystems at scales relevant for strategic-level planning. Increasingly, this network of valuable inventory plots is being called upon to answer questions relevant to forest land management which is below the spatial and temporal scales for which the sample was originally designed. Information is needed on resources lost and recovery rates within disturbance boundaries, on significant change in carbon sources and sinks, as well as on the state of the forests within individual counties, districts, or other small management

units. There is strong interest in exploring methods to integrate extant inventory data with remotely sensed data through models that expand the capacity to estimate forest attributes over smaller domains in space and time.

A standard estimator that combines inventory data with remotely sensed data is the generalized regression estimator (Cassel et al., 1976) in which the inventory data are modeled and predicted over the domain of interest and then the observed data and predictions are aggregated to construct an estimator. Post-stratification (PS), a common estimation technique for national forest inventories (NFI) such as FIA, is a special case of the generalized regression estimator that incorporates a single, categorical auxiliary variable into the estimator (Särndal et al., 1992). Since the generalized regression estimator only makes use of data within the domain of interest, it is called a *direct estimator*. And, although leveraging auxiliary data typically improves the precision of a direct estimator, it tends to still not achieve adequate levels of precision when the sample size of the inventory data in the domain is small (Rao and Molina, 2015). Therefore, we consider here *indirect estimators* with their defining characteristic of borrowing strength from data outside the domain of interest. These domains are often classified as *small areas* and we will use the terms *domain* and *small area* interchangeably. When the indirect estimator explicitly relies on a model to link the data in the desired small area with data in other related small areas it is called a *small area estimator*. These linking models can be built either at the area level or unit (i.e., plot) level, depending on data availability and the strength of the relationships between the inventory data and remote sensing data at these two resolutions. We study area level models here because the inventory and remotely-sensed data we consider have strong linear trends at the area level and violate normality assumptions at the unit level. These estimators are constructed under either a frequentist framework where the quantities of interest are fixed, unknown values or a Bayesian framework where they are considered random variables. Key advantages of the Bayesian approach are that it allows the modeler to directly consider uncertainty between the small areas and to obtain distributions, not just point estimates and standard error estimates, for the parameters of interest.

A frequently utilized, and frequentist-based, indirect estimator is the empirical best linear unbiased prediction (EBLUP) estimator, which uses a linking model with random area-specific effects to borrow strength from related areas (Rao and Molina, 2015). The suitability of area and unit level EBLUP estimators to the small area applications found in NFIs have been studied extensively (Goerndt et al., 2011; Breidenbach and Astrup, 2012; Magnussen et al., 2017; Mauro et al., 2017; Coulston et al., 2021). This paper considers the Bayesian analog to the EBLUP, a hierarchical Bayesian (HB) estimator. These HB estimators are not commonly used in forest inventory research; however, they have been applied in a variety of other application areas ranging from poverty mapping to agriculture to transportation to employment (You et al., 2003; Vaish et al., 2010; Wang et al., 2012; Molina et al., 2014) to name a few. Within the NFI literature, Ver Planck et al. (2018) explored an area level HB estimator for estimating forest attributes and did find improvements in precision over the Horvitz-Thompson (HT) direct estimator.

In this paper, we explore the performance of the PS, the area level EBLUP, and the area level HB estimators at estimating the mean value of four response variables: basal area (m^2 per hectare), count of trees per hectare, above-ground biomass (kg per hectare), and net volume of trees (m^3 per hectare), excluding rotten or form defects, across the Interior West (IW) of the US. We generate estimates within the subregion (ecosubsection) of a hierarchical system of ecological divisions. For both the EBLUP and HB approaches, we consider the impact of borrowing strength from two resolutions from upper hierarchical levels: ecosection and ecoprovince, at scales of thousands of acres and millions of acres, respectively. Leveraging the flexibility of the HB, we study the impacts of varying how prior information on the homogeneity of the modeled small areas is incorporated into the estimator. We find that when borrowing strength to the ecosection level and including weakly informative prior information about small area homogeneity with the area level HB estimator we can reduce variance substantially compared to other common estimators. Explorations of potential bias through comparison with the post-stratified estimator display almost no introduction of bias with this estimator. However, since we do not know the true mean of the response variable of interest, caution is warranted when making strong conclusions about bias.

2. METHODS

2.1. Region of Study and FIA Data

This manuscript focuses on estimating the mean of several key forest attributes for the ecosubsections in the IW region of the United States (**Figure 1**), which encompasses the states of Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming. The inventory data were collected by FIA using a geographically-based systematic sampling design, where each plot represents about 2,500 ha of land (Bechtold and Patterson, 2005) and cover a 10 year measurement cycle from 2007 to 2017. This sample of 86,065 inventory plots were downloaded on February 6, 2019 from the FIA database, version FIADB_1.8.9.99 (last updated Dec 3, 2018). Our analyses include the use of four variables from the FIA database as response variables: basal area (m^2 per hectare), count of trees per hectare, above-ground biomass (kg per hectare), and net volume (m^3 per hectare). For remotely-sensed auxiliary variables, we consider a forest/non-forest classification used for post-stratifying in the IW (Blackard et al., 2008) and the 2016 National Land Cover Database percent tree canopy cover map (Yang et al., 2018), which has a spatial resolution of 30 m. Although each FIA plot consists of 4 subplots, response variables represent the aggregation of information at the plot level, and just the FIA plot center was intersected with the two auxiliary data layers. As input to the area-level estimators described below, both the percent canopy cover and proportions of forest and non-forest classes are averaged to the small area level for the IW. FIA data retrievals and processing of auxiliary data were done through the R package `FIESTA` (Frescino et al., 2015).

Borrowing strength in small area applications often occurs using political boundaries, such as counties within a state. But in

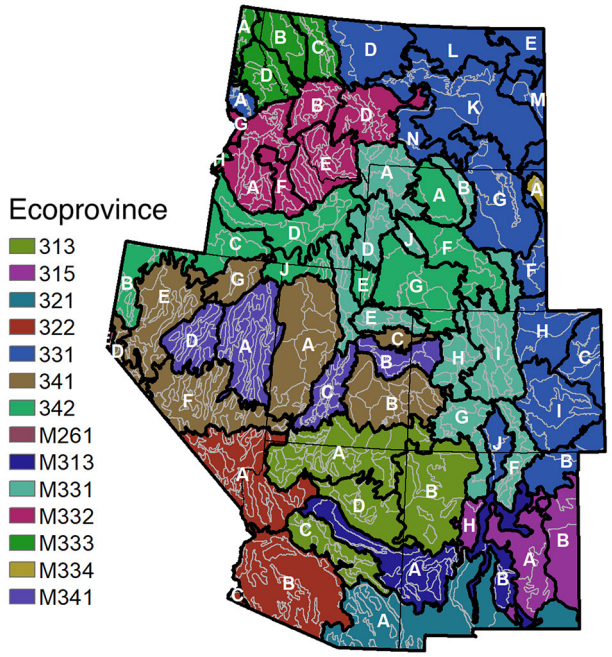


FIGURE 1 | To illustrate borrowing strategies, assume our target areas of interest are the ecosubsections within the IW-RMRS. Additional ecosections can be borrowed for those ecosubsections at the ecosectional level (shaded with the same color and denoted by the same letter) or borrowed at the ecoprovincial level (shaded with the same color). In the case of ecosectional borrowing, separate models are fit for the targeted ecosubsections. In the case of ecoprovincial borrowing, only one model is fit for these same ecosubsections.

order to borrow from "similar" domains, there is an opportunity to use any one of many classification systems that exist in the US to help guide borrowing in a more ecologically sensible fashion. As one prominent example, Cleland et al. (2007) delineated ecological units across the conterminous US using biological and physical information such as potential natural vegetation, geology, soils, climate, and hydrology. These ecological units were developed in a nested hierarchical structure. Ecoprovinces identify major vegetation cover types and land forms. Ecosections delineate more homogeneous areas within the ecoprovinces based on more detailed physical and biological components of the environment. Ecosubsections provide another step toward homogeneity at an even finer scale, with the number of FIA plots in each ecosubsection ranging from 1 to 2,200 in the Interior West. **Figure 1** illustrates how the nested hierarchical structure of this ecological classification system facilitates borrowing at different ecological scales. We investigate how borrowing strategies affect the performance of the indirect estimators by comparing estimates and standard errors of the area level estimators applied to ecosubsections when borrowing occurs at the ecoprovincial vs. ecosectional levels.

2.2. Estimators

We consider the PS estimator, which is a direct estimator, and two indirect estimation approaches, the EBLUP and HB, based on an area level, linear mixed model. For the HB method, we explore four different estimators, which vary based on how prior knowledge is incorporated to see how that impacts the estimator's precision and bias. All data analysis is conducted using the statistical software package `R` (R Core Team, 2020). In particular, the PS estimator is fit with the `mase` package (McConville et al., 2018), the EBLUP estimators are fit with `sae` (Molina and Marhuenda, 2015), and the HB estimators are fit with `mcmcsae` (Boonstra, 2021).

In order to explore these estimators in depth, we now introduce relevant notation. First, suppose we have m small areas we wish to estimate. Next, the indices are as follows: i indexes over units sampled; j indexes over small areas (in our case, ecosubsections); and k indexes over post-strata. Now, recall the goal of producing estimates of the mean of some response variable y, such as trees per hectare, in a small area. So, let μ_{y_j} be the population mean of the study variable in ecosubsection j in the IW. To denote the estimator produced for μ_{y_j} we use $\hat{\mu}_{y_j}$ with a superscript denoting which estimator is being used. We also use $\hat{V}(\hat{\mu}_{y_j})$ to denote the estimator of the variance of $\hat{\mu}_{y_j}$. The set s_j of size n_j includes all units sampled within ecosubsection j. We use the shorthand "iid" when referring to independent and identically distributed random variables and "ind" for independent random variables.

2.2.1. Direct Estimation via Post-stratification

We implement the PS estimator, which is commonly used by FIA and other NFIs, and is considered a direct estimator of μ_{y_j} since it only uses the inventory and auxiliary data within ecosubsection j. With the set of weights, $\{w_{jk}\}_{k=1}^{K}$, representing the proportion of pixels in each post-stratum for ecosubsection j, the PS estimator of μ_{y_j} is represented as follows:

$$\hat{\mu}_{y_j}^{PS} = \sum_{k=1}^{K} w_{jk} \hat{\mu}_{y_{jk}}^{HT} \tag{1}$$

and is a weighted average of the post-strata HT estimators, given by $\hat{\mu}_{y_{jk}}^{HT} = n_{jk}^{-1} \sum_{i \in s_{jk}} y_i$ where s_{jk} is the subset of the sample in ecosubsection j that falls in post-stratum k and n_{jk} is the corresponding sample size (Särndal et al., 1992). Since we have equal probability sampling, the post-strata HT estimators equal the post-strata sample means. For the IW, ignoring adjustments for non-response, the post-strata classes are forest and non-forest so $K = 2$. Post-stratification can certainly be conducted using more than 2 classes (e.g., Rintoul et al., 2020) but here we applied post-strata consistent with that used in the IW production inventory processes.

The variance estimator for $\hat{\mu}_{y_j}^{PS}$ is given by:

$$\hat{V}\left(\hat{\mu}_{y_j}^{PS}\right) = \frac{1}{n_j}\left(\sum_{k=1}^{K} w_{jk} n_{jk} \hat{V}\left(\hat{\mu}_{y_{jk}}^{HT}\right) + \sum_{k=1}^{K} (1 - w_{jk}) \frac{n_{jk}}{n_j} \hat{V}\left(\hat{\mu}_{y_{jk}}^{HT}\right)\right) \tag{2}$$

(Equation 7.6.6 in Särndal et al., 1992 without the finite population correction) where the HT variance estimator of $\hat{\mu}_{y_{jk}}^{HT}$ is given by

$$\hat{V}\left(\hat{\mu}_{y_{jk}}^{HT}\right) = \frac{1}{n_{jk}(n_{jk}-1)} \sum_{i \in s_{jk}} \left(y_i - \hat{\mu}_{y_{jk}}^{HT}\right)^2. \quad (3)$$

Since the number of pixels for the post-strata map is significantly larger than the sample size, the finite population correction would be negligible and is therefore omitted from the variance estimator calculation. For the IW, the PS estimator is typically more efficient than the HT estimator since many of the desired response variables are more homogeneous within the forest/non-forest post-strata. [For the forest/non-forest used here, Blackard et al. (2008) report an accuracy of 91% correctly classified based on an independent test set, with errors of omission and commission for forest at 17 and 18%, respectively, and for non-forest at 7 and 6%, respectively]. Therefore, we use the PS estimator in the subsequent indirect estimators when a direct estimator is needed.

2.2.2. Indirect Estimation via Small Area Models

When the sample sizes in the domains of interest are small, direct estimation techniques often do not provide sufficiently small variances, even with the use of auxiliary data, to make informative inferences. Indirect estimators increase the effective sample size by borrowing strength from data outside, with greater gains made when the larger area has the same characteristics, in terms of the response variables and their relationships with the auxiliary data, as the small areas of interest.

One common technique for borrowing strength is to explicitly use a linking model with a random-area specific effect, in addition to the sampling model which describes the data generation. Combining the linking model and the sampling model results in a mixed model approach to estimating the parameters of interest in the small areas. We consider a linear mixed model, which can be estimated using the EBLUP or using HB when additional assumptions are made on model parameters.

2.2.2.1. The Area Level EBLUP Estimator

For our parameters of interest, μ_{y_j}, we assume the following linking model:

$$\mu_{y_j} = \beta_o + \beta_1 \bar{X}_j + \nu_j \quad (4)$$

where $\bar{X}_j$ is the average percent tree canopy cover for ecosubsection j and the area-specific random effects satisfy the following conditions:

$$\nu_j \overset{\text{iid}}{\sim} \text{N}(0,\ \sigma_\nu^2).$$

And, we assume the PS estimators were generated from the following data generation model:

$$\hat{\mu}_{y_j}^{PS} = \mu_{y_j} + \epsilon_j \quad (5)$$

where $\epsilon_j \overset{\text{ind}}{\sim} N(0, \sigma_j^2)$. Inserting Equation (4) into Equation (5) gives the following area level mixed model, also known as the Fay-Herriot model (Fay and Herriot, 1979):

$$\hat{\mu}_{y_j}^{PS} = \beta_o + \beta_1 \bar{X}_j + \nu_j + e_j \quad (6)$$

where

$$\nu_j \overset{\text{iid}}{\sim} \text{N}(0,\ \sigma_\nu^2), \quad e_j \overset{\text{ind}}{\sim} \text{N}(0, \sigma_j^2), \quad \text{and} \quad \nu_j \perp\!\!\!\perp e_j.$$

To obtain an estimator of μ_{y_j} from this model, we use an EBLUP approach. This requires estimating the within-area and between area variances and the model coefficients. For $j = 1, 2, \ldots, m$, the within-area variations, σ_j^2 are set to $\hat{V}\left(\hat{\mu}_{y_j}^{PS}\right)$, the estimated variances of the PS estimates. The between-area variation, σ_ν^2, is estimated using a method of moments estimator (Ch 6.1.2 in Rao and Molina, 2015) and the estimated model coefficients $\hat{\beta}_o$ and $\hat{\beta}_1$ are the EBLUPs of β_o and β_1, respectively. The equation for the variance estimator of the EBLUP estimator of μ_{y_j} is given in the **Appendix**.

The EBLUP estimator of μ_{y_j} can be expressed as a weighted average of the direct estimator and an area level regression-synthetic estimator:

$$\hat{\mu}_{y_j}^{EBLUP} = \hat{\gamma}_j \hat{\mu}_{y_j}^{PS} + (1 - \hat{\gamma}_j)(\hat{\beta}_o + \hat{\beta}_1 \bar{X}_j) \quad (7)$$

where

$$\hat{\gamma}_j = \frac{\hat{\sigma}_\nu^2}{\hat{V}\left(\hat{\mu}_{y_j}^{PS}\right) + \hat{\sigma}_\nu^2}. \quad (8)$$

Notice that the EBLUP estimator is a composite of an indirect and a direct estimator where the weighting term accounts for local variation. In particular, $\hat{\gamma}$ is the ratio of between-area variation and total variation. When the small areas are fairly heterogeneous, the EBLUP will rely more heavily on the direct, PS estimator, which only relies on data within the small area of interest. The estimator leans more on outside information when the variance estimator of the PS estimator is large compared to the variability between the small areas. In this case, it relies on the fixed effect component of the estimated regression line, which is called a regression-synthetic estimator.

2.2.2.2. The Area Level Hierarchical Bayesian Estimator

So far, we have explored common frequentist approaches to small area estimation. However, the primary focus of this paper is to study the performance of the HB for small area estimation. Under the Bayesian paradigm, the parameter of interest, μ_{y_j}, and other model parameters, are treated as random variables instead of fixed, unknown values. Leveraging Bayes' Theorem, this technique synthesizes information gained from the data via a likelihood function with prior knowledge about the parameter of interest and model parameters to obtain a posterior distribution for the parameters:

$$P(\mu_{y_j}, \beta_o, \beta_1, \sigma_\nu^2 \mid \text{data}) \propto P(\text{data} \mid \mu_{y_j}, \beta_o, \beta_1, \sigma_\nu^2) \cdot P(\mu_{y_j}, \beta_o, \beta_1, \sigma_\nu^2) \quad (9)$$

A marginal posterior distribution for μ_{y_j} is found by integrating out the model parameters or by Markov chain Monte Carlo (MCMC) methods. Typically the posterior mean of

the distribution, $\mathrm{E}\left[\mu_{y_j} \mid \text{data}\right]$, serves as the estimator of the parameter, with precision provided by the posterior variance, $\mathrm{Var}\left[\mu_{y_j} \mid \text{data}\right]$.

For the area level HB estimator, we start with Equation (6), as was done for the area level frequentist EBLUP, and apply a HB approach. This transformation involves rewriting the data generation model, also referred to as the likelihood function, as a conditional normal distribution where we condition on the parameter of interest and model parameters:

$$\hat{\mu}_{y_j}^{PS} \mid \mu_{y_j},\ \beta_o,\ \beta_1,\ \sigma_v^2 \sim \mathrm{N}\left(\mu_{y_j},\ \hat{V}\left(\hat{\mu}_{y_j}^{PS}\right)\right)$$

and the distribution of μ_{y_j} as a conditional normal distribution where we condition on the model parameters:

$$\mu_{y_j} \mid \beta_o,\ \beta_1,\ \sigma_v^2 \sim \mathrm{N}(\beta_o + \bar{X}_j\beta_1,\ \sigma_v^2).$$

The HB approach also requires specifying prior distributions for $\beta_o,\ \beta_1$, and σ_v^2. For the model coefficients, we assume a flat prior:

$$f(\beta_o, \beta_1) \propto 1.$$

For the between-area variation parameter, we consider two prior distributions, an uninformative improper uniform distribution:

$$f(\sigma_v^2) \propto 1$$

and a unit-scale half-Cauchy distribution:

$$\sigma_v \sim \text{half-Cauchy}(\text{scale} = 1).$$

Note that the half-Cauchy distribution is applied to the between-area standard deviation, not the between-area variance. Lastly, we assume the model parameters are independent, namely, $f(\beta_o, \beta_1, \sigma_v^2) = f(\beta_o)f(\beta_1)f(\sigma_v^2)$.

Now that the HB model has been specified, we can attain the small area estimator and variance estimator. For the estimator in ecosubsection j, the Bayes estimator for μ_{y_j} is:

$$\hat{\mu}_{y_j}^{HB} = E[\mu_{y_j} \mid \hat{\mu}_{y_j}^{PS}]. \tag{10}$$

For the variance in ecosubsection j, the variance of the posterior distribution is used:

$$\hat{V}(\hat{\mu}_{y_j}^{HB}) = V(\mu_{y_j} \mid \hat{\mu}_{y_j}^{PS}). \tag{11}$$

The estimator and variance estimator are obtained through MCMC methods with the `mcmcsae` R package (Boonstra, 2021). Using MCMC methods allow for posterior distributions to be well-approximated by sampling from a probability distribution. We use 1,000 sampling iterations (the length of each Markov Chain), 3 Markov Chains, and a burn-in period length of 250 to obtain the results of each HB model we fit.

Figure 2 represents the area level HB estimator as a probabilistic graphical model (PGM). This diagrammatic view can be helpful in understanding the relationships between the parameter of interest, the data, model parameters, and other key random variables included in the model. For example, the arrows in **Figure 2** can give us the distribution for $\hat{\mu}_{y_j}^{PS}$ and show us that it depends on the parameter of interest (μ_{y_j}) and model parameters (β_o, β_1, and σ_v^2). Not only can we quickly see how distributions are conditioned through the use of a PGM, we can also less formally view how variables are related to each other and gain a deeper understanding of how strength is borrowed for this area level HB estimator. If we remove the formality of some parameters representing random variables, we can even use **Figure 2** to visualize how strength is borrowed with the area level EBLUP. Recall that the area level EBLUP is specified with the same linking model and thus strength is borrowed from the same places. Thus, **Figure 2** not only depicts the components of the area level HB model, but also the area level EBLUP, albeit in a less formal way.

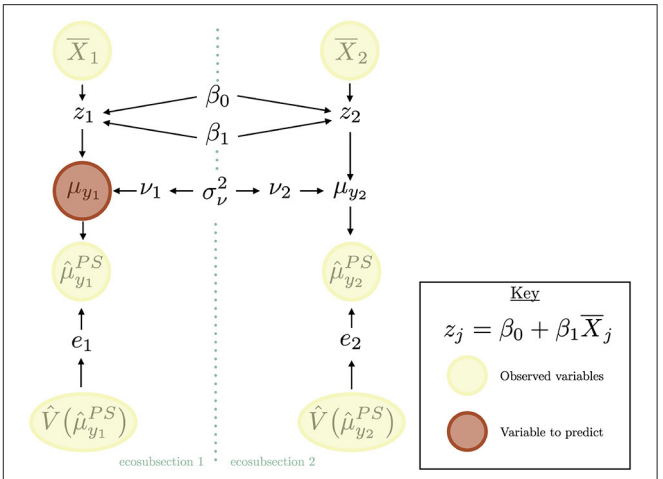


FIGURE 2 | The area level HB estimator depicted as a PGM, which helps us understand the relationships between the data, model parameters, parameter of interest, and other key random variables. Quantities in yellow circles represent observed (known) variables. The quantity in red represents the quantity we would like to estimate. In this diagram, the quantity we would like to estimate is the mean of the response variable in the first ecosubsection. The seafoam dotted line represents the division between ecosubsection 1 and ecosubsection 2. Quantities on this dotted line represent random variables that effect both ecosubsections. The arrows represent marginal and conditional distributions of random variables.

2.3. Methods Summary

We use seven estimators–the PS estimator, two area level EBLUPs, and four area level HB estimators–to produce estimates for the average of basal area (m^2 per hectare), tree count per hectare, above-ground biomass (kg per hectare), and net volume (m^3 per hectare). The EBLUPs and HB estimators use one explanatory variable, the average percent tree canopy cover of the ecosubsection, to produce estimates. Estimation occurs at the ecosubsection level, and thus we have produced 11,928 estimates (seven estimators, four response variables, and 426 ecosubsections). The model-based estimators are fit either within an ecoprovince or an ecosection, and hence each ecosubsection only borrows strength out to either the ecoprovince or ecosection

level, not the entire IW region. In order to assess the quality of these estimators, we summarize the findings over the entire study region and for particular regions.

The data span the entire IW; however, we are forced to exclude a small portion of ecosubsections from our analyses. These ecosubsections contain either no or very close to no sampled areas with non-zero values for the variables of interest: that is, areas which are in extremely non-forested areas. These areas have to be excluded due to their within-area variance being zero or so close to zero that the software does not recognize that the number was positive.

3. RESULTS

3.1. Estimator Performance

The indirect estimators that we implement perform similarly, on average, to the PS estimator. **Figure 3** displays each indirect estimator's estimate on the y-axis and the PS estimate on the x-axis for the basal area response variable and **Figure 4** does so for the count per hectare response variable. Notably, **Figure 3** shows a strong linear relationship between the indirect and direct estimates (which is also observed for volume and biomass) whereas in **Figure 4** this linear relationship begins to deteriorate for larger values of average canopy cover. This is due to the relationship between the explanatory variable (average canopy cover) and the PS estimator of the average tree count per hectare exhibiting more variability for those larger values, violating the model assumption of homoskedasticity. **Figure 5** displays this larger variability for the tree count per hectare variable and showcases that the PS estimates consistently fall below the regression line for the largest average canopy cover values. This violation of the homoskedasticity assumption seems to have introduced bias into our indirect estimates. This represents a good cautionary tale that while indirect estimators can provide significant reductions in variance, they can be biased when the model is incorrectly specified. For the remainder of the paper, we focus on basal area, where the linear model specification seems most appropriate.

Figures 3, **4** also display that the flat prior HB estimator produces very similar estimates to the EBLUP for both the estimators that borrow strength out to the ecosection level and to the ecoprovince level. **Figure 6** displays this relationship in further detail. Notably, the flat prior HB estimates and standard errors are very similar to the EBLUP. This is expected as we add no prior information to the flat prior HB estimators. By adding no information and specifying the same model we should and do see extremely similar results.

While it is reassuring for the flat prior HB estimator to reinforce the results of the EBLUP, the full benefits of the HB estimators are not gained without careful thought into how prior information is incorporated. In our case, we specify a half-Cauchy prior with scale of one, which is considered a weakly informative prior, on the between-area variation parameter. This distribution places more probability mass over smaller values for our between-area variation, signifying that we expect the between-area variation to be low. This prior is commonly used for the between-area standard deviation parameter in hierarchical models, especially when the number of small areas

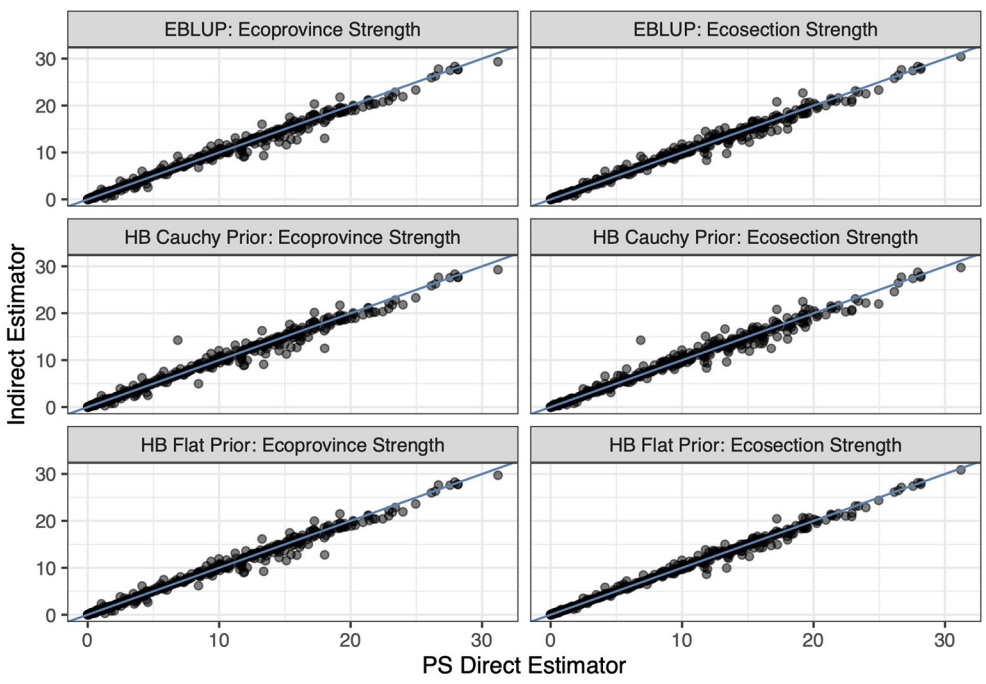


FIGURE 3 | The six indirect estimators compared to the PS direct estimator. Each point represents two estimates of basal area for an ecosubsection, its y-coordinate representing the indirect estimate of basal area for that ecosubsection and its x-coordinate representing the PS direct estimate of basal area. The blue line is the identity line.

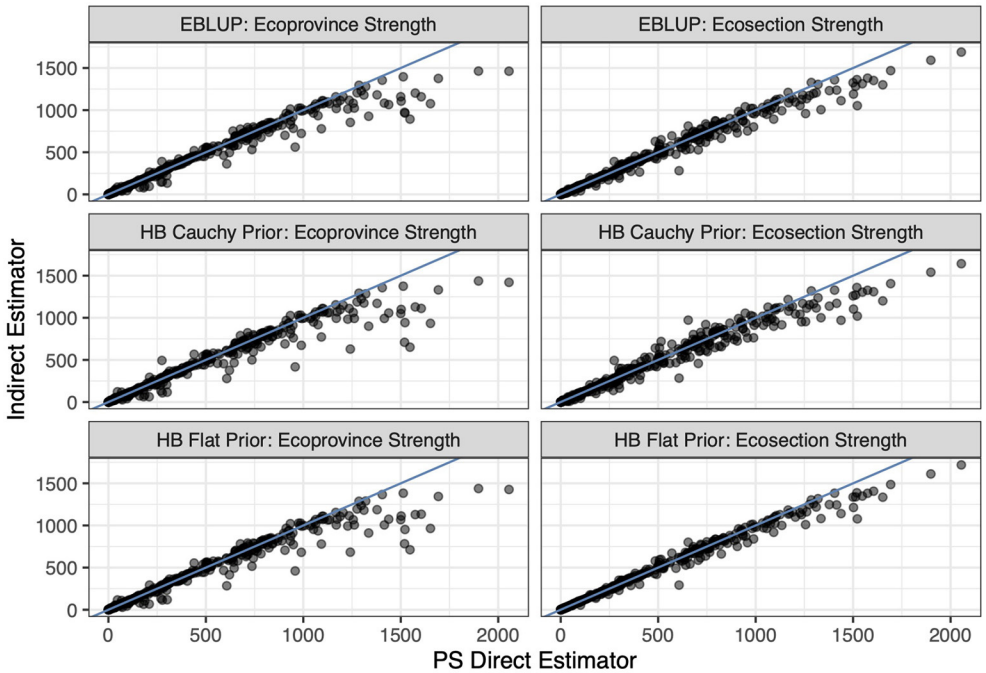


FIGURE 4 | The six indirect estimators compared to the PS direct estimator. Each point represents two estimates of tree count per hectare for an ecosubsection, its y-coordinate representing the indirect estimate of tree count per hectare for that ecosubsection and its x-coordinate representing the PS direct estimate of tree count per hectare. The blue line is the identity line.

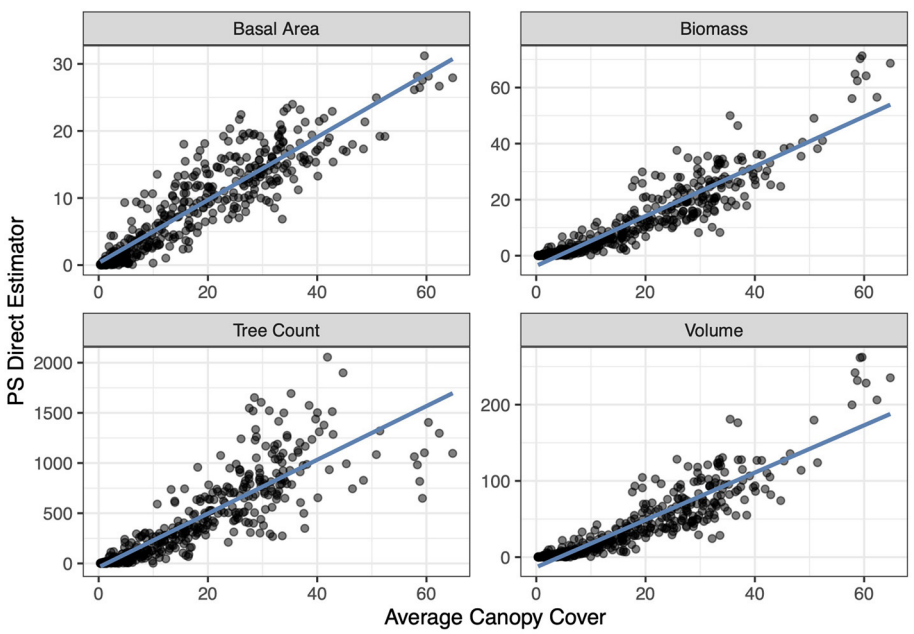


FIGURE 5 | The relationship between the response variables and the explanatory variable (average canopy cover) for each response variable at the ecosubsection level across the IW. The x-coordinate represents the population value of average canopy cover based on remotely sensed data in a given ecosubsection, and the y-coordinate represents the post-stratified estimate of a response variable in a given ecosubsection. The blue line is the ordinary least squares regression line.

is small and so the data provide little information about the group-level variance (Gelman, 2006). Further, we know that ecosections should be more homogeneous than ecoprovinces and this prior information should reinforce the homogeneity we see in the data. We still chose to place a half-Cauchy prior on the between-area variation when borrowing out to the

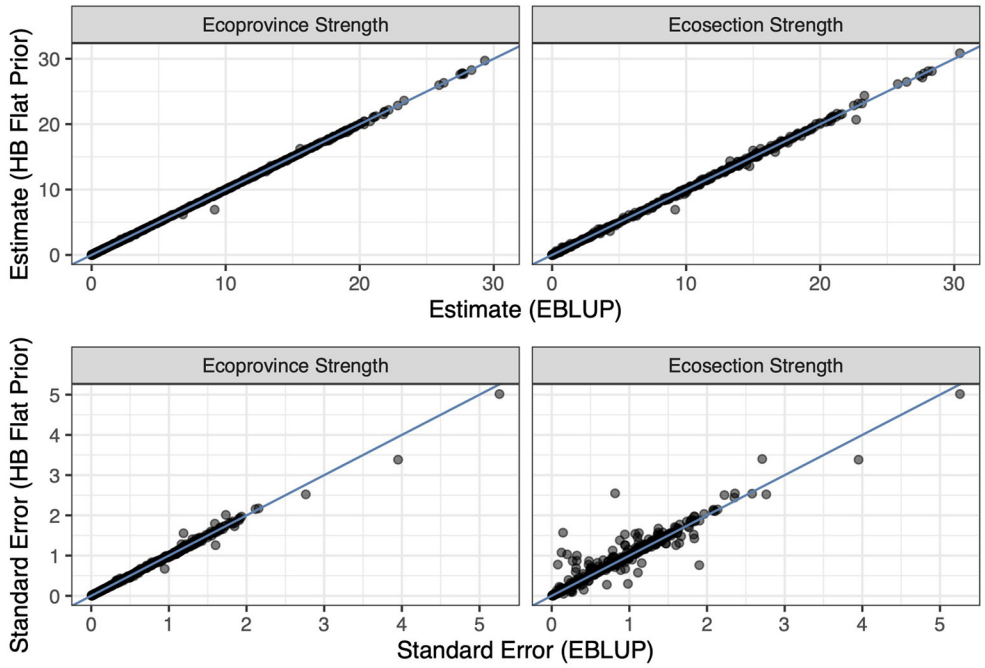


FIGURE 6 | The flat prior HB estimators compared with the EBLUP estimators at each level of strength. The top left plot displays the estimates for each estimator at the ecoprovince level, the top right plot displays the estimates for each estimator at the ecosection level, the bottom left plot displays the standard errors for each estimator at the ecoprovince level, and the bottom right plot displays the standard errors for each estimator at the ecosection level. Each plot contains either estimates or standard errors for the basal area response variable. The blue line in each plot is the identity line.

ecoprovince level, as these regions are defined by ecologists as more homogeneous than the rest of the study region (McNab et al., 2007).

Figure 7 displays the reduction in variance when we change the prior on the between-area variation from flat to half-Cauchy in both the ecosection and ecoprovince approaches. The variance is reduced much more significantly when we use a half-Cauchy prior for estimators that borrow strength to the ecosection level because the number of small areas is smaller. In particular, one can observe that most areas where variance is reduced a large amount have less ecosubsections that they borrow strength from (light purple dots).

Outside of this graphical representation, we can look numerically at the mean and median percent reduction in variance when moving from a flat prior HB estimator to one with the half-Cauchy prior. **Table 1** displays both the mean and median percent reduction in variance of basal area for the ecosection and ecoprovince level HB estimators. Borrowing to the more homogeneous ecosection level with the half-Cauchy prior on the between-area variation leads to the greater reductions in variance. While this reduction in variance is compelling, it is possible that the weakly informative prior introduced bias to the estimator.

To understand where bias may be introduced in our estimates, **Figure 8** displays the estimates made by the HB estimators with a half-Cauchy prior compared to those made by the PS estimator, an estimator that is unbiased under resampling regardless of model accuracy. Here, we see a high level of agreement between the two estimators, which suggests the HB estimators are not systematically biased. However, it is important to note that it is the PS estimator, under resampling, that is unbiased, not a given PS estimate. We also saw strong agreement between the estimators from the half-Cauchy prior and those from the flat prior, signifying a robustness to the choice of prior distribution for the between-area variation.

We can also investigate indications of bias numerically, with the percent relative difference (PRD) metric. The PRD between two estimators is defined as followed:

$$PRD(\hat{\mu}_1,\ \hat{\mu}_2) = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\mu}_2} \cdot 100\%.$$

When we examine the PRD between the PS estimator and the half-Cauchy prior HB estimator for the basal area response variable we see that the average PRDs are −0.007% and 0.756% at the ecoprovince and ecosection level, respectively. The median PRDs between for these estimators are −0.12% and −0.225% at the ecoprovince and ecosection level, respectively. The low PRD values provide additional evidence that we are not introducing much systematic bias with the use of the auxiliary data and prior on the between-area variation parameter.

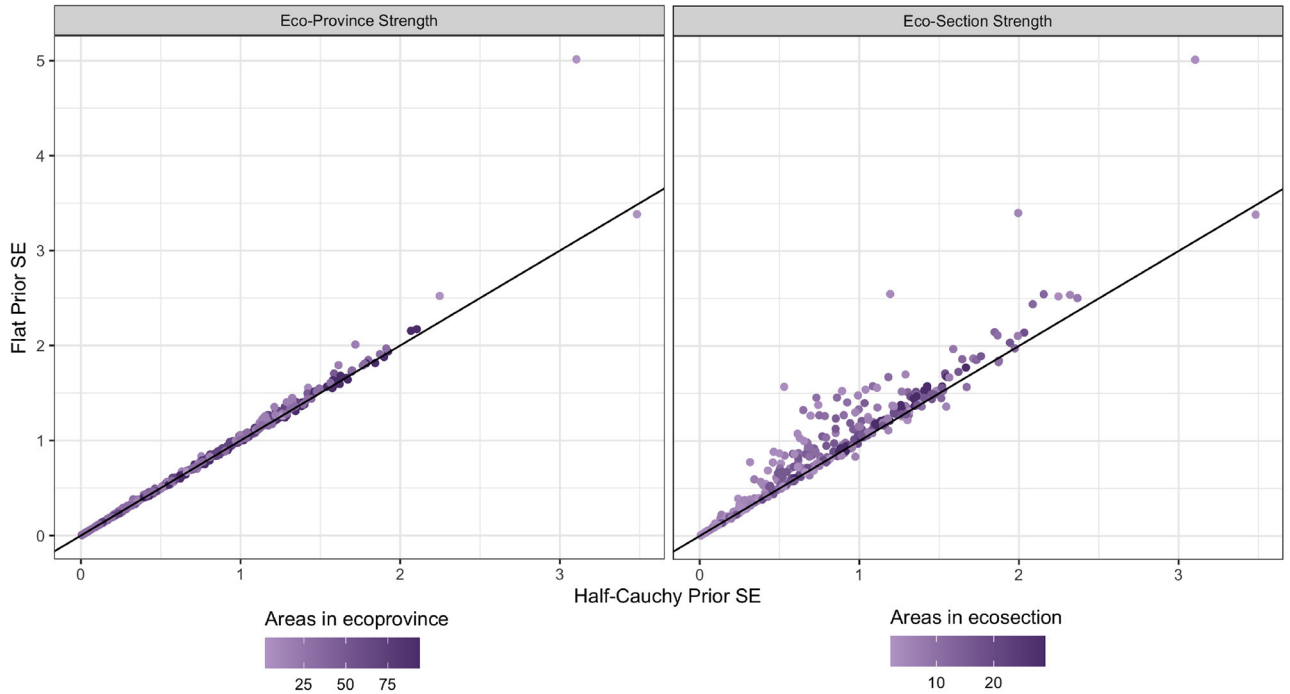


FIGURE 7 | Basal area standard errors for each HB estimator. The x-coordinate of each dot represents the standard error when a half-Cauchy prior is placed on between-area variation, and the y-coordinate represents the standard error when a flat prior is placed on between-area variation. The color of the dot represents the number of ecosubsections that strength was borrowed out to for the given dot. The black line is the identity line. The plot on the right shows the estimators that borrow strength to the ecosection level, while the plot on the left shows the ecoprovince level estimators.

TABLE 1 | Percent reduction in variance of basal area estimates from flat prior to half-cauchy.

Strength	Metric	Percent reduction
Ecosection	Mean	14.580
	Median	7.313
Ecoprovince	Mean	3.384
	Median	2.792

3.2. Case Study: The South Central Highlands (M331G) and the Utah High Plateau (M341C)

We now explore the effects of adding prior information to the HB estimators at a micro level: by examining two ecosections. The South Central Highlands and the Utah High Plateau ecosections both exist in mountainous ecoprovinces in the IW. **Figure 9** displays both ecosections. These two ecosections are located relatively close to each other in the IW, yet the addition of the half-Cauchy prior when we borrow strength to the ecosection level has a very different effect within each ecosection. To understand how the estimators perform differently across these two ecosections, we explore the mean estimates for basal area, and corresponding standard errors, within these two ecosections.

Figure 10 displays the HB estimates and illustrates that, for both ecosections, the basal area estimate is about the same for both priors on the between-area variation parameter. This again showcases a robustness to how the prior information is specified for between-area variation.

While the estimate values show high agreement, it should be noted that the standard error estimates changed more drastically when we changed the prior on between-area variation, as seen in **Figure 11**. Interestingly, the standard errors in eocsection M331G are reduced significantly when the half-Cauchy prior is used compared to the flat prior, while the standard errors in ecosection M341C hardly change. This is likely due to a couple of factors. First of all, the estimated variance of the PS estimates in ecosection M331G is lower than the estimated variance of the PS estimates in M341C (32.044 and 44.384, respectively). That is, based on the data, there is less between-area variation in ecosection M331G. By placing a prior which has high probability density for small values of σ_v^2 we have reinforced the pattern seen in the data. Additionally, M331G is borrowing from less small areas and therefore will lean more on the weakly informative prior which preferences smaller values for the between-area variation.

4. DISCUSSION

We consider six indirect, area level small area estimators and one direct estimator across the IW region of the United States. The two HB estimators with flat priors on between-area variation

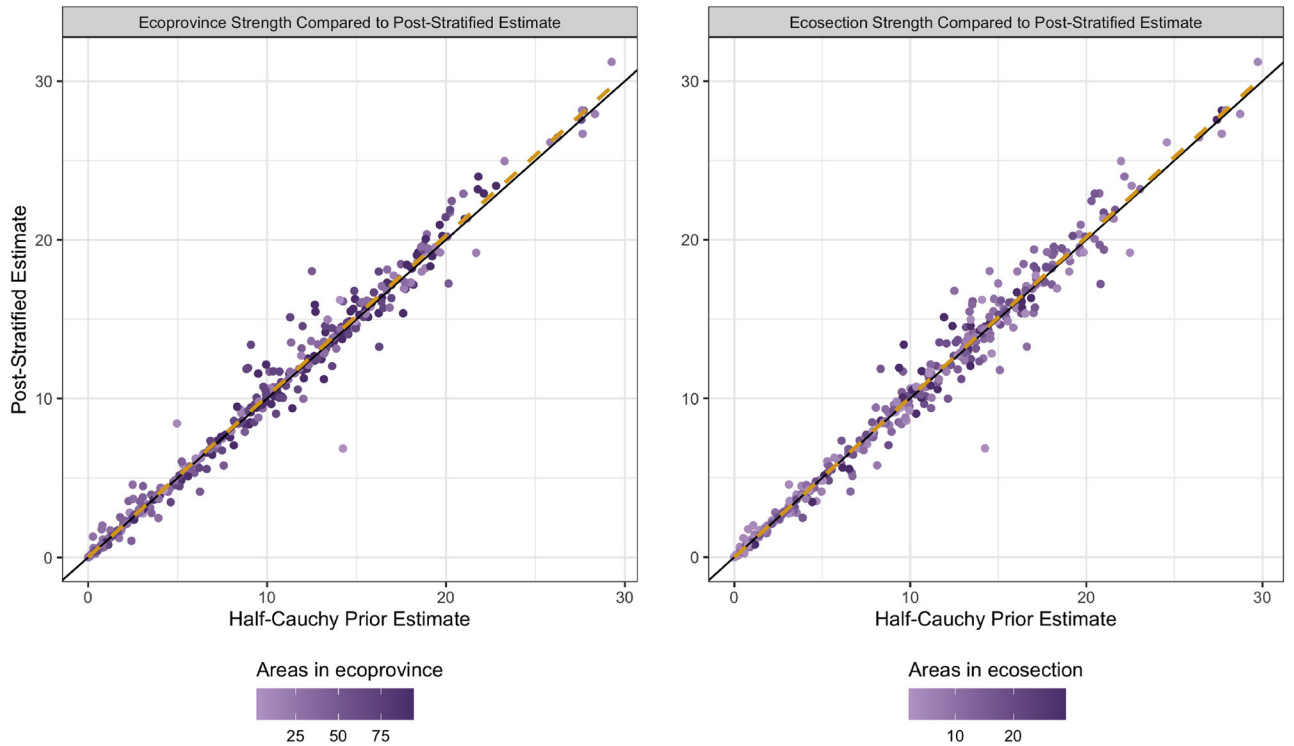


FIGURE 8 | Basal area estimates for the half-Cauchy HB estimator and post-stratified estimator. The x-coordinate of each dot represents the estimate when a half-Cauchy prior is placed on between-area variation, and the y-coordinate represents the post-stratified estimate. The color of the dot represents the number of ecosubsections that strength was borrowed out to for the given dot. The black line is the identity line. The dashed yellow line is the ordinary least squares regression line. The plot on the right shows the estimators that borrow strength to the ecosection level, while the plot on the left shows the ecoprovince level estimators.

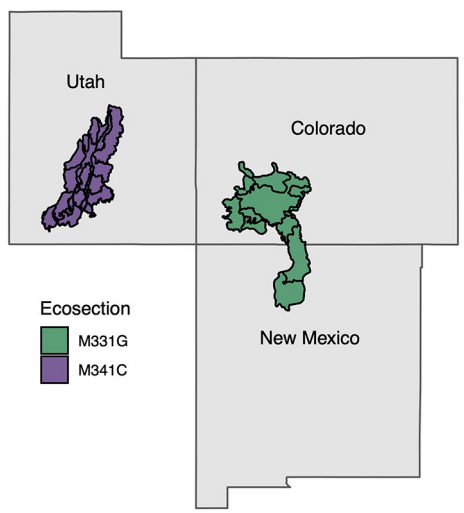


FIGURE 9 | The South Central Highlands (M331G) and the Utah High Plateau (M341C) ecosections on a map of the three states that they collectively occupy. The purple area represents the Utah High Plateau and the green area represents the South Central Highlands. The black lines divide the ecosubsection within each ecosection.

mimic the two analogous area level EBLUP estimators in both estimates and variances. When supplying the HB estimators with a half-Cauchy prior for the between-area variation parameter, we see a reduction in variance when borrowing strength out to both the ecoprovince and ecosection level, with more reduction observed at the latter level of strength.

Table 2 displays the relative efficiency of each estimator implemented in this article compared the standard HT direct estimator for the basal area response variable. We define relative efficiency of a given estimator as the variance estimator of that estimator divided by the variance estimator of a direct estimator. The first column of **Table 2** makes it clear that incorporating informative auxiliary data into a direct estimator, the PS estimator in this case, does improve its efficiency. These improvements mimic FIA's production process with just 2 post strata assigned at the plot level, not at the subplot level. However, greater gains can be had by moving to an indirect estimator. In particular, the HB estimator with a half-Cauchy prior on between-area variation borrowing strength to the ecosection level has the highest mean and median relative efficiency. Notably, the ecosection-level, half-Cauchy prior, HB estimators relative efficiency is greater than the ecoprovince-level, half-Cauchy prior, HB estimators relative efficiency. This gain in relative

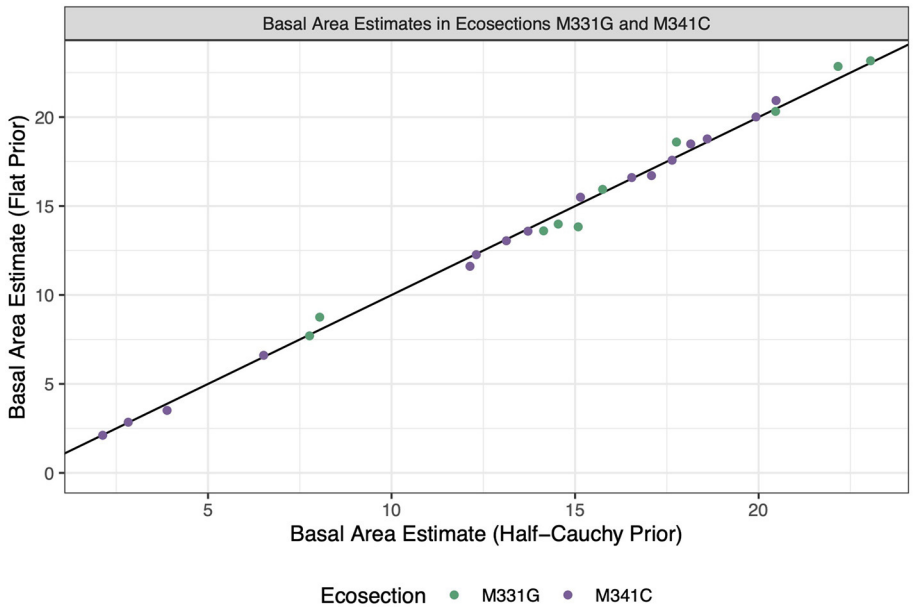


FIGURE 10 | Estimates for each HB estimator. The x-coordinate of each dot represents the estimate when a half-Cauchy prior is placed on between-area variation, and the y-coordinate represents the estimate when a flat prior is placed on between-area variation. The color of the dot represents the ecosection that a given ecosubsection is in. The black line is the identity line.

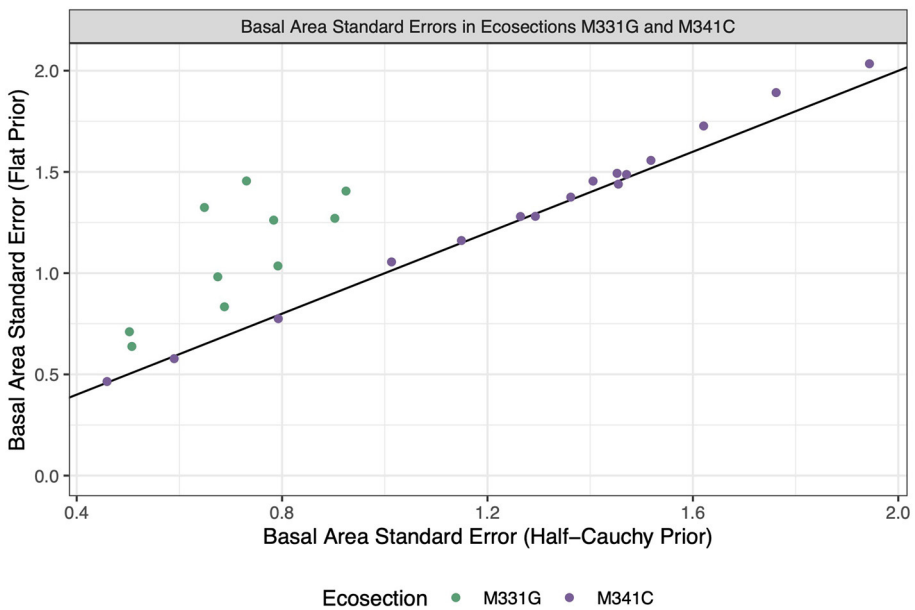


FIGURE 11 | Standard errors for each HB estimator. The x-coordinate of each dot represents the estimate when a half-Cauchy prior is placed on between-area variation, and the y-coordinate represents the estimate when a flat prior is placed on between-area variation. The color of the dot represents the ecosection that a given ecosubsection is in. The black line is the identity line.

efficiency is likely due to the half-Cauchy prior being a reasonable depiction of the between-area variation of ecosubsections within a given ecosection in the IW.

Figure 12 shows the relative efficiency of basal area estimates for both the ecosection- and ecoprovince-level half-Cauchy prior HB estimators compared to the PS direct estimator (as opposed to the Horvitz Thompson in **Table 2**). We can see more drastic improvements in relative efficiency in the more forested Northern parts of the IW, and the relative efficiency is sometimes below the PS estimator in extremely unforested areas.

TABLE 2 | Relative efficiency of each estimator compared to the Horvitz-Thompson (basal area response).

Metric	Post-strat	HB cauchy: ecosection	HB cauchy: ecoprovince	HB flat: ecosection	HB flat: ecoprovince	EBLUP: ecosection	EBLUP: ecoprovince
Mean	1.40	1.87	1.86	1.63	1.80	1.38	1.78
10% quantile	2.00	5.45	3.64	2.65	3.24	3.41	3.03
Median	1.31	1.80	1.74	1.55	1.70	1.59	1.67
90% quantile	1.01	1.10	1.17	1.07	1.15	0.99	1.13

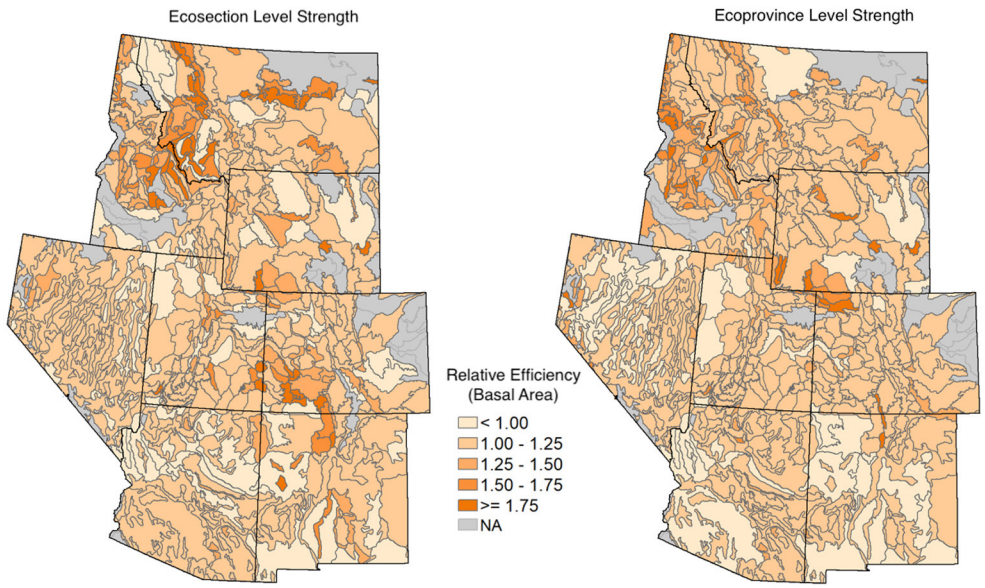


FIGURE 12 | The relative efficiency of basal area estimates from the HB estimators with a half-Cauchy prior on between-area variation compared to the PS direct estimator. On the left, we have the HB estimator that borrows strength out to the ecosection level, and on the right, we have the HB estimator that borrows strength out to the ecoprovince level. Darker orange areas correspond to higher relative efficiency. The gray areas are ecosubsections where we were unable to obtain HB estimates.

This might be due to artificially low variance estimates for the PS estimator, which can occur when almost all sampled units in an ecosubsection have values of 0 for the response variables. In the case of an estimator that borrows strength, such as the HB estimators, we will likely borrow strength to some areas that have larger direct estimates of response variables, giving us a larger variance.

The efficiency gains of the HB estimators with informative priors on between-area variation over the more common EBLUP and PS (see **Table 2** and **Figure 12**) imply that these estimators can attain the same level of precision but with less sampled plots. However, the benefits of a HB approach do not stop there. Conveniently, the Bayesian paradigm allows for more intuitive inferential statements than provided by frequentist methods. Since the Bayesian methods provide a distribution for our parameter of interest, we can make probabilistic statements about the location of the parameter, whereas the frequentist approach only allows us to talk about the behavior of our method under repeated sampling.

Considering the performance and all the characteristics of these estimators, the results of this work provide some guidance on when to consider which of these estimators. If one only has a small number of areas that they borrow strength out to, and those areas are believed to have a good amount of homogeneity between them, a HB estimator with a half-Cauchy prior on the between-area variation might be preferred. On the other hand, if one has the ability to borrow strength to a large number of groups that may not be too homogeneous, keeping a flat prior on between-area variation should be considered. This suggests that the HB estimator that borrows to the ecosection level and uses the half-Cauchy prior on between-area variation may be a viable estimator for FIA applications. Further testing with alternative responses and auxiliary data in other parts of the country is warranted.

Further work will include investigations of the unit level HB estimator, particularly with an eye to handling non-Gaussian data. Researchers have explored unit level modeling of non-Gaussian data types, such as zero-inflated data (Krieg et al., 2016) and other non-Gaussian data (Parker et al., 2020a). In particular, Parker et al. (2020b) discusses the benefits of unit level models, both in terms of potential efficiency gains and incorporating various levels of spatial aggregations. We hope to investigate the utility of these unit level models in a forest inventory setting. At both the area and unit level we will also explore extensions to

the HB estimators through spatially structured variance models. Ver Planck et al. (2018) explores area level HB estimators with conditional autoregressive random effects and conditional autoregressive random effects with smoothed sampling variance and found that these spatially structured variance models can help reduce the variance of the estimator. We hope to explore these spatially structured variance models further and investigate how they perform with different prior information supplied.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: the data include confidential plot data, which can not be shared publicly. FIA data can be accessed through the FIA DataMart (https://apps.fs.usda.gov/fia/datamart/datamart.html). Requests for data used here or other requests including confidential data should be directed to FIA's Spatial Data Services (https://www.fia.fs.fed.us/tools-data/spatial/index.php).

AUTHOR CONTRIBUTIONS

GW, KM, GM, and TF: conceptualization. GW and KM: methodology and writing. GW: analysis. TF: data curation. GW and TF: data visualization. GW, KM, and GM: review and editing. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported by the USDA Forest Service, Forest Inventory and Analysis Program (via agreement 19-JV-11221638-112) and by Reed College.

ACKNOWLEDGMENTS

The authors would like to thank the USDA Forest Service, Forest Inventory and Analysis Program for the data.

REFERENCES

Bechtold, W. A., and Patterson, P. L. (2005). *The Enhanced Forest Inventory and Analysis Program–National Sampling Design and Estimation Procedures*. USDA Forest Service, Southern Research Station.

Blackard, J., Finco, M., Helmer, E., Holden, G., Hoppus, M., Jacobs, D., et al. (2008). Mapping US forest biomass using nationwide forest inventory data and moderate resolution information. *Remote Sens. Environ.* 112, 1658–1677. doi: 10.1016/j.rse.2007.08.021

Boonstra, H. J. (2021). *mcmcsae: Markov Chain Monte Carlo Small Area Estimation*. Available online at: https://CRAN.R-project.org/package=mcmcsae

Breidenbach, J., and Astrup, R. (2012). Small area estimation of forest attributes in the Norwegian National Forest Inventory. *Eur. J. Forest Res.* 131, 1255–1267. doi: 10.1007/s10342-012-0596-7

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* 63, 615–620. doi: 10.1093/biomet/63.3.615

Cleland, D. T., Freeouf, J. A., Keys, J. E., Nowacki, G. J., Carpenter, C. A., and McNab, W. H. (2007). *Ecological Subregions: Sections and Subsections for the Conterminous United States*. Available online at: https://www.fs.fed.us/research/publications/misc/73326-wo-gtr-76d-cleland2007.pdf

Coulston, J. W., Green, P. C., Radtke, P. J., Prisley, S. P., Brooks, E. B., Thomas, V. A., et al. (2021). Enhancing the precision of broad-scale forestland removals estimates with small area estimation techniques. *Forest. Int. J. Forest Res.* 94, 427–441. doi: 10.1093/forestry/cpaa045

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, 269–277. doi: 10.1080/01621459.1979.10482505

Frescino, T. S., Patterson, P. L., Moisen, G. G., and Freeman, E. A. (2015). FIESTA-an R estimation tool for FIA analysts," in *Forest Inventory and Analysis (FIA) Symposium 2015* (Portland, OR: US Department of Agriculture, Forest Service, Pacific Northwest Research Station), 72.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* 1, 515–534. doi: 10.1214/06-BA117A

Goerndt, M. E., Monleon, V. J., and Temesgen, H. (2011). A comparison of small-area estimation techniques to estimate selected stand attributes using LiDAR-derived auxiliary variables. *Can. J. Forest Res.* 41, 1189–1201. doi: 10.1139/x11-033

Hidiroglou, M., and You, Y. (2016). Comparison of unit level and area level small area estimators. *Survey Methodol.* 42, 41–61.

Krieg, S., Boonstra, H., and Smeets, M. (2016). Small-area estimation with zero-inflated data – a simulation study. *J. Off. Stat.* 32:963–986. doi: 10.1515/jos-2016-0051

Magnussen, S., Mauro, F., Breidenbach, J., Lanz, A., and Kändler, G. (2017). Area-level analysis of forest inventory variables. *Eur. J. Forest Res.* 136, 839–855. doi: 10.1007/s10342-017-1074-z

Mauro, F., Monleon, V. J., Temesgen, H., and Ford, K. R. (2017). Analysis of area level and unit level models for small area estimation in forest inventories assisted with LiDAR auxiliary information. *PLoS ONE* 12:e0189401. doi: 10.1371/journal.pone.0189401

McConville, K., Tang, B., Zhu, G., Cheung, S., and Li, S. (2018). *MASE: Model-Assisted Survey Estimation*. Available online at: https://cran.r-project.org/package=mase

McNab, W. H., Cleland, D. T., Freeouf, J. A., Keys, J. E., Nowacki, G. J., and Carpenter, C. A. (2007). *Description of "Ecological Subregions: Sections of the Conterminous United States*. United States Department of Agriculture. doi: 10.2737/wo-gtr-76b

Molina, I., and Marhuenda, Y. (2015). sae: an R package for small area estimation. *R J.* 7, 81–98. doi: 10.32614/RJ-2015-007

Molina, I., Nandram, B., and Rao, J. (2014). Small area estimation of general parameters with application to poverty indicators: a hierarchical bayes approach. *Ann. Appl. Stat.* 8, 852–885. doi: 10.1214/13-AOAS702

Parker, P. A., Holan, S. H., and Janicki, R. (2020a). Computationally efficient Bayesian unit-level models for non-Gaussian data under informative sampling. *arXiv [Preprint]. arXiv.2009.05642*

Parker, P. A., Janicki, R., and Holan, S. H. (2020b). Unit level modeling of survey data for small area estimation under informative sampling: a comprehensive overview with extensions. *arXiv [Preprint]. arXiv.1908.10488*.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/

Rao, J. N., and Molina, I. (2015). *Small Area Estimation*. Hoboken, New Jersey: John Wiley & Sons.

Rintoul, M. A., Maebius, S., Alvarado, E., Lloyd-Damnjanovic, A., Toyohara, M., McConville, K. S., et al. (2020). "An alternative post-stratification scheme to decrease variance of forest attribute estimates in the interior west," in *Celebrating Progress, Possibilities, And Partnerships: Proceedings of the 2019 Forest Inventory and Analysis (FIA) Science Stakeholder Meeting* (Asheville, NC: U.S. Department of Agriculture Forest Service, Southern Research Station), 268–276. Available online at: https://www.fs.usda.gov/treesearch/pubs/63184

Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York, NY: Springer-Verlag.

Vaish, A. K., Chen, S., Sathe, N. S., Folsom, R. E., Chandhok, P., and Guo, K. (2010). Small area estimates of daily person-miles of travel: 2001 National Household Transportation Survey. *Transportation* 37, 825–848. doi: 10.1007/s11116-010-9279-8

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

Wang, J. C., Holan, S. H., Nandram, B., Barboza, W., Toto, C., and Anderson, E. (2012). A Bayesian approach to estimating agricultural yield based on multiple repeated surveys. *J. Agric. Biol. Environ. Stat.* 17, 84–106. doi: 10.1007/s13253-011-0067-5

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

You, Y., Rao, J. N., and Gambino, J. (2003). Model-based unemployment rate estimation for the Canadian labour force survey: a hierarchical Bayes approach. *Survey Methodol.* 29, 25–32.

Conflict of Interest: GW is 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.

A. APPENDIX

A.1. Area Level EBLUP Variance Estimator

The variance of the Area level EBLUP is expressed by the following equation (Molina and Marhuenda, 2015; Rao and Molina, 2015):

$$\hat{V}(\hat{\mu}_j^{EBLUP}) = g_{1j} + g_{2j} + 2g_{3j} - b \tag{A1}$$

where

$$g_{1j} = \hat{\gamma}_j \hat{V}\left(\hat{\mu}_{y_j}^{PS}\right),$$

$$g_{2j} = \hat{\sigma}_v^2 \left(1 - \hat{\gamma}_j\right)^2 \mathbf{z}_\mathbf{j}' \left(\sum_j \hat{\gamma}_j \mathbf{z}_\mathbf{j} \mathbf{z}_\mathbf{j}'\right)^{-1} \mathbf{z}_\mathbf{j},$$

$$g_{3j} = 2m\left(\hat{V}\left(\hat{\mu}_{y_j}^{PS}\right)\right)^2 \left(\hat{\sigma}_v^2 + \hat{V}\left(\hat{\mu}_{y_j}^{PS}\right)\right)^{-3} \left(\sum_j \left(\hat{\sigma}_v^2 + \hat{V}\left(\hat{\mu}_{y_j}^{PS}\right)\right)^{-1}\right)^{-2},$$

$$b = 2m\hat{\sigma}_v^2 \left(\sum_j \left(\hat{\gamma}_j\right)^2 - \left(\sum_j \hat{\gamma}_j\right)^2\right)\left(\sum_j \hat{\gamma}_j\right)^{-3}\left(1 - \hat{\gamma}_j\right)$$

where

$$\mathbf{z_j} = \begin{bmatrix} 1 \\ \overline{X}_j \end{bmatrix}.$$

One can intuitively think about each $g_{\#j}$ as follows: g_{1j} accounts for within-area variation, g_{2j} accounts for variation in estimating the regression parameter β, and g_{3j} accounts for model-variance estimation (Hidiroglou and You, 2016).

frontiers in Forests and Global Change

ORIGINAL RESEARCH
published: 18 January 2022
doi: 10.3389/ffgc.2021.763414

GREGORY: A Modified Generalized Regression Estimator Approach to Estimating Forest Attributes in the Interior Western US

Olek C. Wojcik[1], *Samuel D. Olson*[1], *Paul-Hieu V. Nguyen*[1], *Kelly S. McConville*[1]*, *Gretchen G. Moisen*[2] *and Tracey S. Frescino*[2]

[1] Department of Mathematics, Reed College, Portland, OR, United States, [2] Rocky Mountain Research Station, USDA Forest Service, Ogden, UT, United States

The national forest inventory within the US has been experiencing a greater need to estimate forest attributes over smaller geographic areas than the inventory was originally designed for. Producing reliable estimates for these areas may require the use of estimation methods beyond post-stratification. Staying within the dominant design-based paradigm, this research explores how model-assisted estimation is impacted by leveraging data outside the area of interest. In particular, we compare the performance of the post-stratified estimator, the generalized regression estimator (GREG), and a modified GREG. Typically the assisting model of the modified GREG is fit over a sample comprising all of the areas of interest. Here we introduce a modified GREG, denoted as GREGORY, which gives the practitioner a high degree of flexibility in selecting the sample subset for constructing the assisting model. We use these estimators to produce county level estimates of the mean of four forest attributes in the Interior Western US. Comparing the relative efficiencies of the estimators, we find that the more complex estimators, GREG and GREGORY, generally improve the precision of the estimates, especially in regions with a high degree of forested land. When using all the data from a 10-year measurement, fitting the model over a larger region does not lead to efficiency gains. To explore the impact of smaller sample sizes, we conduct a simulation study and find that as the sampling intensity decreases, the GREGORY tends to produce more efficient estimates than the GREG, and its variance estimator exhibits less negative bias. The GREG and GREGORY can easily be computed and compared using a new R package, gregRy, available on CRAN.

Keywords: generalized regression (GREG) estimator, post-stratification, model-assisted estimation, ecoregions, improved precision, domain estimation

OPEN ACCESS

Edited by:
Paolo Giordani,
University of Genoa, Italy

Reviewed by:
Stephen Stehman,
SUNY College of Environmental Science and Forestry, United States
Nicholas Nagle,
The University of Tennessee, Knoxville, United States

***Correspondence:**
Kelly S. McConville
mcconville@reed.edu

Specialty section:
This article was submitted to Forest Management, a section of the journal Frontiers in Forests and Global Change

Received: *23 August 2021*
Accepted: *23 December 2021*
Published: *18 January 2022*

Citation:
Wojcik OC, Olson SD, Nguyen P-HV, McConville KS, Moisen GG and Frescino TS (2022) GREGORY: A Modified Generalized Regression Estimator Approach to Estimating Forest Attributes in the Interior Western US. Front. For. Glob. Change 4:763414. doi: 10.3389/ffgc.2021.763414

1. INTRODUCTION

The US Forest Inventory and Analysis Program (FIA) is responsible for monitoring forest ecosystems across the United States. Established in 1930, the initial focus of this program was to estimate the extent and volume of merchantable trees for harvest. But today the extensive data collected nationwide are valuable for assessing biomass and carbon storage, fuels and fire

risk, wildlife habitat, effects of insect and disease outbreaks, forest health and trends in forest conditions. Along with these new uses, FIA is experiencing a greater demand for estimates of forest attributes over smaller geographic areas. Specifically, the "Agricultural Act of 2014" (U.S. Department of Agriculture, 2014) calls for FIA to implement procedures to improve precision in sub-state estimates, pushing the inventory to provide information at scales beyond which it was originally intended. Producing reliable estimates for these smaller areas requires considering additional data sources and new estimation methods beyond FIA's current techniques.

One standard estimation approach is the generalized regression estimator (GREG), which has the capacity of combining inventory data and remote-sensing data using a wide range of predictive modeling techniques (Särndal et al., 1992; Breidt and Opsomer, 2017). The GREG is a *direct estimator*, since it only uses data within the domain of interest and is *design-based* in that randomness comes solely from sample selection. The GREG is asymptotically unbiased, regardless of how well the model captures the true relationship between the inventory and auxiliary data. This useful feature is why the estimator is classified as *model-assisted* and not *model-based*.

Using a variety of assisting models, the GREG has been applied and studied rather extensively in the forest inventory literature (Baffetta et al., 2009; McRoberts, 2010; Gregoire et al., 2011; Moser et al., 2017; McConville et al., 2020). A thorough summary of forest inventory estimators that utilize models, including model-assisted estimators like the GREG, can be found in Ståhl et al. (2016). Most of the focus in these articles is on large areas with adequate sample sizes within the domain of interest. For areas with few sampled ground plots, the model estimates may not capture the true relationship well and may be highly variable. A solution is to leverage sample data outside the domain of interest to estimate the GREG's assisting model, resulting in what is sometimes referred to as a modified GREG (Rao and Molina, 2015). Most commonly the entire sample across all domains of interest is used to fit the model for the modified GREG. Here we consider estimating the models over large homogeneous regions and then combining the model predictions within the domain of interest. We call this estimator GREGORY for GREG *O*ver *R*esolutions of Y, where Y stands for the inventory data, to emphasize that the additional regions leveraged should depend on their homogeneity with the inventory data in the domain of interest. Although the GREGORY leverages data from outside the domain of interest, Rao and Molina (2015) still classify it as a direct estimator, since it only applies model parameter estimates to the plot data within the domain of interest and is still design-based in that randomness comes solely from sample selection. As with the GREG, the GREGORY is model-assisted, an important feature to national statistical agencies.

While the modified GREG, or GREGORY, has been proposed in the survey statistics literature (section 2.5 in Woodruff, 1966; Rao and Molina, 2015), it does not, to the best of our knowledge, appear to have been investigated deeply in the forest inventory literature. In this article, we hope to provide some insights into the utility of the GREGORY for forest estimation. Through a case study focused on estimating county level means of forest attributes in the US Interior West (IW), we attempt to measure how the estimator precision changes when the model-estimating now leverages additional data outside the domain of interest. Additionally, we investigate how precision gains from estimating the model over these broader samples change and the bias of the standard variance estimator as the sample size decreases.

We focus on a design-based, model-assisted approach for small domain estimation and consider only direct estimators in this article. Although a wide range of model-based methods and indirect estimators (Empirical Best Linear Unbiased Prediction, Hierarchical Bayes) exist, the design-based approach to estimation is still the prevailing choice for many national forest inventories because of its freedom from model assumptions. Therefore, it is important to understand the viability of a model-assisted estimator when the sample size is small and how leveraging more data impacts the performance of the estimator when compared to post-stratification, the standard estimation technique for larger regions.

2. METHODS

In this article, our domains of interest are counties in the IW and we focus on estimating the county level mean of four forest inventory variables: basal area (square-foot per acre), count of trees per acre, above-ground biomass (pounds per acre), and net volume (cubic-foot per acre). These inventory variables are all strongly and positively correlated with one another (with Pearson correlation coefficients between 0.42 and 0.6 with count of trees per acre and 0.85 and above for all other combinations of variables). Let U_d denote the spatial domain of county d, which has been discretized into N_d units based on the resolution of the auxiliary data and is enumerated by $\{1, 2, \ldots, N_d\}$. We write the true, unknown mean of U_d for a given inventory variable, y, as $\mu_{y_d} = {N_d}^{-1} \sum_{i \in U_d} y_i$. Our goal is to estimate μ_{y_d} for $d = 1, 2, \ldots, D$ where D equals the 280 counties with plots in the IW.

2.1. Data Sources

Computing the estimators requires data on the response variables, any predictor layers for estimating the assisting models (via GREG or GREGORY), a post-stratification layer for the PS estimator, and a layer depicting ecologically similar regions for leveraging data for the GREGORY. For county d, the set of sample plots is given by s_d, which is a subset of U_d, and the sample size is denoted by n_d. Field plot data were collected by FIA on a quasi-systematic sample of ground plots over a 10 year period (2007–2017). FIA data in the western US are collected on a 10-year measurement cycle. Specifically, plot data are collected under an annual, non-overlapping panel design, where each panel consists of one-tenth of the sample plots distributed roughly equidistant throughout the population (Reams et al., 2005). After 10 years, data on all plots have been collected and re-measurement of plots resumes in the first panel. With a base sampling intensity of one plot per every 6,000 acres, our IW sample represents one 10-year measurement cycle and includes data from 86,057 field plots. The plot data include our four response variables: basal area, count of trees per acre, above-ground biomass, and net volume, along with the RMRS-FIA

post-strata classifications and weights. The current IW post-stratification scheme is a forest/non-forest classification based on a forest probability map (Blackard et al., 2008). This layer is no longer being maintained or updated, hence being phased out of FIA estimation processes. So this variable is not considered as potential auxiliary data in the GREG or GREGORY. The inventory data were downloaded on February 6, 2019 from the FIA database, version FIADB_1.8.9.99 (last updated Dec 3, 2018).

Our predictor variable comes from the 2016 National Land Cover Database (NLCD) Tree Canopy Cover (TCC) map, which provides estimates of the percent tree canopy cover for the entire IW at a resolution of 30 by 30 meters^2 (Yang et al., 2018). Therefore, the discretization of county d is done at a 30-m resolution and the population size, N_d, is given by the number of pixels from the NLCD TCC map in county d. In addition to the unit level pixel TCC data, denoted by $\{x_i\}_{i\in U_d}$, we extract the subset, $\{x_i\}_{i\in s_d}$ where unit i is the pixel that is spatially closest to the center of field plot i.

Since the GREGORY allows the assisting model to leverage data outside the domain of interest, we must also determine what subset of s should be used for each county. While the model could be estimated using s, the entire IW sample, we focus on estimating the model over the ecological provinces given by Cleland et al. (2007) since they delineate the landscape into ecological units across the conterminous US based on major vegetation cover types and land forms. See **Figure 1** for the eco-provinces in the IW.

FIA data retrievals and processing of auxiliary data were done through the R package `FIESTA` (Frescino et al., 2020). In summary, we have the following data for each county d and each response variable:

- $\{y_i, x_i, z_i, f_i\}_{i\in s_d}$ where the data for plot i includes y_i, the value of the forest inventory/response variable, x_i, the TCC value, z_i, the eco-province, and f_i, the post-strata classification.
- $\{x_i\}_{i\in U_d}$, the TCC values for each unit in county d.
- $\{w_{pl}\}_{l=1}^{14}$, a set of weights where each weight represents the proportion of county d in a given eco-province.
- $\{w_{sl}\}_{l=1}^{2}$, a set of weights where each weight represents the proportion county d in a given post-stratum.

2.2. Estimators

In this section, we formally introduce the GREG and its extension, the GREGORY. We also present the post-stratified estimator (PS), which is featured in our analyses since it is the standard estimator used in FIA's production processes. Additionally, we address variance estimation and provide two variance estimators. All data analysis was done in the statistical software package R (R Core Team, 2020) and the estimators were computed using the `gregRy` package (Olson and Wojcik, 2021).

2.2.1. The Generalized Regression Estimator

The GREG for μ_{y_d} is given by

$$\hat{\mu}_{y_d,GREG} = \frac{1}{n_d}\sum_{i\in s_d}\left(y_i - \hat{m}(\boldsymbol{x}_i)\right) + \frac{1}{N_d}\sum_{i\in U_d}\hat{m}(\boldsymbol{x}_i) \qquad (1)$$

where for county d, $\hat{m}(\boldsymbol{x}_i)$ is the model prediction for unit i based on the predictor vector $\boldsymbol{x}_i^T = (1, x_i)$. When we assume a linear regression assisting model, then $\hat{m}(\boldsymbol{x}_i) = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ with estimated least squares regression coefficients, $\hat{\boldsymbol{\beta}}^T = (\hat{\beta}_0, \hat{\beta}_1)$, given by

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}}\sum_{i\in s_d}\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2.$$

In Equation (1), the first term, the average residuals component, ensures that the estimator is asymptotically unbiased since it compensates for any under- or overestimation caused by the second term, which provides the average predicted value. This second term is commonly called the synthetic estimator (Rao and Molina, 2015). Notice that the GREG is only constructed using data within the domain of interest, U_d, and in particular that the estimated regression coefficients are computed using only s_d. When n_d is small, the variance of the estimated coefficients may be large, which in turn increases the variance of $\hat{\mu}_{y_d,GREG}$. Another potential concern is bias. If n_d is small, the property of asymptotic unbiasedness of the estimator may no longer hold. The GREGORY attempts to overcome these issues by fitting the model using not just s_d but also using ecologically similar sample data.

2.2.2. The Generalized Regression Estimator Over Resolutions of Y

For the GREGORY, the estimator form is still given in Equation (1) but now the models are estimated over a larger region. To differentiate between the different data sources, we call the sample data used in estimation, s_d, the *estimation sample* while that used in modeling is called the *modeling sample*. For our data application, the resolution of the modeling samples are eco-provinces and so the estimated model prediction for unit i is given by a weighted sum of regression models,

$$\hat{m}(\boldsymbol{x}_i) = \boldsymbol{x}_i^T\left(\sum_{l=1}^{P} w_{pl}\hat{\boldsymbol{\beta}}_l\right),$$

where the estimated regression coefficient vector for province l come from

$$\hat{\boldsymbol{\beta}}_l = \text{argmin}_{\boldsymbol{\beta}}\sum_{i\in s}\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2 I(z_i = l).$$

Recall that z_i specifies the eco-province of unit i. By separating the estimation and modeling samples, we are able to estimate the models using larger sample sizes and eco-province samples that are likely more ecologically homogeneous than those created by the arbitrary political boundaries of counties. If an estimation sample is nested in a modeling sample, then $\hat{m}(\boldsymbol{x}_i)$ reduces to a single regression equation. While we focus on weighting simple linear regression models here, more nuanced correlation structures that allow for spatial and/or temporal autocorrelation could be incorporated through a mixed-model approach.

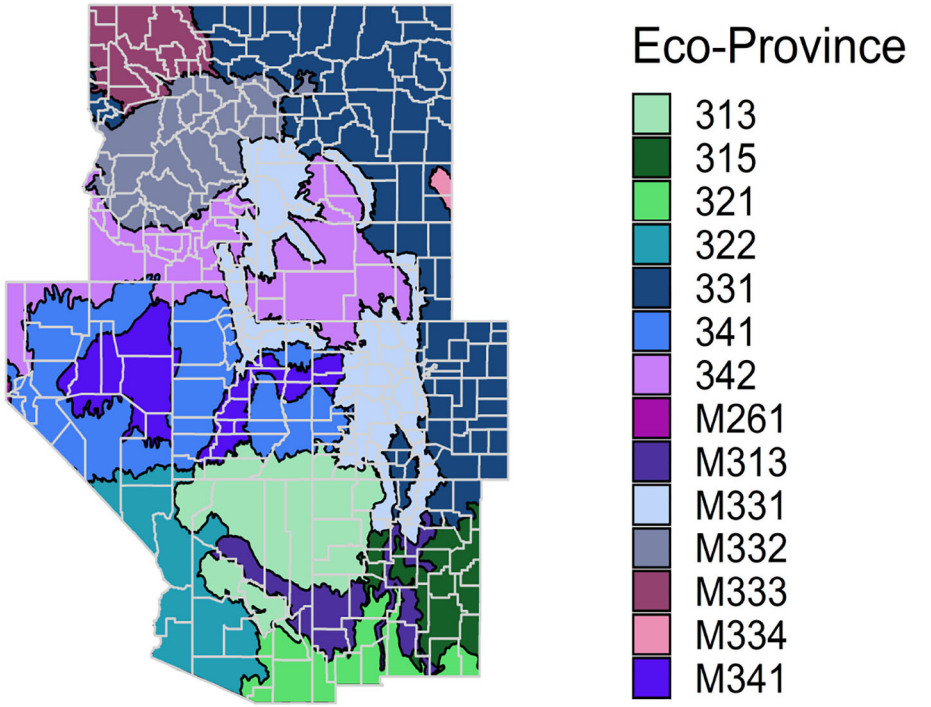


FIGURE 1 | A map of counties in the Interior West, colored by eco-provinces.

2.2.3. Post-stratification

The PS is a special case of a GREG where a single categorical predictor is used in the regression assisting model. In this case, the estimator of μ_{y_d} simplifies to a weighted sum of post-strata means,

$$\hat{\mu}_{y_d,PS} = \sum_{j=1}^{2} w_{sj}\bar{y}_j,$$

where $\bar{y}_j = n_{dj}^{-1}\sum_{i\in s_d} y_i I(f_i = j)$, the sample mean of y for post-stratum j and n_{dj} is the number of sampled plots in post-stratum j for county d. Recall that w_{sj} is the proportion of county d in post-stratum j.

2.2.4. Variance Estimation

Särndal et al. (1992) provide the standard variance estimator of the GREG,

$$\widehat{V}(\hat{\mu}_{y_d}) = \left(1 - \frac{n_d}{N_d}\right)\frac{1}{n_d}\frac{1}{n_d - 1}\sum_{i\in s_d}\left(y_i - \hat{m}(\boldsymbol{x}_i)\right)^2, \quad (2)$$

which can also be used to estimate the variance of the GREGORY. However, the form of the variance estimator relies on large sample approximations and does not account for model estimation variation. Note that the model coefficients of the GREG are chosen to minimize the sum of the squared errors over s_d. Therefore, equation (2) will always report a smaller value for GREG than GREGORY, *by construction*. However, the true variance of the GREGORY may in fact be smaller than the variance of the GREG since its modeling sample is typically larger and therefore its model estimation variance is likely smaller. To compare the efficiency of the estimators, we want a variance estimator that accounts for both the variability in the residuals and the variability induced by fitting the model. Therefore, in our data application we estimate the variance of the estimators not using Equation (2) but using the following bootstrap variance estimator

$$\widehat{V}_B(\hat{\mu}_{y_d}) = \left(\frac{n_d}{n_d - 1}\right)\left(\frac{N_d - n_d}{N_d - 1}\right)\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\mu}_{y_d}^{(b)} - \bar{\hat{\mu}}_{y_d})^2$$

where $\hat{\mu}_{y_d}^{(b)}$ is the bth bootstrap estimate and $\bar{\hat{\mu}}_{y_d} = B^{-1}\sum_{b=1}^{B}\hat{\mu}_{y_d}^{(b)}$ is the average of the bootstrapped estimates. See Mashreghi et al. (2016) for more details on using bootstrap methods in survey estimation.

Returning to the standard variance estimator, it is important to understand the degree of its negative bias since it is commonly used in practice. For simple models and moderately large sample sizes where model estimation variability accounts for little of the overall variance, the standard variance estimator tends to be slightly negatively biased. For more complex models, Kangas et al. (2016) found that this variance estimator can significantly underestimate the true variance. In the simulation study, we explore and compare the bias of the standard variance

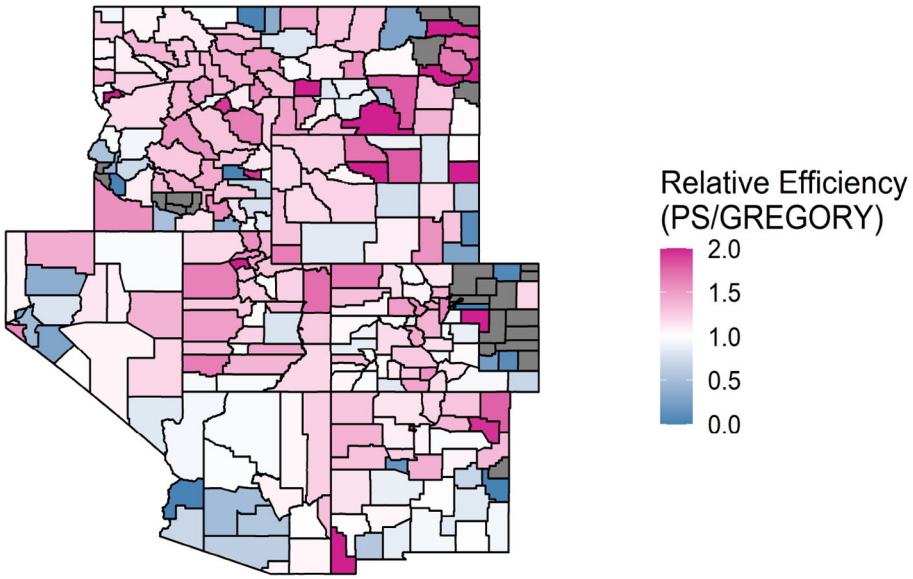


FIGURE 2 | A map of the relative efficiencies of the PS to the GREGORY when estimating the average trees per acre for each county in the Interior West. Values above 1 indicate that the GREGORY is more efficient. Values greater than 2 were truncated to 2 to increase the readability of the map. A county is gray if the RE is 0, due to all plots containing values of 0 trees per acre.

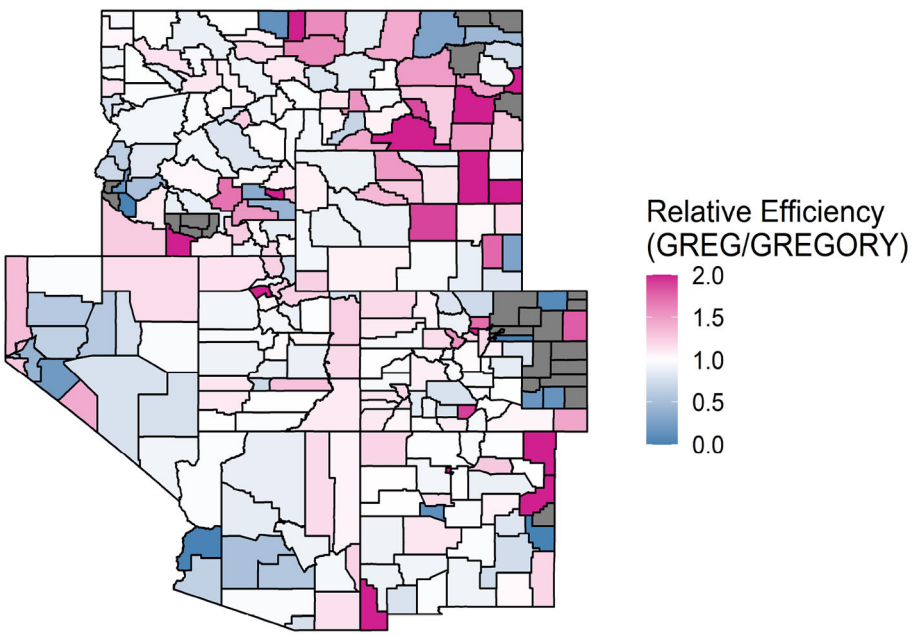


FIGURE 3 | A map of the relative efficiencies of the GREG to the GREGORY when estimating the average trees per acre for each county in the Interior West. Values above 1 indicate that the GREGORY is more efficient. Values greater than 2 were truncated to 2 to increase the readability of the map. A county is gray if the RE is 0, due to all plots containing values of 0 trees per acre.

estimator for both the GREG and the GREGORY across a range of sampling fractions. This allows us to study how the size of the modeling sample impacts the biasedness of the variance estimator.

3. RESULTS

In the data application and simulation study, we compare how county level models vs. eco-province level models impact the

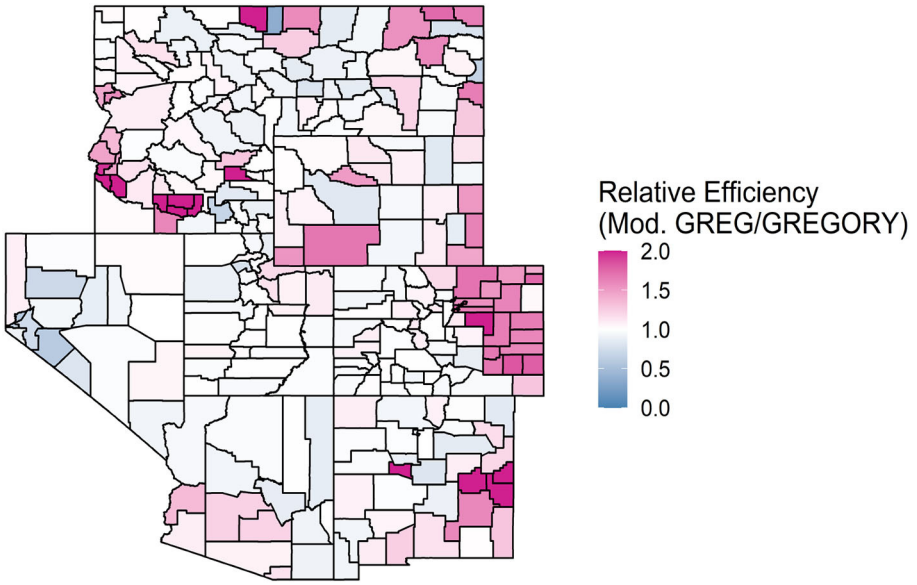


FIGURE 4 | A map of the relative efficiencies of the modified GREG to the GREGORY when estimating the average trees per acre for each county in the Interior West. Values above 1 indicate that the GREGORY is more efficient. Values greater than 2 were truncated to 2 to increase the readability of the map.

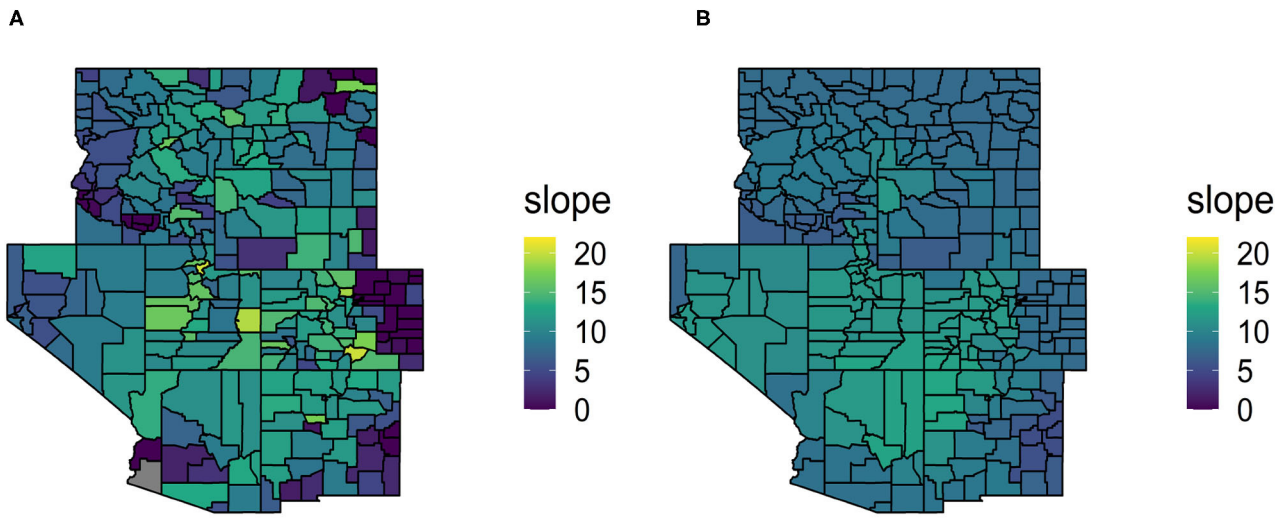


FIGURE 5 | Map **(A)** contains the estimated slopes of the linear regression model of trees per acre based on TCC for each county in the GREG when the models are fit at the county level. Map **(B)** contains the estimated slopes for each county in the GREGORY when the models are fit at the province level.

model-assisted estimator, especially as we vary the sampling intensity within the counties. While we considered four response variables, we present the results for estimating the average trees per acre in this section. We found similar patterns for the other three response variables.

3.1. Case Study

Our data application investigates the impact of leveraging more data when estimating the model for a modified GREG. We focus on producing county level estimates of the mean trees per acre for the IW and compare the performance of the PS and GREG, which uses only data within the domain, to the GREGORY, which uses additional data from outside the domain. We also consider the modified GREG which uses the entire IW sample for model fitting.

The first step is to determine the resolution of the model samples for the GREGORY, which could range from using just the sampled plots in the county of interest to the entire set of sampled plots in the IW to something in between. Using just the sampled plots, as the GREG does, runs the risk of high variance

in its estimates, especially for small sample sizes. On the other hand, using the entire sample, as the modified GREG does, could also be ill-advised if it means lumping together heterogeneous landscapes where the relationship between TCC and trees per acre may vary. And, though GREG is asymptotically unbiased, concerns of bias arise from the small sample sizes of areas being estimated. To reduce the bias of the eventual estimate in finite samples, it would be ideal to estimate GREGORY's model using plot data from areas that have a similar relationship between TCC and trees per acre as that found in the county of interest. Keeping this in mind, we constructed modeling samples based on ecology, in particular, the eco-provinces given by Cleland et al. (2007). When considering at what level to estimate our models, we were motivated to utilize the eco-province level after considering the different levels of ecologies used by FIA. The principal map unit design criterion for eco-provinces is the dominant potential natural vegetation, compared to more granular levels such as eco-sections, which are delineated by the physical and biological components of an ecology such as climate, physiography, lithology, soils, and potential natural communities (McNab et al., 2007).

Figures 2–4 allow for a spatial look at how the relative efficiencies of the estimators, given by the ratio of the estimated bootstrapped variances, compare to one another when estimating the average count of trees per acre for each county. A county is gray if all county plots had a response value of 0 and therefore a variance estimate of 0 for PS or GREG. GREGORY and modified GREG circumvent this issue by using data from outside of these problematic counties. As seen in **Figure 2**, the GREGORY has a lower variance estimate than the PS for most counties (71%). This trend was similar when comparing GREG to PS. However, we see from **Figure 3** that GREG and GREGORY are roughly matched in the number of counties in which one outperforms the other (with GREGORY outperforming 53% of the time). This implies that constructing the model over a larger resolution did not, generally, reduce the variance of the estimates.

We can expand the modeling sample even further, as the modified GREG does, and compare that to the GREGORY, as seen in **Figure 4**. While GREGORY only outperformed the modified GREG 54% of the time, the precision gains were rather large for some counties and the precision losses were not as extreme. On average, the estimated variance of the modified GREG is 1.14 times the estimated variance of the GREGORY, suggesting that building the model over ecologically homogeneous samples can improve the efficiency of the estimator.

It should be noted that we did see a higher degree of variability in the estimated slopes ($\hat{\beta}_1$) for the county level models than the eco-province level models (see **Figure 5**). We conjecture that this extra variability did not translate into higher variance estimates because the predictive accuracy of the estimated model is a much more dominant component of the variance. This actually provides justification for the standard variance estimator, given in Equation (2), only being a function of the prediction errors and not accounting for model estimation variability. In the next section, we conduct a simulation study to more concretely compare the variability and bias of the estimators as the sample size shrinks.

TABLE 1 | Table of the counties included as domains in the simulation.

County	State	Number of plots
Beaverhead county	Montana	597
Bonner county	Idaho	202
Catron county	New Mexico	507
Clearwater county	Idaho	215
Custer county	Idaho	493
Duchesne county	Utah	250
Eureka county	Nevada	403
Flathead county	Montana	542
Gallatin county	Montana	279
Garfield county	Colorado	292
Grand county	Colorado	201
Grant county	New Mexico	258
Gunnison county	Colorado	340
Idaho county	Idaho	810
Lander county	Nevada	456
Lemhi county	Idaho	479
Lewis and Clark county	Montana	274
Lincoln county	Montana	381
Madison county	Montana	377
Meagher county	Montana	235
Missoula county	Montana	264
Park county	Wyoming	343
Park county	Montana	282
Park county	Colorado	230
Powell county	Montana	231
Ravalli county	Montana	242
Rio blanco county	Colorado	289
Routt county	Colorado	225
Saguache county	Colorado	226
San Miguel county	New Mexico	215
Sanders county	Montana	278
Sevier county	Utah	205
Shoshone county	Idaho	274
Teton county	Wyoming	304
Uintah county	Utah	270
Valley county	Idaho	389
White Pine county	Nevada	929

Also listed are the number of plots from the county that are used. Only plots within provinces M313, M331, M332, M333, and M341 were included.

3.2. Simulation Study

The application in the previous section compares *estimated* variances and so observed differences may be due to random variability and are not necessarily indicating that one estimator is truly more precise than another in a given county. To better understand how the modeling sample size impacts the estimator's bias and precision, we conducted a simulation study. We treated

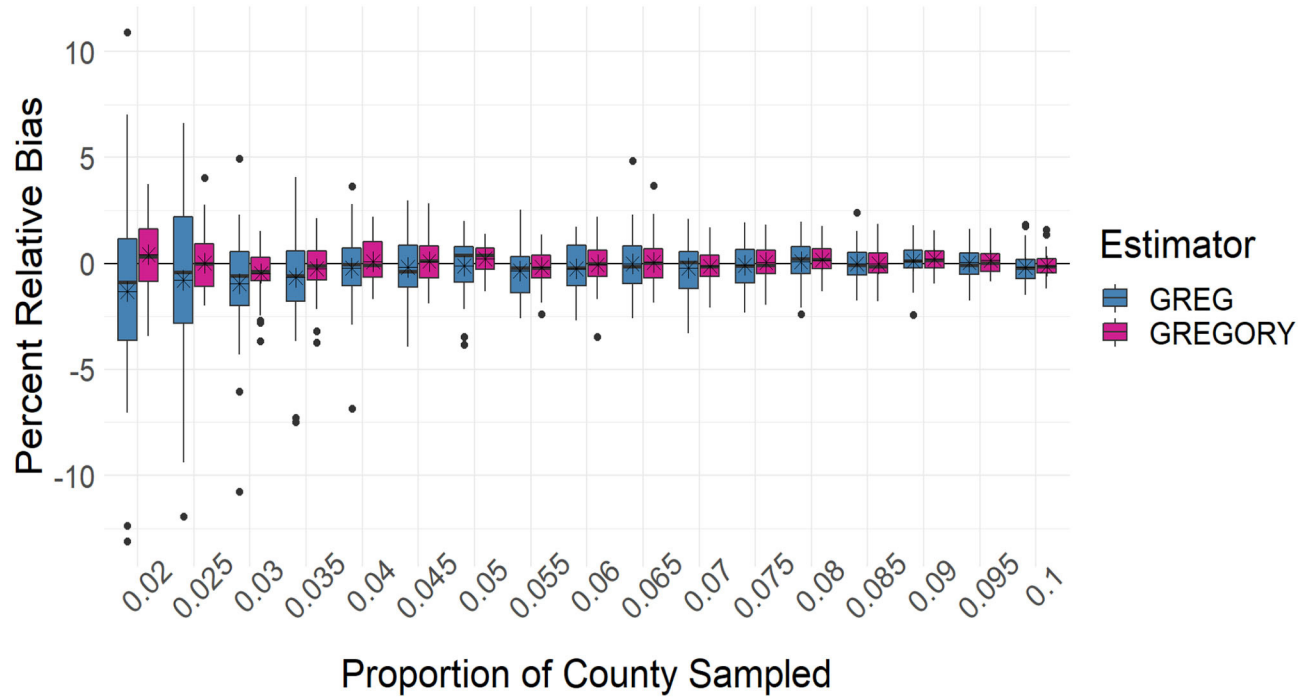


FIGURE 6 | Boxplots of the percent relative bias of the county estimates of average trees per acre across the sampling intensities. The average percent relative bias is denoted by the stars.

part of the IW as the true, finite population and drew 1000 Monte Carlo samples from the population. By using the plot data as the population, we know the true mean trees per acre for each county and therefore can obtain both the percent relative bias and the empirical mean squared error for the estimators, along with the percent relative bias of the standard variance estimator, by averaging across the samples. Due to the computational intensity of the bootstrap variance estimator, we only measure the bias of the standard variance estimator, given by Equation (2), in this study.

To ensure we had enough data and sampling variability, we selected for the population the 5 IW Mountain eco-provinces which each had at least 3,000 plots. Within these 5 eco-provinces, we selected the counties which had at least 200 plots and where a majority of the county plots were in the selected eco-provinces. It should be noted that we did not include plots from outside these 5 eco-provinces, even if they were in one of the selected counties. **Table 1** contains information on the 37 counties that comprised the finite population. For each replicate sample, we randomly sampled p% of each county. To explore the effect of sampling intensity, we varied p from 2 to 10 in 0.5 increments.

While the GREG and GREGORY are asymptotically unbiased, the estimators are applied in practice to samples with finite sample sizes. Therefore, it is important to study the degree of bias in the estimators and their variance estimators, especially as a function of sample size. **Figures 6**, **7** capture the percent relative bias of the estimators across the sampling fractions and sample sizes. Both estimators exhibit little bias for the moderate to large sampling intensities but for the smallest intensities the GREG's percentage relative bias across the 37 counties is rather variable, more so than the GREGORY's.

Figures 8, **9** compare the mean square error of the GREGORY and the GREG across the sampling fractions and sample sizes. For the lower sampling fractions, the GREG MSE is more variable and larger, on average, than the MSE of the GREGORY. From **Figure 8**, we see that the GREGORY is typically more efficient than the GREG for smaller sample sizes and then the estimators perform similarly once a county has at least 30–40 sampled plots. This result demonstrates an advantage to using GREGORY in settings where data are sparse.

The distributions of the percent relative bias of the standard variance estimator, given in Equation (2), are displayed in **Figures 10**, **11**. The variance estimators for both the GREG and GREGORY are negatively biased for the smaller sampling intensities but the GREGORY is less so. And by a sampling fraction of around 6.5%, or a sample size of at least 20, the variance estimator of the GREGORY exhibits little bias, while the GREG variance maintains some amount of negative bias, even for the largest sample sizes.

4. CONCLUSION

This paper considers how the variability of a direct estimator is impacted when the assisting model is built

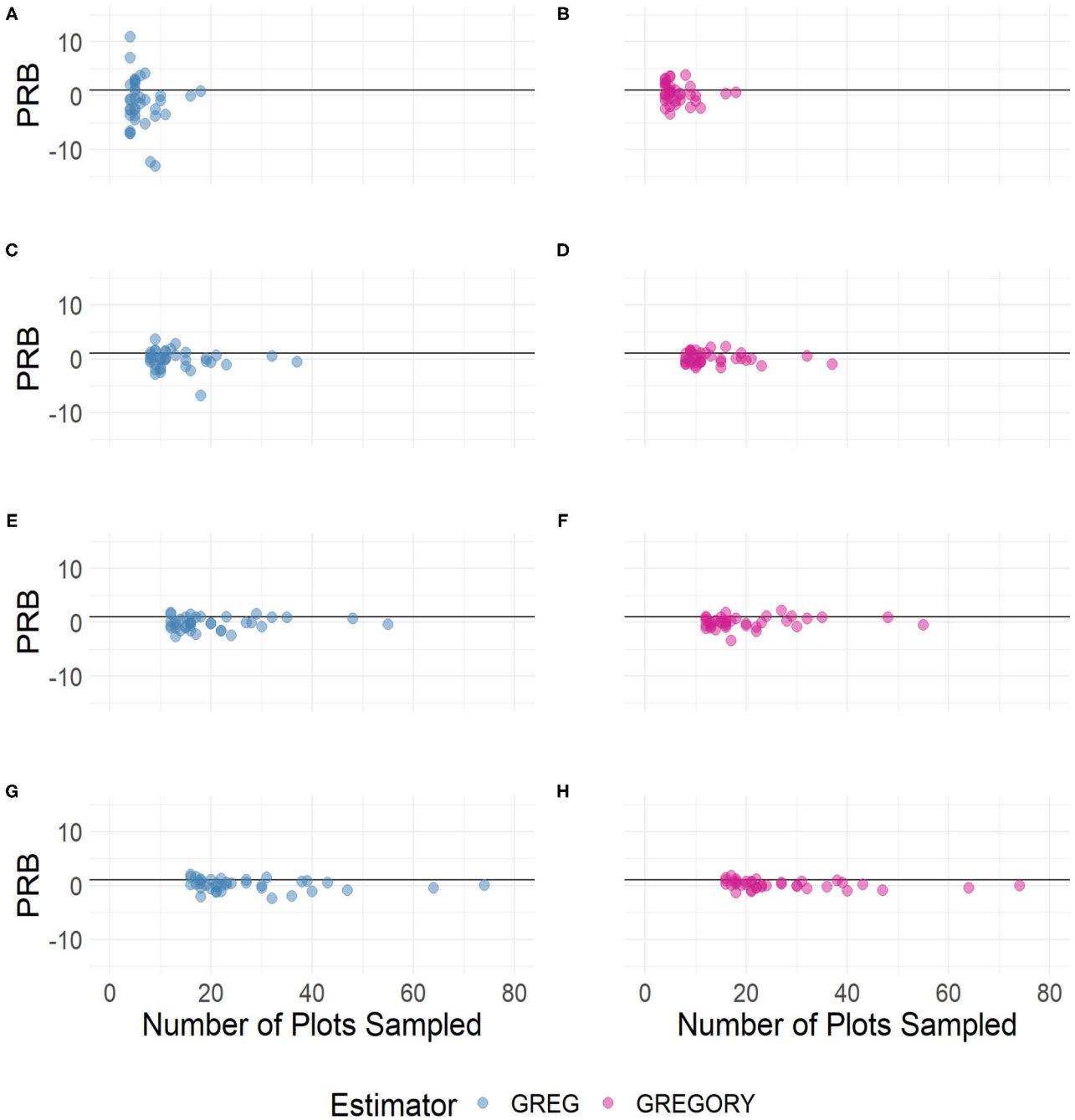


FIGURE 7 | The percent relative bias of the county estimates of average trees per acre by sample size across the following sampling intensities: **(A,B)** 2%, **(C,D)** 4%, **(E,F)** 6%, and **(G,H)** 8%.

using data from a larger region, some of which falls outside the domain of interest. We found that efficiency gains are achieved from these larger modeling samples when the sample size within the domain of interest is small.

A key interest for a practitioner is under what conditions to use GREGORY instead of GREG. We believe this primarily comes down to four questions. First, does survey data exist beyond the domains of interest that samples similar domains? Here, we used eco-province boundaries to identify similar areas

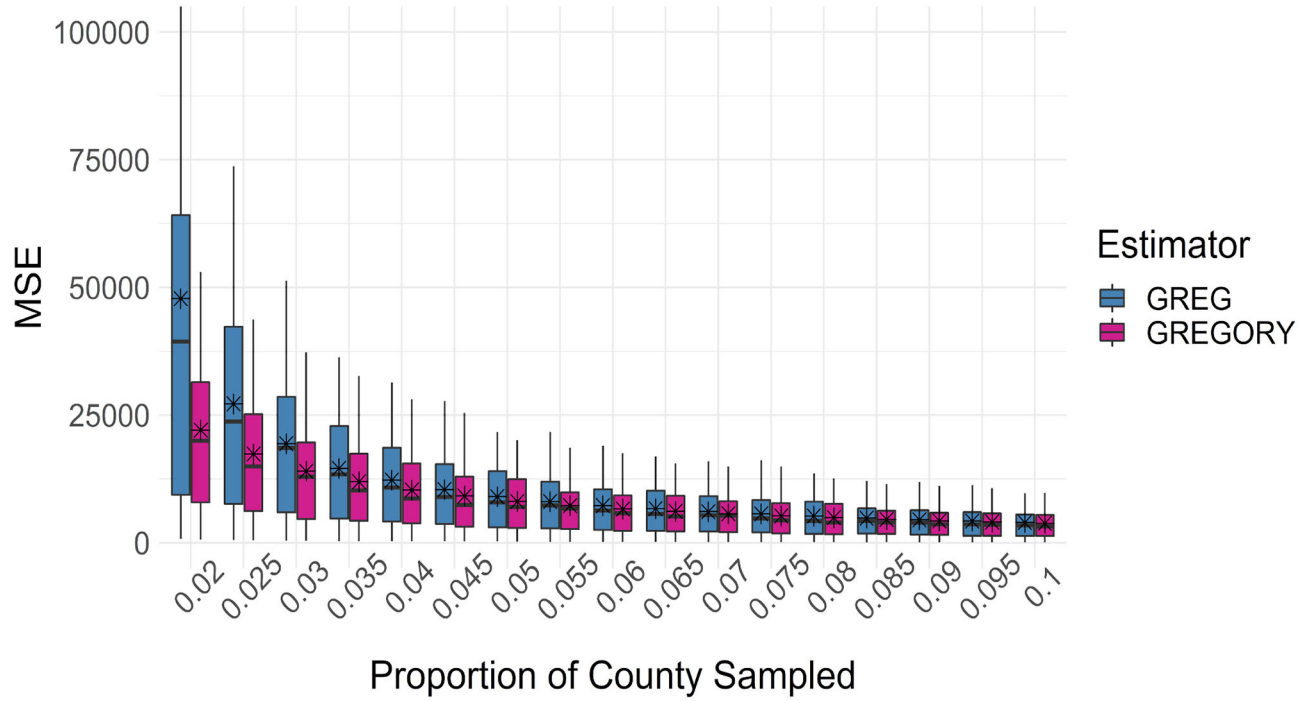


FIGURE 8 | Boxplots of the mean squared error of the county estimates of average trees per acre across the sampling intensities. The average mean squared error is denoted by the stars.

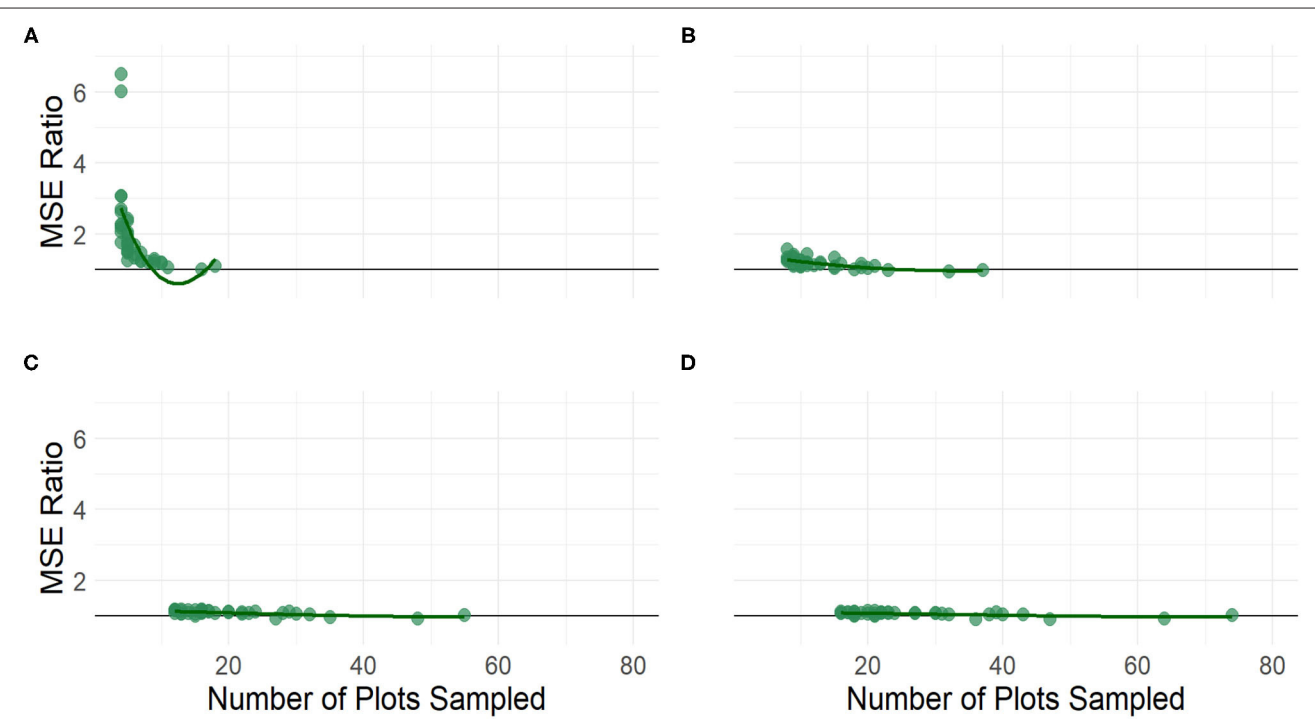


FIGURE 9 | The MSE ratio of the GREG to the GREGORY when estimating the average trees per acre at the county level by sample size across the following sampling intensities: **(A)** 2%, **(B)** 4%, **(C)** 6%, and **(D)** 8%. A loess smoother is included.

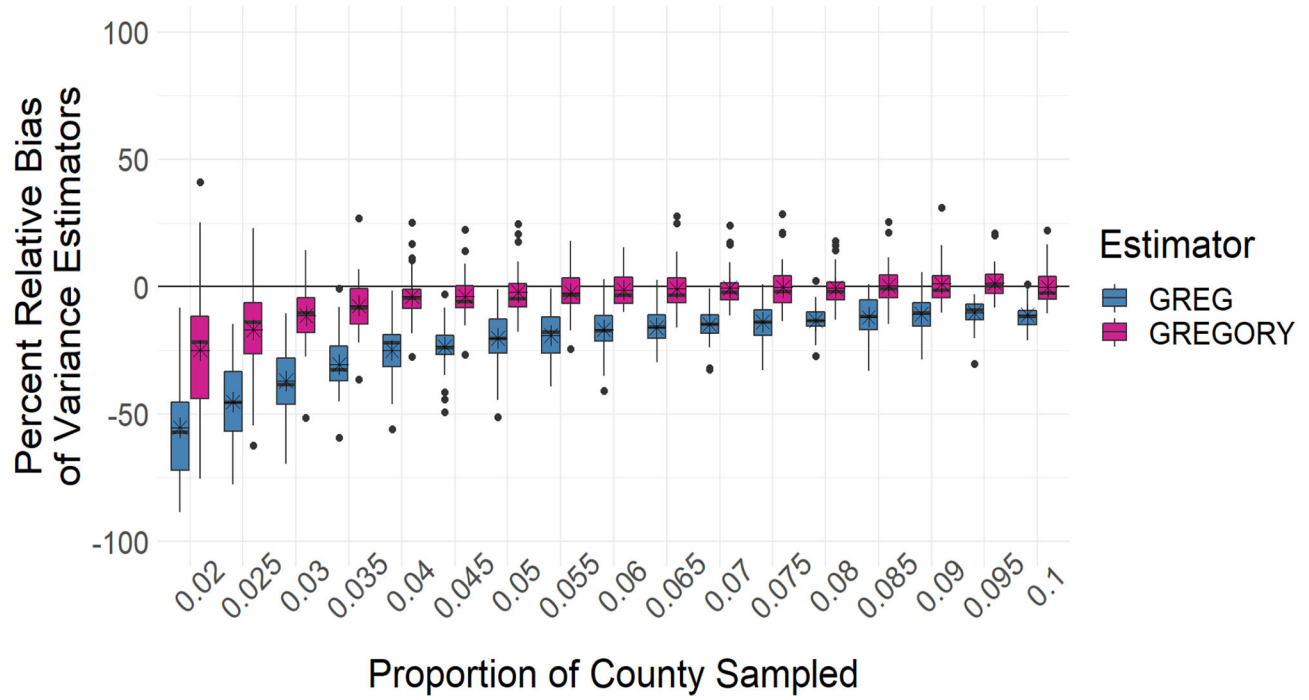


FIGURE 10 | Boxplots of the percent relative bias of the standard variance estimator when estimating the average trees per acre across the sampling intensities. The average percent relative bias is denoted by the stars.

that resulted in less variable estimates and avoided the problem of introducing bias caused by modeling from completely different populations. If the entire sample region is ecologically similar, then the modified GREG, which utilizes all the sample data to fit the model, should be considered. Second, does borrowing over this larger region result in diverse models? In our case, adjacent eco-provinces in the Interior Western US are often dramatically different due to topography, but borrowing over a larger area that is quite homogeneous could have little impact on the performance of estimators. Third, are there domains of interest with small sample sizes? In our application in the Interior West, enough data were available and the GREG was adequate for the situation. However, our simulation results show that GREGORY generally produced less biased estimates and better relative precision than GREG as sample size decreased. And fourth, how will the uncertainty of the estimates be calculated? We found that the standard variance estimator exhibited less negative bias for the GREGORY and eventually showed little bias for moderate sample sizes. On the other hand, the standard variance estimator of the GREG continued to exhibit negative bias across all sampling intensities. Lastly, we'd note that for very small sample sizes, a practitioner should consider model-based methods which more directly leverage information from outside the domain as these methods are likely to be more efficient.

Whether fitting a GREG or a GREGORY, there are additional considerations for a practitioner about what assisting model to employ and what auxiliary data to incorporate. These choices should be guided by extensive exploratory data analyses and visualizations. For the GREGORY, we fit separate linear models for each eco-province and then for each county, weighted the eco-province estimated model coefficients by the proportion of the eco-province in the county. There are many other potential approaches, such as building a single model with eco-province indicator functions or taking a mixed-model approach with eco-province random effects. Time spent up front thinking about the model, how the estimated model coefficients may vary across subsets, the inclusion of relevant ancillary data, spatial variations in the data, and domain sample sizes may be profitable by increasing the precision of particular small area estimates, in addition to motivating the choice between GREG and GREGORY.

For understanding operational implications for FIA, GREGORY should be evaluated as an alternative to post-stratification for more response variables, over different geographic regions, and using alternative auxiliary information. Further, much work is underway to expand forest inventory capacity to address new user needs through small area estimation. Through GREGORY, new investigations can determine just how far FIA can push direct, model-assisted estimators suitable for generic inference to meet small domain needs before turning to model-based methods.

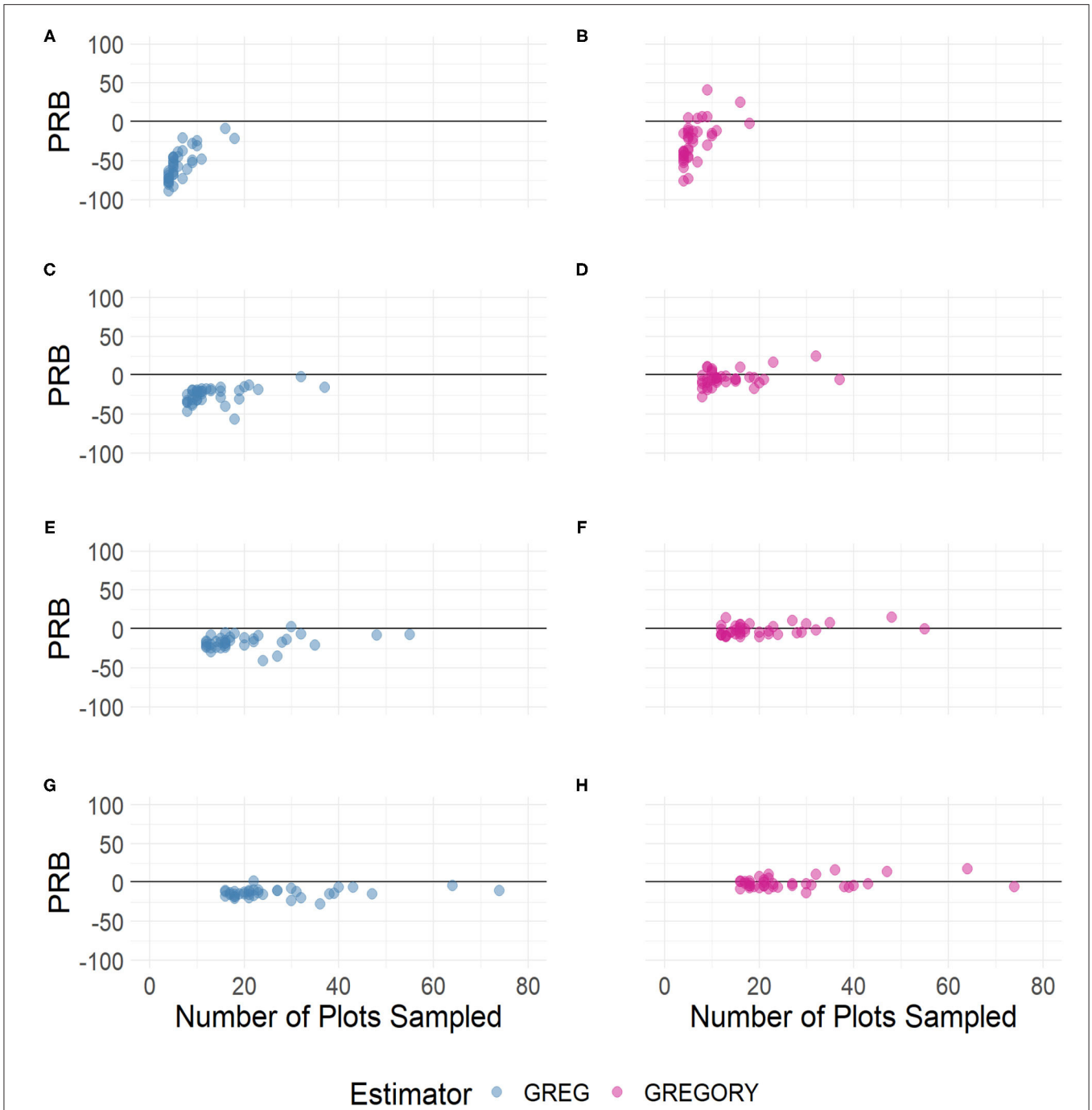


FIGURE 11 | The percent relative bias of the county standard variance estimator when estimating the average trees per acre by sample size across the following sampling intensities: **(A,B)** 2%, **(C,D)** 4%, **(E,F)** 6%, and **(G,H)** 8%.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: the data include confidential plot data, which can not be shared publicly. FIA data can be accessed through the FIA DataMart (https://apps.fs.usda.gov/fia/datamart/datamart.html). Requests for data used here or other requests including confidential data should be directed to FIA's Spatial Data Services (https://www.fia.fs.fed.us/tools-data/spatial/index.php). Requests to access these datasets

should be directed to https://apps.fs.usda.gov/fia/datamart/datamart.html, https://www.fia.fs.fed.us/tools-data/spatial/index.php.

AUTHOR CONTRIBUTIONS

OW, SO, P-HN, KM, GM, and TF: conceptualization. OW, SO, P-HN, KM, and GM: methodology and review and editing. OW, SO, and P-HN: analysis and data visualization. TF: data curation. OW, SO, P-HN, and KM: writing. OW and SO: software. KM: supervision. GM: funding acquisition. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported by the USDA Forest Service, Forest Inventory and Analysis Program (via agreement 19-JV-11221638-112) and by Reed College.

ACKNOWLEDGMENTS

The authors would like to thank the USDA Forest Service, Forest Inventory and Analysis Program for the data. The authors would also like to sincerely thank the handling editor and two reviewers for their thorough and constructive comments and suggestions. The reviews really helped us create a clearer and more complete final version of the article.

REFERENCES

Baffetta, F., Fattorini, L., Franceschi, S., and Corona, P. (2009). Design-based approach to k-nearest neighbours technique for coupling field and remotely sensed data in forest surveys. *Remote Sens. Environ.* 113, 463–475. doi: 10.1016/j.rse.2008.06.014

Blackard, J., Finco, M., Helmer, E., Holden, G., Hoppus, M., Jacobs, D., et al. (2008). Mapping US forest biomass using nationwide forest inventory data and moderate resolution information. *Remote Sens. Environ.* 112, 1658–1677. doi: 10.1016/j.rse.2007.08.021

Breidt, F. J., and Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Stat. Sci.* 32, 190–205. doi: 10.1214/16-STS589

Cleland, D. T., Freeouf, J. A., Keys, J. E., Nowacki, G. J., Carpenter, C. A., and McNab, W. H. (2007). *Ecological Subregions: Sections and Subsections for the Conterminous United States*. Available online at: https://www.fs.fed.us/research/publications/misc/73326-wo-gtr-76d-cleland2007.pdf

Frescino, T. S., Moisen, G. G., Patterson, P. L., Toney, C., and Freeman, E. A. (2020). "Demonstrating a progressive FIA through fiesta: a bridge between science and production," in *Brandeis, T. J., comp. Celebrating progress, possibilities, and partnerships: Proceedings of the 2019 Forest Inventory and Analysis (FIA) Science Stakeholder Meeting; November 19-21, 2019* (Knoxville, TN. e-Gen. Tech. Rep. SRS-256; Asheville, NC: US Department of Agriculture Forest Service, Southern Research Station), 199–200.

Gregoire, T. G., Ståhl, G., Nsset, E., Gobakken, T., Nelson, R., and Holm, S. (2011). Model-assisted estimation of biomass in a LiDAR sample survey in hedmark county, norway. *Can. J. Forest Res.* 41, 83–95. doi: 10.1139/X10-195

Kangas, A., Myllymki, M., Gobakken, T., and Nsset, E. (2016). Model-assisted forest inventory with parametric, semiparametric, and nonparametric models. *Can. J. Forest Res.* 46, 855–868. doi: 10.1139/cjfr-2015-0504

Mashreghi, Z., Haziza, D., and Lger, C. (2016). A survey of bootstrap methods in finite population sampling. *Stat. Surveys* 10, 1–52. doi: 10.1214/16-SS113

McConville, K. S., Moisen, G. G., and Frescino, T. S. (2020). A Tutorial on model-assisted estimation with application to forest inventory. *Forests* 11:244. doi: 10.3390/f11020244

McNab, W. H., Cleland, D. T., Freeouf, J. A., Keys, J. E., Nowacki, G. J., and Carpenter, C. A. (2007). *Description of ecological subregions: Sections of the conterminous united states*. General Technical Report (GTR), U.S. Department of Agriculture, Forest Service.

McRoberts, R. E. (2010). Probability-and model-based approaches to inference for proportion forest using satellite imagery as ancillary data. *Remote Sens. Environ.* 114, 1017–1025. doi: 10.1016/j.rse.2009.12.013

Moser, P., Vibrans, A. C., McRoberts, R. E., Nsset, E., Gobakken, T., Chirici, G. O., et al. (2017). Methods for variable selection in LiDAR-assisted forest inventories. *Forestry* 90, 112–124. doi: 10.1093/forestry/cpw041

Olson, S., and Wojcik, O. (2021). *gregRy: GREGORY Estimation*. Available online at: https://cran.r-project.org/web/packages/gregRy/index.html

R Core Team (2020). *R: A language and environment for statistical computing*. Vienna : R Foundation for Statistical Computing. Available online at: https://www.R-project.org/

Rao, J. N., and Molina, I. (2015). *Small Area Estimation*. Hoboken, NJ: John Wiley & Sons.

Reams, G. A., Smith, W. D., Hansen, M. H., Bechtold, W. A., Roesch, F. A., and Moisen, G. G. (2005). "The forest inventory and analysis sampling frame," in *The Enhanced Forest Inventory and Analysis Program—National Sampling Design and Estimation Procedures* (Asheville, NC: US Dep't of Agriculture, Forest Service, Southern Research Station), 11–26.

Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York, NY: Springer-Verlag.

Ståhl, G., Saarela, S., Schnell, S., Holm, S., Breidenbach, J., Healey, S. P., et al. (2016). Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation. *Forest Ecosyst.* 3:5. doi: 10.1186/s40663-016-0064-9

U.S. Department of Agriculture (2014). *Farm Bill*. Available online at: https://www.congress.gov/113/plaws/publ79/PLAW-113publ79.pdf

Woodruff, R. S. (1966). Use of a regression technique to produce area breakdowns of the monthly national estimates of retail trade. *J. Am. Stat. Assoc.* 61, 496–504.

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. Photogrammetry Remote Sens.* 146, 108–123. doi: 10.1016/j.isprsjprs.2018.09.006

Conflict of Interest: The 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Forests and Global Change

ORIGINAL RESEARCH
published: 03 March 2022
doi: 10.3389/ffgc.2022.763422

Examining *k*-Nearest Neighbor Small Area Estimation Across Scales Using National Forest Inventory Data

David M. Bell[1], Barry T. Wilson[2]*, Charles E. Werstak Jr.[3], Christopher M. Oswalt[4] and Charles H. Perry[2]*

[1] Pacific Northwest Research Station, US Forest Service, United States Department of Agriculture, Corvallis, OR, United States, [2] Northern Research Station, US Forest Service, United States Department of Agriculture, St. Paul, MN, United States, [3] Rocky Mountain Research Station, US Forest Service, United States Department of Agriculture, Ogden, UT, United States, [4] Southern Research Station, US Forest Service, United States Department of Agriculture, Knoxville, TN, United States

OPEN ACCESS

Edited by:
Arshad Ali, Hebei University, China

Reviewed by:
Parvez Rana, Natural Resources Institute Finland (Luke), Finland
Grace Jopaul Loubota Panzou, Marien Ngouabi University, Democratic Republic of Congo

***Correspondence:**
David M. Bell
david.bell@usda.gov
Barry T. Wilson
barry.wilson@usda.gov

Specialty section:
This article was submitted to Forest Management, a section of the journal Frontiers in Forests and Global Change

Received: *23 August 2021*
Accepted: *02 February 2022*
Published: *03 March 2022*

Citation:
Bell DM, Wilson BT, Werstak CE Jr, Oswalt CM and Perry CH (2022) Examining k-Nearest Neighbor Small Area Estimation Across Scales Using National Forest Inventory Data. Front. For. Glob. Change 5:763422. doi: 10.3389/ffgc.2022.763422

National forest inventories (NFI), such as the one conducted by the United States Forest Service Forest Inventory and Analysis (FIA) program, provide valuable information regarding the status of forests at regional to national scales. However, forest managers often need information at stand to landscape scales. Given various small area estimation (SAE) approaches, including design-based and model-based estimation, it may not be clear which is most appropriate for the user's application. In this study, our objective was to assess the uncertainty in tree aboveground live carbon (ALC) estimates for differing modes of SAE across multiple scales to provide guidance for appropriate scales of application. We calculated means and variances for ALC with design-based (Horvitz-Thompson), model-assisted (generalized regression), and model-based (k-nearest neighbor synthetic) estimators for estimation units over a range of sizes for 30 subregions in California, United States. For larger areas (10,000–64,800 ha), relative efficiencies greater than one indicated that the generalized regression estimator (GREG) generated estimates with less error than the Horvitz-Thompson estimator (HT), while the bias-adjusted synthetic estimator relative efficiency compared to either the Horvitz-Thompson or model-assisted estimators exceeded one for areas 25,000 ha and smaller. Variance estimates from the unadjusted synthetic estimator underestimated the total error, because the estimator ignores bias and thus only addresses model variance. Across scales (250–64,800 ha, 0–27 plots per area of interest), 93% of the variation in the synthetic estimator's relative standard error was explained by forest area, forest dominance, and regional variation in forest landscapes. Our results support model-assisted estimation use except for small areas where few plots (<10 in the current study) are available for generating estimates in spite of biases in estimates. However, users should exercise caution when interpreting model-based estimates of error as they may not account for model mis-specification, and thus induced bias. This research explored

multiple scales of application for SAE procedures applied to NFI data regarding carbon pools, potentially supporting a multi-scale approach to forest monitoring. Our results guides users in developing defensible estimates of carbon pools, particularly as it relates to the limits of inference at a variety of spatial scales.

Keywords: aboveground live carbon, California (USA), estimation, forest, forest inventory and analysis, national forest inventory (NFI), small area estimation, variance

INTRODUCTION

National forest inventories (NFI), such as the one conducted by the USDA Forest Service Forest Inventory and Analysis (FIA) program, provide valuable information regarding the status of forests at regional to national scales. For example, FIA data are critical to generating estimates of carbon stocks and fluxes and developing and testing ecosystem models in support of planning and reporting of carbon stocks and dynamics in the United States (Tinkham et al., 2018). Such data may also be essential for regional assessments, such as forest resource reports describing status and trends in forest attributes like forest area, tree species composition, stand structure, and forest carbon pools (e.g., Brodie and Palmer, 2020). NFI data can also be integrated with remote sensing to generate maps of forest attributes as a basis for improving the quality and efficiency of estimates (McRoberts and Tomppo, 2007; Lister et al., 2020). For example, USDA Forest Service monitoring of status and trends in late-successional and old-growth forests in Oregon, Washington, and California relies both on design-based estimates as well as predictions generated by integrating FIA data with Landsat satellite imagery using nearest neighbor imputation (Ohmann et al., 2012; Davis et al., 2015). Thus, the national consistency in NFI data generates efficiencies for assessment, planning, and monitoring (*sensu* Wurtzebach et al., 2019), but the utility of NFIs for generating reliable forest attribute estimates at stand to landscape scales remains challenging.

While NFI is vital to supporting strategic planning, forest managers often need information at stand to landscape scales in support of tactical decision making. For example, the USDA Forest Service's 2012 planning rule increases the emphasis on adaptive planning, a recognition of the central role of broad-scale monitoring, and the consideration of climate change, landscape-scale restoration, ecosystem services, and other values (Nie, 2018). This implies an increasing emphasis for National Forest planning on forest conditions from stand scales (10–100s of hectares) to landscape scales (1,000–100,000s of hectares). NFIs are not always designed to answer questions at these scales (e.g., one FIA plot per 2,428 ha) and the minimum area for estimation used by some authors can be relatively coarse (e.g., roughly 27 plots over 64,800 ha EMAP hexagons; Woodall et al., 2006; Menlove and Healey, 2020), impractical for guiding forest management decisions at stand- and landscape-scales.

Many estimation procedures utilizing NFI data are available to users interested in quantifying forest conditions over smaller areas of interest, referred to here as small area estimation (SAE) (Rao and Molina, 2015), though they may vary in terms of both variance and bias (Goerndt et al., 2012). It is important to note that SAE does not necessarily refer to a specific geographic scale of inference, but rather situations under which few if any plots are available for direct estimation based on available forest inventory data (Rao and Molina, 2015). At finer scales relevant to some types of forest management and planning questions, auxiliary data can be integrated with plot data to improve estimation or make it more flexible. Auxiliary data can be used to improve estimator efficiency through model-assisted estimation and models can be used to relate plot data to auxiliary data upon which we can base the development of forest attribute maps or hybrid approaches (Ståhl et al., 2016). From design-based to model-based inference, there is a tradeoff between reliance on probability samples vs. models as the foundation of inference, though selection of a specific estimation procedure depends on the objectives of the study.

Design-based methods provide unbiased estimators for users and are appropriate at relatively broad spatial scales where often 100s or 1,000s of plots are available. For example, the Horvitz-Thompson estimator (HT) (Horvitz and Thompson, 1952) has been commonly used for estimation of forest attribute means and variances with forest inventory data as it is simple to compute and design unbiased (Williams, 2001, Bechtold and Patterson(eds), 2005, McConville et al., 2020, Stanke et al., 2020). However, strong relationships between auxiliary data and forest attributes of interest may lead users to explore other estimation procedures. Model-assisted estimation, such as generalized regression estimators (GREGs) (Deville and Särndal, 1992), leverages models to support design-based inference, thus providing unbiased estimators that are appropriate for smaller scales than direct estimators based on existing inventory data can support (Goerndt et al., 2012; McConville et al., 2020). For example, simulation results indicated that GREGs are more efficient than Horvitz-Thompson estimators as they leverage the auxiliary information to reduce uncertainties (McConville et al., 2020). Synthetic estimation relies on a model alone and, using model-based inference, can thus provide estimates over areas with few or no plots. But bias in synthetic estimators depends on a variety of factors, including data used for fitting models, vegetation characteristics, model assumptions, and other sources of the error (McRoberts, 2012; Chen et al., 2016). For example, the development of a synthetic k-nearest neighbor estimator for variance over areas of interest provides one avenue with which to generate mean and variance estimates for small areas with insufficient plot support to leverage design-based and model-assisted methods (McRoberts et al., 2007). Therefore, while many estimation

methods for forest attributes have been used, it may not always be clear to users which is most appropriate at a given scale of inference.

While the emergence of predictive mapping of forest attributes based on NFI data or other plot networks and remote sensing (e.g., Ohmann and Gregory, 2002; Tomppo et al., 2008; Saatchi et al., 2011; Beaudoin et al., 2014; Du et al., 2014) may provide information at fine scales (e.g., 30-m pixels for maps based on Landsat satellite imagery), simply summing pixels to generate aggregate means or totals does not constitute a small area estimate as there is no characterization of uncertainty. The development of CONUS-level nearest neighbor imputed maps of forest attributes based on FIA plot data, climate, and multispectral remote sensing (e.g., Wilson et al., 2012, 2013) motivates a need to move beyond simply aggregating pixels to compare model-based estimation for *k*-nearest neighbors (*k*NN) techniques (e.g., McRoberts et al., 2007; McRoberts, 2012) with model-assisted and design-based estimation across scales and diverse forest conditions (*sensu* Ståhl et al., 2016). Such an assessment is necessary to identify whether there are clear patterns in the performance and comparability of estimation procedures as a function of estimation unit area and forest heterogeneity, which can both influence the quality of *k*NN estimates when aggregated (Bell et al., 2018). Information regarding the biophysical drivers of uncertainty could inform how users interact with the data, by providing *a priori* information on the appropriate scale of inference given their precision needs, the size of the area of interest, and the biophysical characteristics of the landscape being examined. It could also guide additional plot sampling or improvements to modeling approaches to address forest types where forest attribute estimation is particularly challenging.

Due to substantial uncertainties inherent in the estimation of carbon stocks and fluxes (Glenn et al., 2015) and the challenges of monitoring forest attributes for relatively small areas, there is a need to understand the appropriate use of differing estimation methods across scales. The foundation of that understanding should rely on assessments of variation in estimate uncertainty, both in terms of variance and bias, as the area of an estimation unit changes. In this study, our objective was to assess how tree live aboveground carbon (ALC; Mg ha^{-1}) estimates (mean and variance) differed as a function of scale (250–64,800 ha) and estimation method (design-based, model-assisted, and synthetic estimators). Specifically, we ask what is the size of a small area, and thus the size of the associated forest inventory sample, for which a model-based, synthetic *k*NN estimator (SK) would be selected in favor of either the Horvitz-Thompson or GREGs? Using this information, we aim to provide guidance to users for the appropriate scales of application for different estimation methods and a quantification of the error associated with different procedures. We also propose that a unified framework, which leverages multiple estimation procedures depending on the needs of the user, would support simple and transparent estimation, thus expanding the potential population of users of NFI data.

MATERIALS AND METHODS

Study Region

For this study, we focus on the Sierra Nevada Mountains Ecoregion (M261; Cleland et al., 1997, 2007), a 179,376 km^2 region located in California, United States (**Figure 1**). Forest landscapes in M261 are diverse, ranging from low-elevation woodlands to montane mixed conifer forests to high elevation subalpine forests. Therefore, forest landscapes include a variety of forest types characterized by different tree species, forest heterogeneity, and stand structures. As a result, forest carbon pools themselves are spatially heterogeneous, providing a useful area for assessing differing estimation procedures across various conditions.

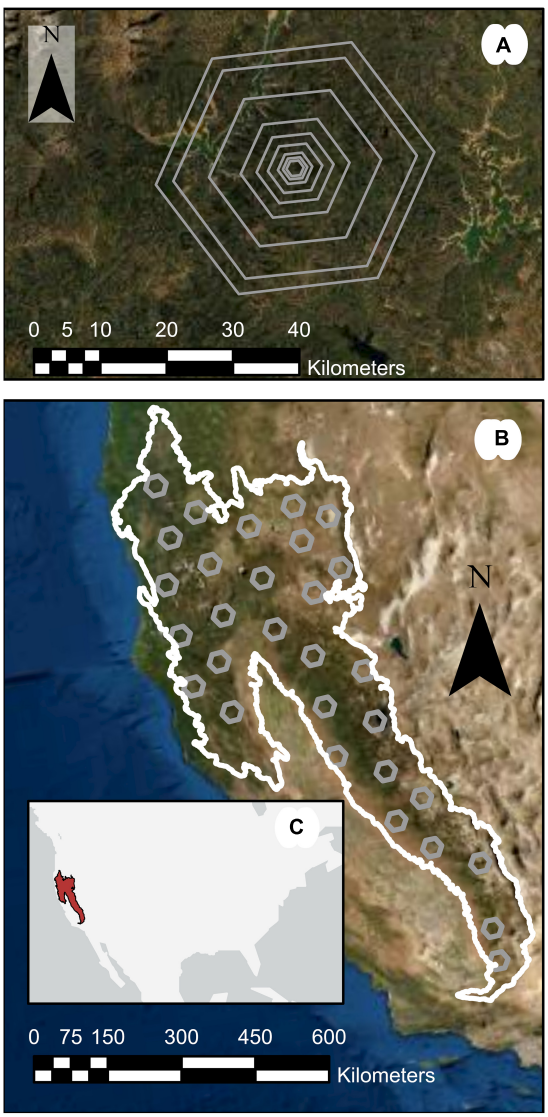


FIGURE 1 | Study region map highlighting **(A)** nested hexagonal areas of interest, **(B)** 1,000,000-ha subregions within M261, and **(C)** an inset showing the study region location in the United States.

In addition to the environmental and ecological heterogeneity in forest landscapes within M261, tracking carbon emissions and sequestration has been of major interest in California, United States. Federal, state, and municipal governments leverage numerous mitigation strategies for emissions reductions and sequestration improvements, such as California's forest offset program (Anderson et al., 2017; Cameron et al., 2017). These types of strategies require reliable information on forest carbon pools at a variety of scales, from all California forest lands down to individual property owners or management units. This study region (M261; **Figure 1**) and others would benefit greatly from an improved capacity to produce carbon pool estimates at a variety of scales as well as guidance with respect to appropriate use of NFI data provided by FIA.

Forest Inventory and Analysis Data

The FIA program is the NFI for the United States and provides a field-based assessment of forest conditions on a uniform triangular grid represented by a hexagonal lattice (one plot per 2,428-ha hexagon) across all lands regardless of ownership (i.e., non-private and private lands) in the United States (Bechtold and Patterson(eds), 2005). Through its design, the FIA plot network is well-suited for analyzing and quantifying forest conditions (e.g., volume, biomass, and carbon) at varying scales over time as the data provides a basis for unbiased estimates of forest conditions in a consistent and timely fashion (Glenn et al., 2015). As determined by aerial photography and other remote sensing, FIA locates a single plot in each 2,428-ha hexagon—either by random or collocated with a preexisting plot (Bechtold and Patterson(eds), 2005), but measures only those plots located on forestlands. On forestlands (i.e., land at least 0.4 ha in size that is at least 10% stocked with trees or formerly having such tree cover and not currently developed for a non-forest land use), field crews visit permanent ground plots and measure a suite of forest and tree variables, including tree species and diameter at breast height (dbh; 1.37 m). Plots consist four sets of nested subplots in a triangular arrangement, with trees 2.5–12.7 cm dbh measured on 2.07-m fixed radius subplots within larger 7.32-m fixed radius subplots used for trees at least 12.7 cm dbh. Therefore, field data are, at their most basic, measurements of tree species, size, and mortality status with associated scaling factors depending on size of the tree and the plot design described above. Additional measurements on FIA plots are plentiful (e.g., seedling counts, tree mortality agents, etc.), but are not used in the current study and are not discussed further.

Individual tree measurements were used to calculate ecosystem- or stand-level statistics, such as tree density, tree basal area, and species diversity. For this study, plot-level ALC was estimated using these tree diameter and species data by applying the Component Ratio Method (Jenkins et al., 2003; Woodall et al., 2011). We used tree measurements from 2014 to 2018 to represent the most recent forest conditions in the study area. While ALC estimates are themselves based on models and thus include error (Clough et al., 2016), we treat these as observations for the purposes of SAE in this study (*sensu* Wilson et al., 2013).

Auxiliary Data

To support the generation of raster maps of imputed plots for the study area by assigning a set of k plots to pixels based on their proximity in feature space (e.g., Ohmann and Gregory, 2002), we identified and developed a suite of auxiliary variables (**Figure 2**). Predictive features, or auxiliary variables, were derived from a digital elevation model (DEM), climate data, and satellite imagery, then resampled to 30-m pixel resolution. Elevation, along with its derivatives, from the 1 arc-second DEM of the National Elevation Dataset (Gesch et al., 2002) formed the set of topographic features used. Topographic derivatives included slope, compound topographic index (Beven and Kirkby, 1979), and potential annual direct incident radiation (McCune and Keon, 2002). Climate variables, derived from the Daymet Version 3 (Thornton et al., 1997, 2016) 1-km gridded monthly summaries, included mean annual growing degree days and mean annual precipitation over the nearly 40-year record. The reflectance bands for each Landsat 8 OLI collection 1 scene collected during 2014–2018 were transformed to the Tasseled Cap (TC) components of brightness, greenness, and wetness (Kauth and Thomas, 1976; Baig et al., 2014). Harmonic regression, based on a 3rd-order Fourier series (Wilson et al., 2018), was employed to characterize the mean shape of the spectro-temporal profile for each pixel and TC component over the 5-year period. A 3rd-order Fourier series requires 7 model coefficients: one for the fundamental frequency, as well as a pair for each of the three harmonics (i.e., comprised of a sine and cosine term). Given that a series was fitted to each of the three TC profiles, a total of 21 model coefficients were estimated.

Generating Tree Aboveground Live Carbon Estimates

Central to this manuscript is the comparison of multiple estimation techniques for areas of different sizes in terms of ALC mean and variance estimates. For this study, we examine the Horvitz-Thompson estimator as an example of traditional design-based estimation, GREG as an example of model-assisted estimation, and a synthetic estimator based on the k-nearest neighbors algorithm as an example of model-based estimation.

Horvitz-Thompson Estimator

Horvitz and Thompson (1952) developed an estimator that provides a general framework for direct estimation under multiple sample designs, whether or not auxiliary variables are available. The Horvitz-Thompson (HT) estimator for the population total Y is:

$$\hat{Y}_{ht} = \Sigma\, I_i d_i y_i$$

where, for the ith unit in a population of size N, I_i is a random variable that indicates whether or not the unit is in the sample, d_i is the unit's design weight, and y_i is the observation of the variable of interest for the unit. The design weight of a unit is the inverse of its probability of inclusion in the sample, π_i, or $d_i = \pi_i^{-1}$. The inclusion probabilities are determined by the sample design, which defines whether or not the sample units are to be drawn, for example, from a simple random sample (SRS), systematic

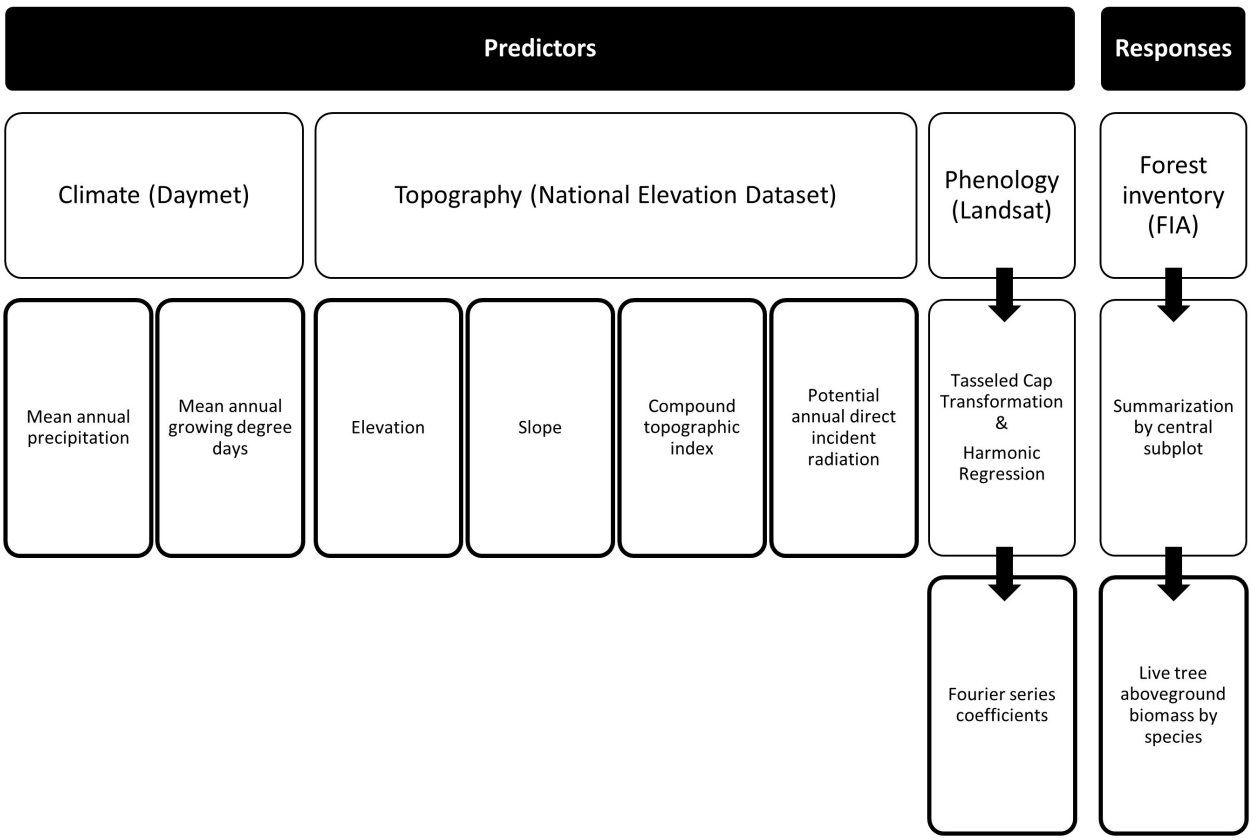


FIGURE 2 | Summarization of the source and generation of predictor (auxiliary) and response (forest inventory) variables used in the development of kNN mapping of ALC.

sample, or cluster sample. Yates and Grundy (1953) developed an estimator of the variance of the HT estimator,

$$\hat{Var}(\hat{Y}_{ht}) = \Sigma\Sigma(\pi_i\pi_j - \pi_{ij})/\pi_{ij}(y_i/\pi + y_j/\pi_j)^2$$

where π_{ij} is the joint inclusion probability of units i and j.

Generalized Regression Estimator

One approach to estimation when auxiliary variables are available is to use a model-assisted estimator. One example is known as the calibration estimator, or the generalized regression (GREG) estimator (Deville and Särndal, 1992). The GREG estimator is a generalization of a class of estimators, such as the ratio and regression estimators, that use values of one or more auxiliary variables for all population units with an assisting model to calibrate the direct estimator. It still uses the design weights and is therefore fundamentally design-based. As described in Rao (2011), suppose that the parametric superpopulation model that describes the relationship between unit-level observations of the variable of interest and the auxiliary variables is,

$$y_i = x_i{}' \beta + \varepsilon_i$$

where β are the model parameters, x_i are the auxiliary data, and ε_i is the model error. In the current study, we used the predictions from a non-parametric kNN model to replace the $\pi_i'\beta$ term (see section Synthetic k-Nearest Neighbors Estimator). The errors are assumed to be uncorrelated with mean of zero and variance proportional to a known constant q_i.

The GREG estimator of the population total Y is given by,

$$\hat{Y}_{greg} = (\hat{Y}_{ht} - \beta'\hat{X}) + \beta'X$$

where X are the known population totals of the auxiliary variables and $\hat{Y}$ and $\hat{X}$ are the corresponding estimated values for the variable of interest and auxiliary variables using the sampled units and their design weights. The variance is calculated as the Yates-Grundy variance, based on the model residuals. The working model used with the GREG estimator does not need to be a parametric linear model, and could instead be non-linear or, as in our study using the kNN algorithm, a non-parametric model.

Synthetic *k*-Nearest Neighbors Estimator

The model used as the foundation of our synthetic estimator and required as the auxiliary data for our GREG estimator (the β'X term) was based on the kNN algorithm (Fix and Hodges, Jr., 1952). The kNN imputation approach has been used extensively as a flexible, multivariate, and non-parametric method for forest attribute mapping (e.g., Ohmann and Gregory, 2002; Tomppo et al., 2008; Eskelson et al., 2009; McRoberts et al., 2011;

Wilson et al., 2013). Here, we briefly describe the development of raster maps of imputed plots based on *k*NN as well as the SK used for generating areal estimates for the mean and variance of forest attributes. For mapping ALC in our study area, the kNN algorithm was used to impute ALC data to individual pixels where no tree measurements were taken based on their similarity to forest inventory plots with relation to some set of predictors (e.g., **Figure 3**). As the non-parametric *k*NN model was fit using the FIA sample for the entire study region M261 and then used to make predictions for all population units within several domains of the study region, it forms the basis for a synthetic estimate of ALC. While the *k*NN estimator is likely nearly, but not exactly, unbiased across all units in the sample (*sensu* McRoberts et al., 2007; Magnussen et al., 2009), there is no guarantee this holds for a subsample, or any smaller domains.

An ecological ordination of tree species found in the ecological province was conducted using a canonical correspondence analysis (CCA) model (ter Braak, 1986). The set of 27 predictor variables described above used were the four topographic variables (slope, compound topographic index, and potential annual direct radiation), two climate variables, and 21 Fourier series coefficients associated with each pixel at the location of the plots measured during 2014–2018. The response variables used were live tree aboveground biomass per hectare by species for trees located on the central 7.32-m fixed radius subplot of the plots (**Figure 2**), to better match the pixel resolution of the predictor variables. There were 2,251 plots with live trees on forest conditions used to fit the CCA model.

The fitted CCA model coefficients formed the feature space for measuring proximity between each pixel and the set of measured plots (**Figure 3**; Ohmann and Gregory, 2002). All 27 orthogonal canonical variates of the CCA model were used with the *k*NN algorithm. Because the CCA model generates orthogonal axes, this approach avoids multicollinearity when assigning nearest neighbors based on the resulting feature space. There were 3,631 plot locations with a complete record of both predictor and response variables used in the imputation for M261, with non-forest conditions assigned a value of 0 for forest condition and tree variables. The value of *k* used for *k*NN regression, with predicted values being the unweighted mean of the *k*-nearest plots, excluding the nearest plot using the Manhattan distance metric, was selected to minimize mean squared error of predicted total live tree aboveground biomass. The optimal value of *k* for province M261 was 28.

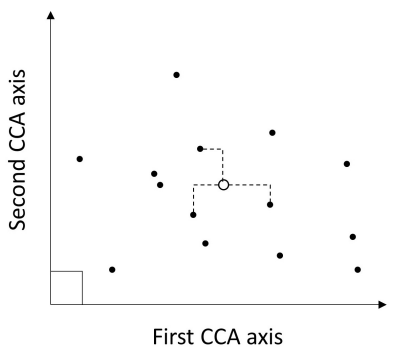


FIGURE 3 | Conceptual diagram highlighting the kNN imputation process for a simple case with two metrics (CCA axes) defining the feature space. After the CCA modeling (solid circles) are located in the feature space, then a pixel to be imputed (open circle) is placed in that space based on the auxiliary data (geospatial predictors). Distances from the focal pixel to all plots in the feature space are calculated and the *k* nearest neighbors are identified by minimizing those distances (indicated by dashed lines). Each of those nearest neighbors was imputed back to the pixel to populate a multiband raster, with the mean ALC for neighbors across bands being the predicted ALC for the pixel.

To generate model-based mean and variance estimates for ALC based on the maps of nearest neighbors, we applied an areal estimation technique for *k*NN imputation (McRoberts et al., 2007; McRoberts, 2012), and our SK. For an AOI, the mean ALC was calculated as the mean pixel-level ALC across all forested pixels and the 28 nearest neighbors for each pixel, excluding the nearest plot. Variance estimation incorporated pixel-level variance in ALC across the 28 nearest neighbors as well as covariance between pixel pairs within an AOI. The covariance between any two pixels depends on the standard deviation in ALC across neighbors for each pixel and the number of plots shared by the two pixels within the list of the $k = 28$ nearest neighbors. Thus, the SK estimator generates model-based mean and variance estimates for ALC, or any other forest attribute of interest for which plot data are available.

Because imputed maps were based on plots that were visited in the field (i.e., forestlands) and we wished to avoid extrapolating beyond the scope of our input data, we used a map of forest type groups (Wilson, 2021) to mask out non-forest lands. As a result, we assume that ALC = 0 for non-forest lands. We also apply the SK estimator only for forestlands, meaning that mean and variance in ALC is for forestlands only. To generate mean ALC for all lands, we multiplied the mean ALC from the SK estimator with the proportion of pixels within an AOI that were forested. Because we assume that ALC = 0 for non-forest lands and is thus not a random variable, variance in ALC for all lands is equal to variance in ALC for forestlands.

For this study, we implemented the SK estimator using R and ArcGIS Pro. Our implementation of the SK estimator utilized an R script embedded within an ArcGIS Pro Model Builder Toolbox. Manipulation of spatial data was handled within ArcGIS Pro 2.6 and mean and variance calculations were processed in R (4.0.2; R Core Team, 2020) within ArcGIS Pro using the arcgisbindings package (version 1.0.1.244; Esri., 2021). ArcGIS Pro required Spatial Analyst and the following R packages: doParallel (version 1.0.16; Microsoft Corporation, and Weston, 2020), raster (version 3.4-5; Hijmans, 2020), rgdal (version 1.5-23; Bivand et al., 2021), rgeos (version 0.5-5; Bivand and Rundel, 2020), and snow (version 0.4-3; Tierney et al., 2018). To accelerate processing time, we adopted a subsampling approach for pixels within an AOI, avoiding the need to assess all pairwise comparisons of individual pixels (McRoberts et al., 2007). An example R script upon which our ArcGIS Pro workflow is based can be found in **Supplementary Material 1**.

To determine whether variance estimates with the subsampling approach converge on the estimate based on all pixels (i.e., stability of variance estimator), we generated five

replicates for each of the 30 subregions of randomly selected pixels for sample proportions from 0.01 to 0.30 and AOI areas of 1,000, 5,000, and 10,000 ha. Initial testing on a high-end workstation indicated that computation time scaled with the square of the number of pixels. Given that constraint, we limited the generation of replicates to intervals of 0.01 for sample proportion between 0.01 and 0.15, but also generated replicates at sample proportions of 0.20, 0.25, and 0.30. The upper value of 0.30 was selected as it was roughly double the recommendation from a previous study (McRoberts et al., 2007). Thus, we attempted to balance reasonable computation time with appropriate coverage of lesser sample proportions which we assumed would be less stable. We then estimated variance for each replicate, sample proportion, and area combination and calculated the percent difference between that estimate and the estimate derived from the one generated when using all pixels in an AOI.

To identify the proportion of pixels that must be subsampled to generate SK variance estimates that converge on the estimate using all pixels within an AOI (i.e., a stable variance estimate), we used ordinary least squares regression (lm function; R Core Team, 2020) to fit a regression model for the absolute value of the proportional difference between sample and full variance SK estimates as a function of forest area within each AOI and proportion of pixels in the AOI sampled to generate variance estimates. Given that we were generating ALC estimates for forest lands, rather than all lands, we used forest area instead of AOI area to reflect the total number of pixels, and thus the amount of information, being used by the SK variance estimator. Additionally, forest area accounts for both AOI area as well as forest dominance (proportion of AOI that was forested). We included proportion of AOI being subsampled to represent the influence of the subsampling procedure. Data exploration indicated that the greatest predictive power for the regression model was achieved when log-transforming both response and predictor variables. We compared regression models with differing combinations of main effects (P and F) using AIC, selecting the model that minimized AIC as the best.

To solve for the proportion of pixels to sample P for values of forest area F in order to generate variance estimates within 1% of the estimate using all pixels in an AOI ($Y = 0.01$), we reorganize the regression equation as

$$P = \frac{0.01 - (\beta_0 + \beta_1 F + \beta_3 F^2)}{\beta_2}$$

When $P > 1$, we set $P = 1$ as this indicates a need to use all pixels in an AOI. Note that forest area is the product of AOI area and forest dominance, such that for any AOI area, the proportion of pixels to be sampled depended on the forest dominance in the AOI.

Comparisons of Tree Aboveground Live Carbon Estimates

To compare ALC estimates generated by the differing approaches across scales, we first defined areas of interest (AOI) across the study region in order to represent a diverse suite of forest conditions (**Figure 1A**). Across M261, we created 30 1,000,000-ha hexagons as subregions covering 500,000–1,000,000 ha each. For each subregion, we selected the 648 km^2 Environmental Monitoring and Assessment (EMAP) hexagons (White et al., 1992) overlapping the centroid of the subregion, resulting in 30 hexagons 64,800 ha in size across the study region M261 (**Figure 1B**). For FIA-based forest attribute estimation, the EMAP hexagons have been identified as providing a balance between fine spatial scale and sufficient numbers of plots to support design-based inference (Woodall et al., 2006; Menlove and Healey, 2020). We then generated hexagons at eight additional scales, centered on the same centroids: 50,000, 25,000, 10,000, 5,000, 2,500, 1,000, 500 ha, and 250 ha. These hexagons were the estimation units for this study.

We compared mean and variance estimates from each of the different methods described above (HT, GREG, and SK) only for the 10,000, 25,000, 50,000, and 64,800-ha AOIs. We compared results from the GREG and SK estimators with the HT estimator results using simple linear regression in order to roughly assess uncertainties in model-assisted and model-based estimators relative to design-based estimators. For each AOI at each scale, we also computed the relative efficiency (RE) of the SK and GREG estimators relative to the HT estimator and to each other, which is simply the ratio of the variances being compared. Two versions of the SK estimator were used for these comparisons. Unadjusted SK is the usual synthetic estimator that, by assuming the modeled relationship between predictor and response variables developed for M261 holds for all domains within it, also assumes unbiasedness for SAE. Adjusted SK uses the design-weighted estimate of the bias provided by the sample to calculate mean square error, where MSE = variance + bias^2. Under most SAE scenarios, this adjustment would not be possible because of small sample sizes.

The SK estimator was applied to all scales described above, though larger areas can require substantial processing time. It should be noted that there are many users interested in estimating means and variances for forest attributes for areas that are much smaller (<10,000 ha). Therefore, we present variance estimates for smaller areas to quantify estimate variance for the SK estimator at scales relevant to forest managers. To examine the variance of ALC estimates across a gradient of AOI area (250–64,800 ha) for the SK estimator, we developed linear mixed effect regressions of the log relative standard error (% of mean ALC estimate) as a function of log forest area, forest dominance, and a random effect for the 1,000,000-ha subregion. Forest dominance was calculated as the proportion of area in an AOI that was forested. We used all the estimates across scales (250–64,800 ha) for the 30 subregions as inputs. We then used the lme function in R (nlme package version 3.1-140; Pinheiro et al., 2019) to fit a model for log relative standard error in the ALC estimates as

$$y_{ij} \sim N\left(\gamma_0 + \gamma_1 A_{ij} + \gamma_2 D_{ij} + \alpha_j, \sigma^2\right)$$

$$\alpha_j \sim N\left(0, \tau^2\right)$$

where y_{ij} was the log relative standard error for AOI i in subregion j, γ_0, γ_1, and γ_2 were regression parameters, A_{ij} was the forest area (ha) in AOI i in subregion j, D_{ij} was the

forest dominance (unitless) in AOI i in subregion j, σ^2 was the process variance for the regression, α_j was the random effect for subregion j, and τ^2 was the variance for the random effects. We fit two additional linear regression models, one with forest area A_{ij} only and one with forest area A_{ij} and forest dominance D_{ij}, in order to examine the explanatory power of each component of the model describing the coefficient of variation.

RESULTS

Small Area Estimate Convergence

We tested the stability of the SK variance estimator as a function of the proportion of pixels sampled. We found that increasing the proportion of pixels sampled quickly led to convergence in variance estimates, supporting the use of only a subset of pixels with an AOI (**Figure 4**). We found that generating variance estimates within 1% of the estimate based on all pixels depended on several factors, including AOI area and proportion of pixels being sampled. For 10,000-ha AOIs, sampling 7% of pixels resulted in most variance estimates being within 1% of the estimate using all pixels, whereas 15% were required to ensure that most estimates were within 0.5%. Proportion of pixels sampled needed to increase for smaller areas to achieve the same convergence in variance estimates, with 5,000 ha AOIs requiring 14% and 1,000 ha AOIs requiring 30% of pixels sampled for most estimates to converge within 1%.

Our regression analysis examining the stability of variance estimates indicated that the best model for log absolute value of the proportional difference between sample and full variance estimates Y explained 33.5% of the variation and included an intercept ($\beta_0 = -1.868 \pm 0.088$ SE), log forest area F ($\beta_1 = -0.516 \pm 0.010$ SE), and log proportion pixels sampled P ($\beta_2 = -0.576 \pm 0.015$ SE).

Predicted proportion sampled increased as AOI area and forest dominance within the AOI decreased, indicating that the stability of the variance estimate depends on the number of pixels being considered. For the purposes of the rest of this study, we set the proportion of pixels sampled for estimating variance using the SK estimator to the values predicted by the 25% forest cover scenario (gray diamonds in **Figure 5**) to increase the likelihood of estimate convergence. Thus, to generate variance estimates using the SK estimator for 250, 500, 1,000, 2,500, 5,000, 10,000, 25,000, 50,000, and 64,800-ha AOIs, we used 1.00, 1.00, 0.82, 0.36, 0.19, 0.10, 0.05, 0.02, and 0.02 for proportion of pixels sampled.

Comparing Estimation Methods Across Scales

Mean ALC estimates based on GREG generally agreed with HT estimates, though that agreement decreased as AOI area decreased (**Table 1**). Regressions of GREG and HT mean ALC estimates across AOIs showed that slopes decreased from 1.010 to 0.818 and R^2 decreased from 0.972 to 0.805 as AOI area decreased from 64,800 to 10,000 ha. The regression intercept also decreased as AOI area decreased. Relative standard errors increased from 16 to 31%, while the RE of GREG vs. HT estimators increased from 1.45 to 1.54 as AOI area decreased.

Comparisons of the SK estimator with HT and GREG estimators indicated a more complex story regarding estimator performance (**Table 2**). Like GREG, regression of SK and HT mean ALC estimates indicated decreasing agreement with decreasing AOI area, with R^2 ranging from 0.920 to 0.758 for 64,800 and 10,000 ha areas, respectively. Slopes from the regression were relatively constant (0.992–1.026) for scales

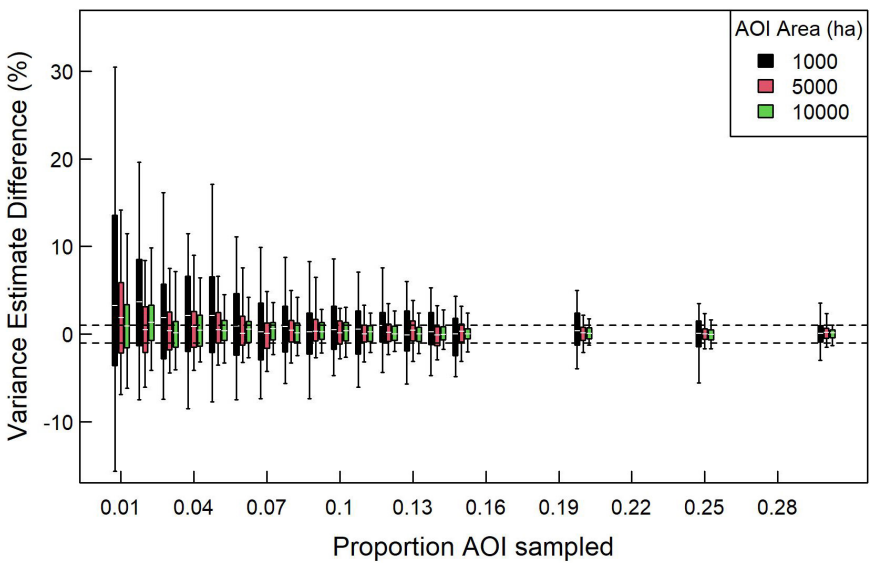


FIGURE 4 | Percent difference between AOI variance estimates based on a sample of pixels vs. using all pixels for AOIs of differing sizes. In this case, we examined 30 1,000, 2,500, and 10,000-ha AOIs distributed across the study region, each with five replicates. White horizontal lines indicate median, boxes indicate 50% intervals, and whiskers indicate 90% intervals. Dashed horizontal lines demarcate –1 and 1% differences.

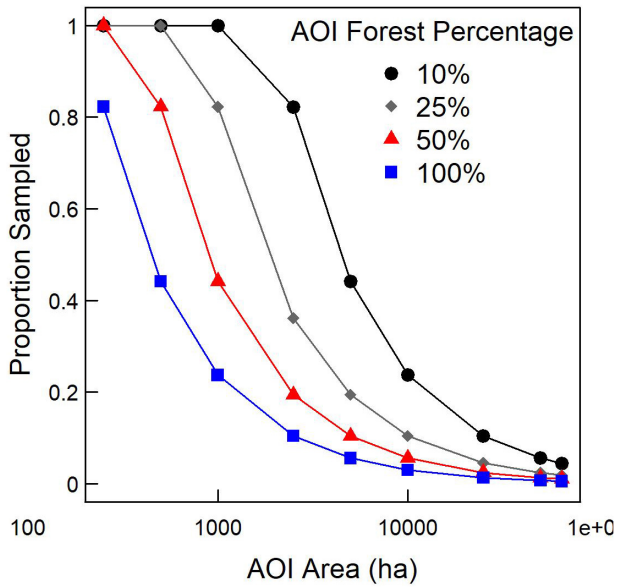


FIGURE 5 | Proportion of pixels sampled for AOIs of differing sizes and forest extent to, on average, generate variance estimates within 1% of the variance estimate using all pixels.

TABLE 1 | Regression results ($y = mx+b$) of the scatterplot of GREG (x) vs. HT (y) estimates across spatial scales, along with median relative standard error (% of estimate) and median RE of the GREG vs. HT estimator.

Area (ha)	m	$b/\bar{y}$	R^2	RSE	RE (HT)
10,000	0.818	0.215	0.805	30.977	1.542
25,000	0.954	0.113	0.928	24.566	1.556
50,000	0.991	0.093	0.946	21.234	1.406
64,800	1.010	0.062	0.972	16.148	1.450

greater than or equal to 25,000 ha, but decreased to 0.905 for 10,000 ha AOIs, while intercepts increased as AOI area decreased. Relative standard errors for unadjusted SK were smaller than other estimators (7.8–8.4%), resulting in RE compared to HT of 4.5–16.8. However, the unadjusted SK estimator RE values do not account for potential biases inherent in the synthetic approach. When we accounted for bias, using the design-weighted estimate of bias, relative root mean square error of adjusted SK increased from 24 to 33% and RE compared to HT increased from 0.649 to 2.269 as AOI area decreased from 64,800 to 10,000 ha. Similarly, adjusted SK estimator RE compared to GREG increased from 0.496 to 1.263 as AOI area decreased from 64,800 to 10,000 ha.

Across subregions, linear mixed effects modeling indicated that the relative standard error for ALC from the SK estimator decreased with forest area and forest dominance within an AOI (**Table 3** and **Figure 6A**). The linear mixed effects model including log forest area, log forest dominance, and a random effect for subregion explained 93% of the variation in the log coefficient of variation, whereas models without random effects or without random effects and forest dominance explained 59 and 39% of the variation, respectively. Mapping random effects indicated that coefficient of variation tended to be lesser in the northwestern, greater in the northeastern, and more variable in the southern portion of the study area (**Figure 6B**).

DISCUSSION

Comparing Estimators (10,000–64,800 ha)

Augmenting NFI data with auxiliary data using either model-assisted or model-based estimation facilitates SAE, but our results emphasize that the appropriate estimation procedure depends upon the area of an AOI, and thus the sample of plots, being considered. In our study, 25,000 ha was the nominal scale below which one would consider changing from the GREG to the adjusted SK estimator, or vice versa: RE for GREG was greatest among estimators tested for areas larger than 25,000 ha and RE for adjusted SK was greatest for areas less than or equal to 25,000 ha (**Tables 1**, **2**). In the case of the FIA data used in this study, 25,000 ha roughly equates to 10 plots whereas the commonly used EMAP hexagons (64,800 ha) would generally contain 27 plots. Even at the 64,800-ha scale, the GREG estimator RE compared to HT was greater than one, indicating that GREG estimators should be preferred given sufficient plot support in an AOI.

Our results highlight a fundamental limitation of the unadjusted SK estimator examined: the lack of appropriate accounting of bias. The synthetic estimator used in this study assumes unbiasedness in pixel predictions (McRoberts et al., 2007). Given that regression slopes close to one and intercepts close to zero highlight agreement, SK mean ALC estimates did not exhibit major systematic lack

TABLE 2 | Regression results ($y = mx+b$) of the scatterplot of the unadjusted synthetic (x) vs. HT (y) estimates across spatial scales, along with median relative standard error (% of estimate) and median relative efficiency of the synthetic vs. HT and GREG estimators.

	Unadjusted synthetic						Adjusted synthetic		
Area (ha)	m	$b/\bar{y}$	R^2	RSE	RE (HT)	RE (GREG)	RRMSE	RE (HT)	RE (GREG)
10,000	0.905	0.073	0.758	7.880	16.843	11.853	33.284	2.269	1.263
25,000	1.024	−0.002	0.874	8.185	11.088	6.450	24.496	1.263	1.226
50,000	1.026	−0.037	0.897	8.356	7.079	5.149	25.023	0.904	0.681
64,800	0.992	−0.047	0.920	8.383	4.542	3.536	24.033	0.649	0.496

Adjusted synthetic results are based on root mean square error, including the HT estimate of bias.

TABLE 3 | Fixed effect and mixed effect linear regression results for predicting the log coefficient of variation in tree aboveground live carbon as a function of area, dominance, and subregion.

Model	β_0	β_1	β_2	Residual standard error	Random effect variance (τ^2)	R^2
Forest area	−1.817 (0.140)	−0.104 (0.008)		0.270		0.39
Forest area and dominance	−1.551 (0.058)	−0.095 (0.007)	−0.537 (0.047)	0.222		0.59
Forest area, dominance, and subregion random effect	−1.803 (0.062)	−0.088 (0.003)	−0.223 (0.048)	0.096	0.050	0.93

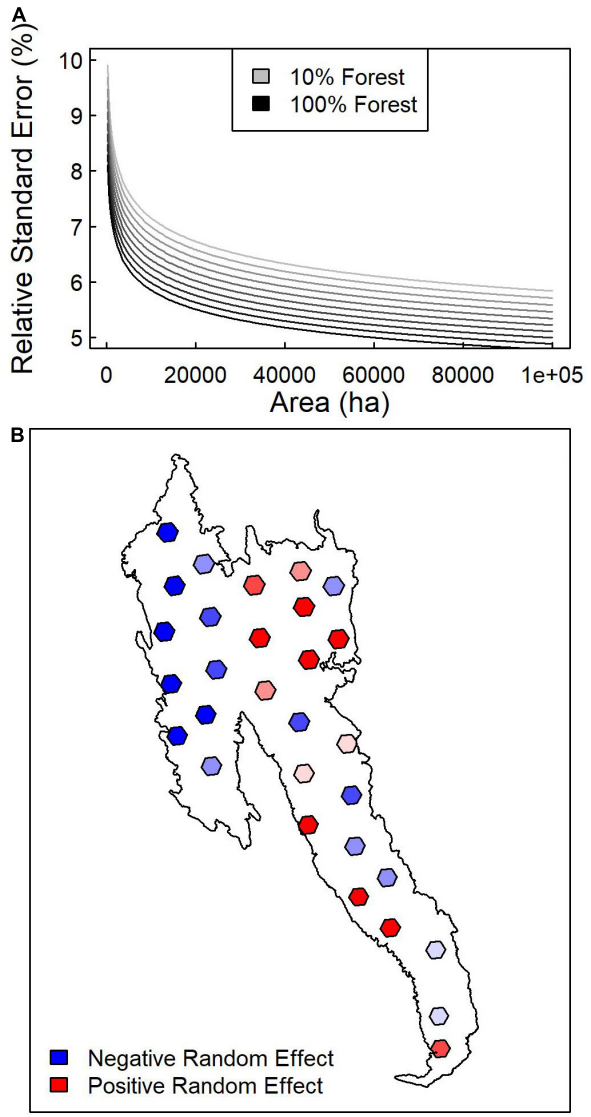


FIGURE 6 | Mixed effects regression results of relative standard error as a function of scale and landscape. **(A)** Mean predictions of coefficient of variation based on forested area and forest dominance (percent of AOI that is forested). **(B)** Spatial distribution of random effects across study area.

of fit, but R^2-values were less than those reported for GREG (**Tables 1**, **2**). This increased error in mean predictions was not reflected in the unadjusted SK variance estimates, which were considerably smaller than GREG or HT variance estimates. Previous examinations of the *k*NN synthetic estimator used in this study indicated that *k*NN using NFI data can be unbiased with respect to the sampling aspects of the estimator, but not necessarily in terms of the bias associated with model mis-specification (McRoberts et al., 2007; Magnussen et al., 2009; McRoberts, 2012). Such bias, for example, motivates the use of empirical best linear unbiased prediction (such as a Fay-Herriot model) or composite estimators that minimize MSE by finding the optimal balance between the low variance of a synthetic estimator and the unbiasedness of a direct estimator (such as a James-Stein estimator) (Breidenbach and Astrup, 2012; Rao and Molina, 2015; Mauro et al., 2017; Coulston et al., 2021). Thus, while the unadjusted SK estimator can produce variance estimates far smaller than other methods (McRoberts et al., 2007; Breidenbach et al., 2010), they reflect only model variance, not bias. While model-based variance estimates can be useful for many applications, focusing primarily on variance estimates without accounting for model bias leads to an overly optimistic view of uncertainty.

Still, it is interesting that the adjusted SK estimator RE compared to HT and GREG support the use of synthetic estimators at smaller scales where few plots were available. One might speculate that improving model fit or accounting for biases among AOIs would improve relative efficiencies and increase the nominal area for which one would select SK vs. GREG estimators. Our results imply that development or application of model-based SAE should incorporate an assessment against GREG at the scales relevant to the individual study to determine whether estimates are improved in a practical sense.

k-Nearest Neighbors Variance Estimates (250–64,800 ha)

In our study, the SK variance estimates relative to the mean ALC was predictable (**Table 3**), indicating that land cover and AOI area determine the precision of estimates derived from the SK estimator. Relative standard error for ALC depended almost entirely on forest area within the AOI, forest dominance, and biogeographic variation at the scale of our 1,000,000-ha subregions. Across forest dominance gradients, average predicted relative standard errors ranged between 0.05 and 0.07 for the largest forest areas (64,800 ha) and 0.08–0.11 for the smallest forest areas (250 ha) (**Figure 6A**). However, geographic variation in random effects imply that broad-scale variation in forest conditions explains roughly one third of the relative standard error (**Figure 6B**). This result is consistent with our previous examination of lidar-based vs. Landsat-based maps of forest

biomass which highlighted increasing differences and decreasing correlation between the two products as one shifted from coniferous to mixed broadleaf-coniferous forest landscapes (Bell et al., 2018). It has also been shown that stratification by forest type prior to lidar-based modeling improves biomass and carbon mapping (Swatantran et al., 2011; Chen et al., 2012), implying that variation in forest composition and structure influences prediction and estimation approaches.

Our results support the general application of a subsampling approach in this SK estimator using *k*NN, but the degree of subsampling depends on the area of the AOI and the amount of forest located within it. Variance estimates based on a subsample of pixels converged on the estimate using all pixels as AOI area increased (**Figure 4**), but that convergence appeared to be delayed with lesser forest dominance (**Figure 5**). The convergence still appears to be quite variable (R^2 = 0.25), indicating other factors may determine convergence for any given AOI. Based on this uncertainty in convergence, we recommend a relatively conservative approach to selecting proportion of pixels to sample. In our case, we assumed that forest dominance (proportion of pixels forested in AOI) was 0.25. Given that our results for 10,000 ha AOIs were similar to a previous study in Minnesota (15% sampling threshold; McRoberts et al., 2007) and are relatively consistent regardless of the area forested in within the AOI (134–9,740 ha), these results may be broadly applicable across landscapes. Still, further application of this method would necessitate examination of convergence as a function of other biophysical factors, such as forest type group, so that users could easily identify the appropriate sub-sampling to apply for stable variance estimation.

CONCLUSION

Forest managers increasingly rely upon spatially explicit, mapped forest attribute data as central source of information for decision-making, but assessments of uncertainty provide a much needed characterization of variance and bias in estimates of stand-, landscape-, and region-level forest attribute estimates (Tomppo et al., 2008; McRoberts, 2012). Though the choice of inferential mode, from design-based to model-based, will always depend on the question being asked (Ståhl et al., 2016), the scale of inference and characteristics of forest ecosystems appear to play a dominant role in estimate uncertainty. This study (**Table 3** and **Figure 6**) and others (e.g., Bell et al., 2015, 2018) show that spatial variation in estimated variance may be predictable as a function of biophysical characteristics of the ecosystems being studied. Advances in model-based estimation that properly account for bias in estimation error (e.g., Mauro et al., 2017; Coulston et al., 2021) could extend the scale at which these approaches outperform model-assisted estimation (e.g., > 25,000 ha). Such advances could be integrated into estimation procedures to guide the selection of estimators to fully characterize both model precision and bias, both of which impact the utility of estimates for users.

We suggest that an improved understanding of synthetic estimator uncertainty across a diversity of forest landscapes could form the basis for a simple, yet transparent workflow for forest attribute estimation. That platform could open the use of regional or national forest inventory data to a broader community of users. These improvements should, in part, aim to incorporate a proper accounting of prediction bias in model-based estimation for small areas. Furthermore, the identification of nominal scales at which users should generally switch from one estimation technique (e.g., GREG) to another (e.g., synthetic) could be incorporated into an integrated approach that guides users on the appropriate estimator to use at the scale of their AOI. However, both producers and users of estimates should bear in mind potential biases in predictions that, in the case of the SK, result in overly precise (i.e., lesser variance) estimates of forest attributes. By exploring multiple scales of application for an SAE procedure applied to NFI data regarding carbon pools, this research lays the groundwork for a multi-scale estimation framework in a simple and transparent manner that guides users in developing defensible estimates and educates users on the limits of inference at a variety of spatial scales.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.fia.fs.fed.us/.

AUTHOR CONTRIBUTIONS

DB led all components of this research. DB, BW, and CW developed and carried out the analyses, including the generation of figures and tables and wrote the majority of the manuscript. All authors contributed to the development of the research questions and the *k*NN-SAE approach and provided comments and edits on the final version of the manuscript.

FUNDING

This research was funded by the USDA Forest Service, including the collection of FIA plot data used in this study.

ACKNOWLEDGMENTS

We would like to thank Gregory Brunner, Peter Eredics, Andrew Leason, Robert Richard, and John Steffenson for their contributions to the development of *k*NN mapping for CONUS.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/ffgc.2022.763422/full#supplementary-material

REFERENCES

Anderson, C. M., Fields, C. B., and Mach, K. J. (2017). Forest offsets partner climate-change mitigation with conservation. *Front. Ecol. Environ.* 15:359–365. doi: 10.1002/fee.1515

Baig, M. H. A., Zhang, L., Shuai, T., and Tong, Q. (2014). Derivation of a tasselled cap transformation based on Landsat 8 at-satellite reflectance. *Remote Sens. Lett.* 5, 423–431. doi: 10.1080/2150704X.2014.915434

Bechtold, W. A., and Patterson, P. L. (eds) (2005). *The Enhanced Forest Inventory and Analysis Program - National Sampling Design and Estimation Procedures*. Gen. Tech. Rep. SRS-80. Asheville, NC: U.S. Department of Agriculture, 85.

Bell, D. M., Gregory, M. J., and Ohmann, J. L. (2015). Imputed forest structure uncertainty varies across elevational and longitudinal gradients in the western Cascade Mountains, Oregon, USA. *For. Ecol. Manag.* 358, 154–164. doi: 10.1016/j.foreco.2015.09.007

Bell, D. M., Gregory, M. J., Kane, V., Kane, J., Kennedy, R. E., Roberts, H. M., et al. (2018). Multiscale divergence between landsat- and lidar-based biomass mapping is related to regional variation in canopy cover and composition. *Carbon Bal. Manag.* 13:15. doi: 10.1186/s13021-018-0104-6

Beaudoin, A., Bernier, P. Y., Guindon, L., Villemaire, P., Guo, X. J., Stinson, G., et al. (2014). Mapping attributes of Canada's forests at moderate resolution through kNN and MODIS imagery. *Can. J. For. Res.* 44, 521–532. doi: 10.1139/cjfr-2013-0401

Beven, K. J., and Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology. *Hydrol. Sci. J.* 24, 43–69. doi: 10.1080/02626667909491834

Bivand, R., Keitt, T., and Rowlingson, B. (2021). *rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R Package Version 1.5-23*. Available online at: https://CRAN.R-project.org/package=rgdal (accessed December 15, 2021).

Bivand, R., and Rundel, C. (2020). *rgeos: Interface to Geometry Engine - Open Source ('GEOS'). R Package Version 0.5-5*. Available online at: https://CRAN.R-project.org/package=rgeos (accessed December 15, 2021).

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

Breidenbach, J., Nothdurftm, A., and Kändler, G. (2010). Comparison of nearest neighbor approaches for small area estimation of tree species-specific forest inventory attributes in central Europe using airborne laser scanner data. *Eur. J. For. Res.* 129, 833–846. doi: 10.1007/s10342-010-0384-1

Brodie, L. C., and Palmer, M. (2020). *California's Forest Resources, 2006-2015: Ten-Year Forest Inventory and Analysis Report*. General Technical Report PNW-GTR-983. Portland, OR: USDA, Forest Service, Pacific Northwest Research Station, 60.

Cameron, D. R., Marvin, D. C., Remucaf, J. M., and Passero, M. C. (2017). Ecosystem management and land conservation can substantially contribute to California's climate mitigation goals. *Proc. Natl. Acad. Sci. U.S.A.* 114, 12833–12838. doi: 10.1073/pnas.1707811114

Chen, Q., Laurin, G. V., Battles, J. J., and Saah, D. (2012). Integration of airborne lidar and vegetation types derived from aerial photography for mapping aboveground live biomass. *Remote Sens. Environ.* 121, 108–117. doi: 10.1016/j.rse.2012.01.021

Chen, Q., McRoberts, R. E., Wang, C., and Radtke, P. J. (2016). Forest aboveground biomass mapping and estimation across multiple scales using model-based inference. *Remote Sens. Environ.* 184, 350–360. doi: 10.1016/j.rse.2016.07.023

Cleland, D. T., Avers, P. E., McNab, W. H., Jensen, M. E., Bailey, R. G., King, T., et al. (1997). "National hierarchical framework of ecological units," in *Ecosystem Management Applications for Sustainable Forest and Wildlife Resources*, eds M. S. Boyce and A. Haney (New Haven, CT: Yale University Press), 181–200.

Cleland, D. T., Freeouf, J. A., Keys, J. E., Nowacki, G. J., Carpenter, C. A., and McNab, W. H. (2007). *Ecological Subregions: Sections and Subsections for the Conterminous United States*. General Technical Report WO-76D. Washington, DC: Washington Office, 76. doi: 10.2737/WO-GTR-76D

Clough, B. J., Russell, M. B., Domke, G. M., and Woodall, C. W. (2016). Quantifying allometric model uncertainty for plot-level live tree biomass stocks with a data-driven, hierarchical framework. *For. Ecol. Manag.* 372, 175–188. doi: 10.1016/j.foreco.2016.04.001

Coulston, J. W., Green, P. C., Radke, P. J., Prisley, S. P., Brooks, E. B., Thomas, V. A., et al. (2021). Enhancing the precision of broad-scale forestland removals estimates with small area estimation techniques. *Forestry* 94, 427–441. doi: 10.1093/forestry/cpaa045

Davis, R. J., Ohmann, J. L., Kennedy, R. E., Cohen, W. B., Gregory, M. J., Yang, Z., et al. (2015). *Northwest Forest Plan – The First 20 Years (1994-2013): Status and Trends of Late-Successional and Old-Growth Forests*. PNW-GTR-911. Portland, OR: Pacific Northwest Research Station. doi: 10.2737/PNW-GTR-911

Deville, J.-C., and Särndal, C. E. (1992). Calibration estimators in survey sampling. *J. Am. Stat. Assoc.* 87, 376–382. doi: 10.1080/01621459.1992.10475217

Du, L., Zhou, T., Zou, Z., Zhou, X., Huang, K., and Wu, H. (2014). Mapping forest biomass using remote sensing and national forest inventory in China. *Forests* 5, 1267–1283. doi: 10.3390/f5061267

Eskelson, B. N. I., Temesgen, H., Lemay, V., Barrett, T. M., Crookston, N. L., and Hudak, A. T. (2009). The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. *Scand. J. For. Res.* 24, 235–246. doi: 10.1080/02827580902870490

Esri. (2021). *arcgisbinding: Bindings for ArcGIS. R Package Version 1.0.1.244*. Available online at: http://esri.com

Fix, E., and Hodges, J. L. Jr. (1952). *Discriminatory Analysis-Nonparametric Discrimination: Small Sample Performance*. Berkeley, CA: California Univ Berkeley. doi: 10.1037/e471672008-001

Gesch, D., Oimoen, M., Greenlee, S., Nelson, C., Steuck, M., and Tyler, D. (2002). The national elevation dataset. *Photogramm. Eng. Remote Sensing* 68, 5–32.

Glenn, C. A., Waddell, K. L., Stanton, S. M., and Kuegler, O. (eds) (2015). *California's Forest Resources: Forest Inventory and Analysis, 2001–2010*. Gen. Tech. Rep. PNW-GTR-913. Portland, OR: U.S. Department of Agriculture, 293.

Goerndt, M. E., Monleon, V. J., and Temesgen, H. (2012). Small-area estimation of county-level forest attributes using ground data and remote sensed auxiliary information. *For. Sci.* 59, 536–548. doi: 10.5849/forsci.12-073

Hijmans, R. J. (2020). *raster: Geographic Data Analysis and Modeling. R Package Version 3.4-5*. Available online at: https://CRAN.R-project.org/package=raster (accessed January 22, 2022).

Horvitz, D. G., and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* 47, 663–685. doi: 10.1080/01621459.1952.10483446

Jenkins, J. C., Chojnacky, D. C., Heath, L. S., and Birdsey, R. A. (2003). National scale biomass estimators for United States tree species. *For. Sci.* 49, 12–35.

Kauth, R. J., and Thomas, G. S. (1976). "The tasselled cap–a graphic description of the spectral-temporal development of agricultural crops as seen by landsat," in *Proceedings, Symposium on Machine Processing of Remotely Sensed Data*, West Lafayette, IN, 159.

Lister, A. J., Andersen, H., Frescino, T., Gatziolis, D., Healey, S., Heath, L. S., et al. (2020). Use of remote sensing data to improve efficiency of national forest inventories: a case study from the United States national forest inventory. *Remote Sens.* 11:1364. doi: 10.3390/f11121364

Magnussen, S., McRoberts, R. E., and Tomppo, E. O. (2009). Model-based mean square error estimators for k-nearest neighbour predictions and applications using remotely sensed data for forest inventories. *Remote Sens. Environ.* 113, 476–488. doi: 10.1016/j.rse.2008.04.018

Mauro, F., Monleon, V. J., Temesgen, H., and Ford, K. R. (2017). Analysis of area level and unit level models for small area estimation in forest inventories assisted with LiDAR auxiliary information. *PLoS One* 12:e0189401. doi: 10.1371/journal.pone.0189401

McConville, K. S., Moisen, G. G., and Frescino, T. S. (2020). A tutorial on model-assisted estimation with application to forest inventory. *Forests* 11:244. doi: 10.3390/f11020244

McCune, B., and Keon, D. (2002). Equations for potential annual direct incident radiation and heat load. *J. Veg. Sci.* 13, 603–606. doi: 10.1111/j.1654-1103.2002.tb02087.x

McRoberts, R. E., and Tomppo, E. O. (2007). Remote sensing support for national forest inventories. *Remote Sens. Environ.* 110, 412–419. doi: 10.1016/j.rse.2006.09.034

McRoberts, R. E., Tomppo, E. O., Finley, A. O., and Juha, H. (2007). Estimating areal means and variances of forest attributes using the k-nearest neighbors technique and satellite imagery. *Remote Sens. Environ.* 111, 466–480. doi: 10.1016/j.rse.2007.04.002

McRoberts, R. E., Magnussen, S., Tomppo, E. O., and Chirici, G. (2011). Parametric, bootstrap, and jackknife variance estimators for the k-nearest neighbors technique with illustrations using forest inventory and satellite

image data. *Remote Sens. Environ.* 115, 3165–3174. doi: 10.1016/j.rse.2011.07.002

McRoberts, R. E. (2012). Estimating forest attribute parameters for small areas using nearest neighbor techniques. *For. Ecol. Manag.* 272, 3–12. doi: 10.1016/j.foreco.2011.06.039

Menlove, J., and Healey, S. P. (2020). A comprehensive forest biomass dataset for the USA allows customized validation of remotely sensed biomass estimates. *Remote Sens.* 12:4141. doi: 10.3390/rs12244141

Microsoft Corporation, and Weston, S. (2020). *doParallel: Foreach Parallel Adaptor for the 'Parallel' Package. R Package Version 1.0.16*. Available online at: https://CRAN.R-project.org/package=doParallel (accessed October 16, 2020).

Nie, M. (2018). The forest service's 2012 planning rule and its implementation: federal advisory committee member perspectives. *J. For.* 117, 65–71. doi: 10.1093/jofore/fvy055

Ohmann, J. L., and Gregory, M. J. (2002). Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, USA. *Can. J. For. Res.* 32, 725–741. doi: 10.1139/x02-011

Ohmann, J. L., Gregory, M. J., Roberts, H. M., Cohen, W. B., Kennedy, R. E., and Yang, Z. (2012). Mapping change of older forest with nearest neighbor imputation and landsat time-series. *For. Ecol. Manag.* 272, 13–25. doi: 10.1016/j.foreco.2011.09.021

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team. (2019). *_nlme: Linear and Nonlinear Mixed Effects Models. R Package Version 3.1-140*. Available online at: https://CRAN.R-project.org/package=nlme (accessed January 13, 2022).

R Core Team (2020). *R: A Language and Environment for Statistical Computing, Version 4.0.2*. Vienna: R Foundation for Statistical Computing.

Rao, J. N. K. (2011). Impact of frequentist and Bayesian methods on survey sampling practice: a selective appraisal. *Stat. Sci.* 26, 240–256. doi: 10.1214/10-STS346

Rao, J. N. K., and Molina, I. (2015). *Small Area Estimation*, 2nd Edn. Hoboken, NJ: John Wiley & Sons, Inc, 441. doi: 10.1002/9781118735855

Saatchi, S. S., Harris, N. L., Brown, S., Lefsky, M., Mitchard, E. T. A., Alas, S., et al. (2011). Benchmark map of forest carbon stocks in tropical regions across three continents. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9899–9904. doi: 10.1073/pnas.1019576108

Ståhl, G., Saarela, S., Schenll, S., Holm, S., Breidenbach, J., Healey, S. P., et al. (2016). Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation. *For. Ecosyst.* 3:5. doi: 10.1186/s40663-016-0064-9

Stanke, H., Finley, A. O., Weed, A. S., Walters, B. F., and Domke, G. M. (2020). rFIA: an R package for estimation of forest attributes with the US forest inventory and analysis database. *Environ. Model. Softw.* 127:104664. doi: 10.1016/j.envsoft.2020.104664

Swatantran, A., Dubayah, R., Roberts, D., Hofton, M., and Blair, J. B. (2011). Mapping biomass and stress in the Sierra Nevada using lidar and hyperspectral data fusion. *Remote Sens. Environ.* 115, 2917–2930. doi: 10.1016/j.rse.2010.08.027

ter Braak, C. J. F. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* 67, 1167–1179. doi: 10.2307/1938672

Thornton, P. E., Running, S. W., and White, M. A. (1997). Generating surfaces of daily meteorological variables over large regions of complex terrain. *J. Hydrol.* 190, 214–251. doi: 10.1016/S0022-1694(96)03128-9

Thornton, P. E., Thornton, M. M., Mayer, B. W., Wei, Y., Devarakonda, R., Vose, R. S., et al. (2016). *Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 3*. Oak Ridge, TEN: ORNL DAAC.

Tierney, L., Rossini, A. J., Li, N., and Sevcikova, H. (2018). *snow: Simple Network of Workstations. R Package Version 0.4-3*. Available online at: https://CRAN.R-project.org/package=snow (accessed October 27, 2021).

Tinkham, W. T., Mahoney, P. R., Hudak, A. T., Domke, G. M., Falkowski, M. J., Woodall, C. W., et al. (2018). Applications of the United States forest inventory and analysis dataset: a review and future directions. *Can. J. For. Res.* 48, 1251–1268. doi: 10.1139/cjfr-2018-0196

Tomppo, E., Olsson, H., Ståhl, G., Nilsson, M., Hagner, O., and Katila, M. (2008). Combining national forest inventory field plots and remote sensing data for forest databases. *Remote Sens. Environ.* 112, 1982–1999. doi: 10.1016/j.rse.2007.03.032

White, D., Kimerling, J., and Overton, S. (1992). Cartographic and geometric components of a global sampling design for environmental monitoring. *Cartogr. Geogr. Inf. Syst.* 19, 5–21. doi: 10.1559/152304092783786636

Williams, M. S. (2001). Comparison of estimation techniques for a forest inventory in which double sampling for stratification is used. *For. Sci.* 47, 563–576.

Wilson, B. T. (2021). *Forest Type Groups of the Continental United States [Map]*. Available online at: https://www.arcgis.com/home/item.html?id=fe77a8a503ca4b9ba1ee2ef3c8ff7b19 (accessed June 29, 2021).

Wilson, B. T., Lister, A. J., and Riemann, R. I. (2012). A nearest-neighbor imputation approach to mapping tree species over large areas using forest inventory plots and moderate resolution raster data. *For. Ecol. Manag.* 271, 182–198. doi: 10.1016/j.foreco.2012.02.002

Wilson, B. T., Woodall, C. W., and Griffith, D. M. (2013). Imputing forest carbon stock estimates from inventory plots to a nationally continuous coverage. *Carbon Bal. Manag.* 8:1. doi: 10.1186/1750-0680-8-1

Wilson, B. T., Knight, J. F., and McRoberts, R. E. (2018). Harmonic regression of Landsat time series for modeling attributes from national forest inventory data. *ISPRS J. Photogr. Remote Sens.* 137, 29–46. doi: 10.1016/j.isprsjprs.2018.01.006

Woodall, C. W., Heath, L. S., Domke, G. M., Nichols, M., and Oswalt, C. (2011). *Methods and Equations for Estimating Aboveground Volume, Biomass, and Carbon for Forest Trees in the U.S. Forest Inventory, 2010*. General Technical Report NRS-88. Newtown Square, PA: U.S. Department of Agriculture. doi: 10.2737/NRS-GTR-88

Woodall, C. W., Perry, C. H., and Miles, P. D. (2006). The relative density of forests in the United States. *For. Ecol. Manag.* 226, 368–372. doi: 10.1016/j.foreco.2006.01.032

Wurtzebach, Z., DeRose, R. J., Bush, R. R., Goeking, S. A., Healey, S., Menlove, J., et al. (2019). Supporting national forest system planning with forest inventory and analysis data. *J. For.* 2019, 1–18.

Yates, F., and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *J. R. Stat. Soc. Ser. B* 15, 253–261. doi: 10.1111/j.2517-6161.1953.tb00140.x

Author Disclaimer: The findings and conclusions in this report are those of the author(s) and do not necessarily represent the views of the USDA Forest Service.

Conflict of Interest: The 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Forests and Global Change

REVIEW
published: 16 March 2022
doi: 10.3389/ffgc.2022.813569

Review and Synthesis of Estimation Strategies to Meet Small Area Needs in Forest Inventory

Garret T. Dettmann[1], Philip J. Radtke[1], John W. Coulston[2], P. Corey Green[1], Barry T. Wilson[3] and Gretchen G. Moisen[4]*

[1] Department of Forest Resources and Environmental Conservation, Virginia Tech, Blacksburg, VA, United States, [2] U.S. Forest Service, Blacksburg, VA, United States, [3] U.S. Forest Service, St. Paul, MN, United States, [4] U.S. Forest Service, Ogden, UT, United States

Small area estimation is a growing area of research for making inferences over geographic, demographic, or temporal domains smaller than those in which a particular survey data set was originally intended to be used. We aimed to review a body of literature to summarize the breadth and depth of small area estimation and related estimation strategies in forest inventory and management to-date, as well as the current state of terminology, methods, concerns, data sources, research findings, challenges, and opportunities for future work relevant to forestry and forest inventory research. Estimation methodologies explored include direct, indirect, and composite estimation within design-based and model-based inference bases. A variety of estimation methods in forestry have been applied to extensive multi-resource inventory systems like national forest inventories to increase the precision of estimates on small domains or subsets of the overall populations of interest. To avoid instability and large variances associated with small sample sizes when working with small area domains, forest inventory data are often supplemented with information from auxiliary sources, especially from remote sensing platforms and other geospatial, map-based products. Results from many studies show gains in precision compared to direct estimates based only on field inventory data. Gains in precision have been demonstrated in both project-level applications and national forest inventory systems. Potential gains are possible over varying geographic and temporal scales, with the degree of success in reducing variance also dependent on the types of auxiliary information, scale, strength of model relationships, and methodological alternatives, leaving considerable opportunity for future research and growth in small area applications for forest inventory.

Keywords: small area estimation, model-assisted estimation, forest sampling, geospatial data, design-based inference, model-based inference

OPEN ACCESS

Edited by:
Isabel Cañellas, Centro de Investigación Forestal (INIA), Spain

Reviewed by:
Hubert Hasenauer, University of Natural Resources and Life Sciences Vienna, Austria
Antonio Carlos Ferraz Filho, Federal University of Piauí, Brazil

***Correspondence:**
Philip J. Radtke
pradtke@vt.edu

Specialty section:
This article was submitted to Forest Management, a section of the journal Frontiers in Forests and Global Change

Received: *12 November 2021*
Accepted: *14 February 2022*
Published: *16 March 2022*

Citation:
Dettmann GT, Radtke PJ, Coulston JW, Green PC, Wilson BT and Moisen GG (2022) Review and Synthesis of Estimation Strategies to Meet Small Area Needs in Forest Inventory. Front. For. Glob. Change 5:813569. doi: 10.3389/ffgc.2022.813569

1. INTRODUCTION

The frequency and sophistication of statistical methods in forest inventory have grown steadily since their earliest adoption by forest researchers, with an overall goal of providing information of sufficiently high quality to inform decision-making (Schumacher, 1945). One ongoing trend involves the use of data collected as a part of broad regional or national forest inventories to produce estimates for areas smaller than the surveys were originally designed to address

(Magnussen et al., 2014). This trend reflects a situation described by W. A. Fuller in a plenary presentation delivered to a 1998 workshop on Environmental Monitoring Surveys Over Time hosted by the University of Washington, Seattle (Fuller, 1999):

> "The client will always require more than is specified at the design stage. For example, the client will explain that they require estimates only at the regional or national level and then, when data are available, ask for county estimates."

While the quotation cites a typical circumstance, many situations arise in forest resource assessment where stakeholders recognize that data from multiple sources and scales could be leveraged to give improved estimates for increasingly small subsets of survey populations—whether based on geographic areas, time periods, or demographic subsets of larger populations from which sample data were collected.

One means of addressing this need is through small area estimation (SAE), a set of statistical methods aimed at providing estimates of parameters of interest for population subsets known as small area domains, typically by linking information from auxiliary sources with sample observations gathered over a larger population that encompasses multiple small area domains. A small area in this case can be described as any geographic, temporal, or categorical domain for which an established approach for direct estimation does not provide adequate precision (Rao and Molina, 2015). Usual quantities of interest include finite population parameters such as area totals or means, especially where tolerances for estimator accuracy have been specified as a part of the sample design. Domains in SAE are typically subsets of larger survey populations, such as those found in regional or national economic, health, or agricultural surveys focused on multiple attributes of interest (Schreuder et al., 1993). These surveys may involve decades of repeated sampling and data collection targeting dozens of attributes, or be limited to a single attribute observed at just one point in time.

SAE methods seek to improve the precision of estimates for small area domains of interest (DOI) using data observed from other domains to increase the amount of sample information available, an approach often described as "borrowing strength." The indirect (i.e., outside-of-domain) data are generally linked to small area DOI through one or more auxiliary variables and a model relationship that holds across multiple domains. In this sense, the domain (d) direct data ($y \in d$) are linked to the indirect data ($y \notin d$) *via* a model relationship involving the auxiliary data (x) and corresponding observations of y. In the example Fuller (1999) described, counties of interest are the small area domains, with the sample data collected in a specific county serving as its source of direct information. Sample observations from surrounding areas—including other counties—are the source of indirect data. In a national forest inventory (NFI) application consistent with Fuller's example both $y \in d$ and $y \notin d$ would be collected on observational units selected by statistical sampling. Auxiliary information to construct a model for $y \sim x$ might come from remote sensing or other geospatial data sets separate from the NFI sample observations, or from other sources including surveys of forest landowners, agencies, or enterprises that keep records of timber harvesting, tree-planting, or other management-related activities. Other arrangements and sources of information are possible, but the overall pattern of direct, indirect, and auxiliary information, and a model $y \sim x$ is present in most SAE applications (Rao, 2008).

Methods now widely associated with SAE largely appeared in technical literature beginning in the 1970s to address needs for increased accuracy when estimating for small area domains within population, social-economic, and public health surveys (Federal Committee on Statistical Methodology, 1993). Since then, SAE has been adopted in applications aimed at estimating incomes (Fay and Herriot, 1979), census groupings, and crop areas (Battese et al., 1988), among others. Methods for SAE have continued to develop over time as new statistical and computational tools have become available, together with widespread availability and cost-effectiveness of modern data sets. Remote sensing technologies such as satellite imagery or aerial laser scanning (ALS) have played a key role in accelerating the application of SAE methodologies to forest inventory settings (Pfeffermann, 2002, 2013; Sugasawa and Kubokawa, 2020; Coulston et al., 2021).

Interest in applying SAE across disciplines has grown over time but most applications in forest inventory began to appear in published work over the past several decades (Burk and Ek, 1982; Anderson and Breidenbach, 2007; Breidenbach et al., 2010; Goerndt et al., 2011). SAE reduces forest inventory estimator errors for small area domains, offering an efficient and cost-effective option for reducing uncertainty compared to increasing sampling intensity by installing additional forest-inventory field plots (Magnussen et al., 2014). SAE techniques have been used to enhance precision of NFI-derived estimates (Breidenbach and Astrup, 2012; Frank et al., 2020), in forest stand inventories (Ver Planck et al., 2018), and from surveys of wood processors or commercial landowner inventories (Green et al., 2020; Coulston et al., 2021), but are not limited to those uses (Affleck and Gregoire, 2015). The ability of SAE to increase estimator precision in small areas where data are otherwise too sparse to satisfy tolerance specifications makes it attractive for applications in forest inventory (Guldin, 2021). For example, the Norwegian NFI has employed national canopy height maps from aerial remote sensing as auxiliary data sources since about 2010 to address needs for better local information in producing municipal forest statistics and forest-management related inventories (Astrup et al., 2019; Breidenbach et al., 2020). SAE has been used with forest inventory data from the U.S. Department of Agriculture Forest Service Forest Inventory and Analysis (FIA) program to generate estimates of forest attributes in small areas such as biofuel supply areas around co-firing power plants (Goerndt et al., 2019). In forests where field plots can be precisely referenced to high-quality geospatial auxiliary data (e.g., ALS), SAE can provide increased precision of estimates for arbitrarily small spatial areas—accounting for spatial correlations in sample data when warranted—even where no direct sample data lie within some DOI (Babcock et al., 2018; Pascual et al., 2018). Current research aims to combine forest biomass data from field plots with canopy-height measurements from the

NASA GEDI spaceborne lidar platform and other remotely-sensed auxiliary information in a SAE framework, signaling the first efforts to obtain global scale forest biomass estimates in an inferential framework (Patterson et al., 2019).

While the potential value of SAE for use in forest inventory and monitoring is high, the number of applications reported in technical literature to date is relatively small. A main objective of this work is to provide an overview of published findings of small area applications in forest inventory including relevant considerations for their implementation in other, perhaps novel, settings. An underlying goal is to provide a backdrop of SAE-related concepts and terminology, as the clear and consistent use of statistical language is important to wider adoption of these tools in future work. A further objective is to provide clarification, without a heavy emphasis on mathematical statistics, where ample terminology and notation can pose challenges to those interested in pursuing small area applications, possibly for the first time. We begin in Section 2 with background on relevant statistical paradigms of design- and model-based inference relevant to SAE, including an important extension of design-based inference known as model-assisted estimation. Section 3 presents terminology for direct and indirect estimators along with an overview of composite estimators used in a majority of SAE methods classified as either unit-level or area-level methods. In Section 4, we summarize key findings of published SAE research in forestry, with synthesis and comparison to design-based approaches including model-assisted estimation. Section 4 also includes some discussion of variance estimators in small-area inventory applications along with emerging topics in SAE. We conclude in Section 5 with some take-home findings of the work.

2. BACKGROUND

2.1. Design-Based Estimators

The design-based framework for inference from statistical sampling is a pillar of many SAE procedures, especially area-level estimators that will be discussed in Section 3.3 below. Sampling is likely familiar to most forest inventory specialists as it provides the basis for establishing statistical properties of estimators in well-designed inventories (Shiver and Borders, 1996; Gregoire and Valentine, 2007; Thompson, 2012). The sample design assigns probabilities to population units (e.g., plots, trees, etc.) in the sampling frame for being selected in a particular random draw from a finite population, with the overall probability of any unit being included in a sample determined from its selection probability in the context of the sampling scheme. While the attributes of each unit—whether sampled or not—are treated as fixed quantities, randomness in design-based methods arises through the process of sampling. Gregoire (1998) explained this detail stating, "in the design-based framework, the population is regarded as fixed whereas the sample is regarded as a realization of a stochastic process." Design-based methods rely on the randomization distribution of sampling and possible estimates that could be obtained by following the sample design and its implementation in the sampling scheme. A key consequence is the lack of reliance on mathematical assumptions of how elements in the population are distributed, or of model relationships assumed about how two or more variables in the population are related (Sterba, 2009).

The usual goal of design-based sampling and subsequent estimation is to obtain reliable estimates of finite population parameters in an inferential framework. In forest inventories, population totals, means, or proportions for various attributes of interest are typical subjects of estimation. No assumptions about the underlying structure or distributions of population units being surveyed are required for valid inference in the design-based framework. Design-based estimators compatible with their sample designs are design-unbiased, meaning the expected value of the estimator over all possible samples equals the true population parameter being estimated. It follows that design-based estimators are design consistent, such that, "both the design bias and the variance go to zero as the sample size increases" (Skinner and Wakefield, 2017).

The Horvitz-Thompson (H-T) estimator is a design-based estimator widely used in forest inventory and introduced in many texts starting with simple random sampling. Any finite population total for attribute y can be estimated using the H-T estimator

$$\hat{\tau}_y = \sum_{i \epsilon s} \frac{y_i}{\pi_i} \tag{1}$$

where $\hat{\tau}_y$ is the estimated population total, s denotes the set of observed values in the sample, y_i is the observed value of attribute y on the i^{th} sample unit, and π_i is the probability that y_i is included in s. In this form design-based estimators rely entirely on observed sample data and sample design weights, i.e., the inclusion probabilities in the denominator of Equation (1), to estimate an attribute of interest. Standard errors of an estimate require pairwise joint inclusion probabilities $P(y_i \cap y_j)$ $(i \neq j)$, calculated as the product of π_i and π_j when random sampling is from non-overlapping population units. Thus, the inference base for H-T estimator makes use of properties of the sample design such as a sampling scheme that draws samples according to design weights, independence of sample observations, and sampling distribution properties including applicability of Student's t-distribution and the Central Limit Theorem (Sterba, 2009). Sampling methods relying on a design-based framework do so largely for its desirable properties of unbiasedness and asymptotic consistency, often sought in inventory settings. These methods, however, rely on direct data—values of y observed directly by sampling from the population of interest—which can be expensive to obtain. Any need for estimates on small subsets of the population will likely result in a need for increased sampling intensity and additional expense (Fuller, 1999).

2.2. Model-Assisted Estimators

Models can be used within a design-based framework to improve estimates in what are often called model-assisted estimators (Särndal et al., 1992; McConville et al., 2017, 2020). Like the H-T estimator (Equation 1), model-assisted estimators are direct estimators in that they rely on values of y only from population DOI. One example is post-stratification which has formed the

basis of forest inventory estimation strategies in the U.S. for many decades (Bechtold and Patterson, 2005). Among the simplest model-assisted estimators are linear combinations of auxiliary variables $\boldsymbol{x}$ – either univariate or multivariate – available for all sampled units and having known population means ($\mu_{\boldsymbol{x}}$) or totals ($\tau_{\boldsymbol{x}}$). Using the H-T estimator (1), the model-assisted estimator can be written as in Stahl et al. (2016)

$$\hat{\tau}_y^{MA} = \sum_{i \in U} \hat{y}_i + \sum_{i \in s} \frac{y_i - \hat{y}_i}{\pi_i} \tag{2}$$

where $\hat{\tau}_y^{MA}$ is the estimated model-assisted population total for attribute y, U denotes the (universal) set of all population units, and predicted values are obtained from a model, i.e., $\hat{y}_i = m(\boldsymbol{x}_i)$, oftentimes a linear regression model. The first term on the right-hand side of Equation (2) requires at least that population totals $\tau_{\boldsymbol{x}}$ for all predictors are known, while the second summand is a H-T estimator of sample deviations from model-predicted values, which should be positive if the model underpredicts for $y \in s$ and negative if the model overpredicts.

Särndal et al. (1992) presented an unbiased estimator for the variance of Equation (2) when the coefficients of a linear model m are known constants, an application known as the difference estimator. The difference estimator is design-based because both the estimator $\hat{\tau}_y^{MA}$ and it's estimated variance are unbiased over all possible samples due to the sample design. Thus, although $\hat{\tau}^{MA}$ is a design-based estimator, the phrase "model-assisted" is used to indicate that a model relationship is involved in the estimation of τ.

The regression estimator, or generalized regression estimator (GREG) is based on the same form as Equation (2) in cases where $\hat{y}$ is predicted from a regression model. GREG has been applied in forest inventory solutions that include post-stratification, ratio estimators, LASSO, ridge, and elastic-net regressions (Stehman, 2009; McConville et al., 2020). Extensions using non-linear, semiparametric, and non-parametric predictive modeling techniques have also been demonstrated in forest inventory applications (Opsomer et al., 2007; Tipton et al., 2013; Kangas et al., 2016). Although taking the same form as the difference estimator, as distinguished by Särndal et al. (1992), the linear predictor $m(\boldsymbol{x})$ in GREG consists of a regression model $y = \boldsymbol{x}\boldsymbol{\beta}$ fit to paired sample values of $\boldsymbol{x}$ and y. Some authors distinguish between the difference estimator and GREG as either external or internal, respectively, based on the sources of information from which their coefficients are derived (Kangas et al., 2016; Stahl et al., 2016). Still within the realm of direct, model-assisted estimators, the term modified GREG is used to distinguish models where coefficients are derived using data outside the population of interest (Rao and Molina, 2015). For GREG to be approximately design-unbiased, sample inclusion probabilities for $\boldsymbol{x}$ and y are used in a weighted-least-squares fit of the model to sample observations. Inclusion probabilities are also used in an approximate variance formulation for the GREG estimator detailed by Särndal et al. (1992, ch. 6). Confidence intervals can be reliably constructed for these estimators, but are usually used for major domains as their variance can be large or unstable in domains having small sample sizes (Särndal, 1984; Lehtonen and Veijanen, 2009). We note that design-based approaches including H-T and model-assisted estimation are sometimes categorized as SAE (see Figure 2.1 in Rahman and Harding, 2017; Hill et al., 2021); however, they are often used where interest lies in only a single population domain, such as in the methods demonstrated by McConville et al. (2020) for estimating forest attributes in a single county in Utah, USA.

2.3. Model-Based Estimators

A second pillar of many SAE procedures is the model-based framework for inference, which we introduce here as being distinct from a purely design-based framework, including model-assisted estimators described in Section 2.2. In model-based estimation, univariate or multivariate statistical models are formulated to establish the basis for assigning probabilities to observed data, for characterizing probabilistic relationships between variables, or for both. Unlike the fixed-quantity view of population units in design-based estimation, model-based approaches treat observable units in a population as realizations or instances of random variables that underlie the observable population. For this reason these conceptual models of population variables and their statistical distributions are sometimes referred to as "superpopulation models" but are just as often simply called models (Gregoire, 1998; Skinner and Wakefield, 2017).

Model-based estimators can be useful when random sampling is impractical, or when assigning inclusion probabilities to sample observations requires some assumption about the statistical or spatial distributions of population units (e.g, Radtke and Bolstad, 2001). Sterba (2009) emphasized the utility of model-based estimators in surveys where selection probabilities were unknown, particularly in some forms of non-random sampling. Perhaps most important in the context of this review, models serve an important purpose in providing a statistical link between sample observations of y and auxiliary data $\boldsymbol{x}$, to make use of auxiliary information in ways that increase the precision of parameter estimates. The parameters of interest in forest inventory likely include population totals or means; additionally, parameters of the models themselves, such as regression parameters or ratios linking auxiliary and sample data may be of interest (Gregoire, 1998). Model-based methodology includes many tools to assist with identifying best models to fit data, to estimate parameters and standard errors for model predictors, and to adopt complex models and analyze more complex data structures than might be possible from sampling alone (Rao and Molina, 2015). While the complex and expansive nature of model-based methodologies leads to a wide array of SAE techniques that make use of models, they place a somewhat greater burden on analysts to adequately address model selection, goodness of fit, or checking of model assumptions—all model concepts that would not be required in design-based approaches.

3. TERMINOLOGY AND METHODS

In Sections 2.1–2.3, estimation was presented as a means of obtaining reliable information about population parameters of interest—either for fixed finite populations or underlying model

superpopulations—along with rationale for making statistical inferences under design-based and model-based paradigms. In moving to settings where the goal involves making reliable estimates for subsets or domains of interest within broader populations, the presentation will be expanded to include ideas and terminology better suited to the purposes and practice of SAE.

3.1. Direct, Indirect, and Composite Estimators

Apart from the design- and model-based modes of inference introduced above, estimators in general can be described as being either direct, indirect, or composite in nature, depending on the sources of information they make use of. Domain-direct estimates use observed values y_i only from sample units in a particular domain, i.e., $i \in s_d$. Domain-specific totals can be estimated directly as

$$\hat{\tau}_d^{DIR} = \sum_{i \epsilon s_d} w_i y_i \tag{3}$$

where *DIR* indicates that $\hat{\tau}$ is a direct estimator, and d denotes the domain of interest. Substituting $w_i = 1/\pi_i$ in Equation 3 shows how the H-T estimator (Equation 1) serves as the domain-direct estimator when data are restricted to sample units selected in domain d, i.e., when $i \in s_d$. The benefit of direct domain estimation is that no explicit model assumptions are required (cf. Sterba, 2009), and sampling weights can be used, allowing for design-unbiased estimates. A fundamental concern of SAE is that direct estimation often leads to unacceptably large or unstable standard errors for domains with small sample sizes (n_d). Additionally, no direct estimates are possible for domains having no sampled units, i.e., when $n_d = 0$.

Indirect domain estimation seeks to remedy the direct domain estimator's shortcoming of large variance when n_d is small by increasing the "effective sample size" using information outside of the domain of interest together with a statistical model (Rao and Molina, 2015, pg. 35). The indirect approach is manifest in what is often called a synthetic estimator, e.g.,

$$\hat{\tau}_d^{SYN} = \boldsymbol{\tau}'_{xd}\hat{\boldsymbol{\beta}} \tag{4}$$

which gives the indirect or synthetic estimate (Schaible, 1993) of the total for the d^{th} small area domain, with domain d auxiliary total $\boldsymbol{\tau}_{xd}$, and regression coefficients $\hat{\boldsymbol{\beta}}$ estimated from $(\boldsymbol{x}, y)$ data sampled across the entire population to borrow strength for estimating on d. Benefits of the synthetic estimator are that it can allow for estimates to be made in non-sampled units, and likely has a smaller variance than direct domain estimates, especially where n_d is small. However, Equation (4) as given does not account for between-domain heterogeneity and thus can be biased for specific domains. Additionally the indirect estimator does not necessarily trend toward the unknown domain total τ_d as n_d increases.

Not all synthetic estimators are regression models, but the example in Equation (4) was chosen to illustrate some additional details. First, when the coefficients are estimated only from data sampled in the domain of interest, i.e., when $\hat{\boldsymbol{\beta}}$ in Equation (4) is replaced by $\hat{\boldsymbol{\beta}}_d$, the estimator is considered to be a model-assisted (GREG) estimator having design-based properties for inference. The same is true when the regression model is fit using sample observations weighted by their design weights using weighted least squares (Särndal et al., 1992; McConville et al., 2020, Section 6.4). Second, a synthetic estimator's inference base may depend on assumptions about its model form being correct, and whether its parameters estimated from one set of domains are suitable for making predictions for other domains. As with any estimator, care should be taken to verify what conditions must be met to satisfy the inference base for GREG.

The synthetic estimator can be used in concert with the direct domain estimator to create a composite estimator that balances the strengths and weaknesses of direct and indirect estimators (Rao and Molina, 2015):

$$\hat{\tau}_d^{COMP} = \gamma_d \hat{\tau}_d^{DIR} + (1 - \gamma_d)\hat{\tau}_d^{SYN} \tag{5}$$

The weighted average of the direct and indirect estimators comprises the composite estimator $\hat{\tau}_d^{COMP}$, where $\hat{\tau}_d^{DIR}$ is a direct estimator, e.g., (3), $\hat{\tau}_d^{SYN}$ is an indirect estimator, e.g., (4), and $\gamma_d \in [0, 1]$ is a domain-specific weighting factor, also known as a shrinkage factor. An optimal solution for minimum $MSE(\hat{\tau}_d^{COMP})$ can be formulated as $\hat{\gamma}_d = MSE(\hat{\tau}_d^{SYN}) / \left(MSE(\hat{\tau}_d^{SYN}) + MSE(\hat{\tau}_d^{DIR})\right)$ (Rao and Molina, 2015, Section 3.3.1). For the optimal solution, as n_d gets large or when $MSE(\hat{\tau}_d^{DIR})$ is small, $\hat{\gamma}_d$ tends toward 1, moving the composite estimator toward $\hat{\tau}_d^{DIR}$ and its favorable properties of unbiasedness and consistency. Similarly $\hat{\tau}_d^{COMP}$ tends toward the synthetic estimator when n_d is small or $MSE(\hat{\tau}_d^{DIR})$ is large. Most SAE approaches favor the composite estimator for its ability to balance the unbiasedness and precision of direct and indirect estimators, respectively, while allowing for flexibility in the choice and formulation of direct estimators and synthetic models.

3.2. Unit Level SAE

Unit-level SAE employs synthetic models that operate at the scale of observational or sample units in the population, typically field plots in forest inventory applications. The synthetic model in a unit-level estimator is used to predict $\hat{y}$ on all population units in a domain d, regardless of how many (or whether) sample data for y were observed in that domain. Predictors from auxiliary variables $\boldsymbol{x}$ and the model relationship $y \sim \boldsymbol{x}$ provide the means of generating predictions for $\hat{y} \in d$. Here $\boldsymbol{x}$ is either known for every unit in d or known in aggregate, as in a case where a domain-specific mean for $\boldsymbol{x}$, denoted $\bar{\boldsymbol{x}}_d$, is known. In either case, paired values of $(\boldsymbol{x}, y)$ are observed on sampled units across the broader population, i.e., the indirect data, and used to fit or train a synthetic model. Errors in the synthetic model predictions are partitioned into two components (see Rao and Molina, 2015, model 4.3.1). A domain-specific error v_d is attributed to synthetic model variance that applies equally to all $y \in d$, and within-domain residual errors e_{di}—independent of v_d—that apply to individual sample units y_i; $i \in d$.

The unit-level composite estimator developed by Battese et al. (1988, hereafter BHF) is an excellent example from which to

study unit-level SAE, as the data and models the authors used to demonstrate the application are integrated into the R "sae" package (Molina and Marhuenda, 2015). BHF used the nested-error linear regression model

$$y_{di} = \boldsymbol{x}'_{di}\boldsymbol{\beta} + v_d + e_{di};\ d = 1, \ldots, D;\ i = 1, \ldots, n_d.$$

to estimate crop areas for corn and soybeans in Iowa, using USDA Statistical Reporting Service field survey data from 1978 as direct observations of y, and Landsat 2 multispectral scanner (57 x 79 m) raster imagery as auxiliary information for $\boldsymbol{x}$. Unit-level approaches rely on matching individual sample unit observations to the auxiliary data; thus, the satellite image pixels in BHF were assigned to corresponding 250-ha (1 sq. mile) field survey units (BHF called the units "segments")—about 555 raster cells per segment. The aggregated Landsat and field survey data formed $(\boldsymbol{x}, y)$ pairs for 37 sampled segments in a population of $D = 12$ counties, with each county treated as a small-area domain of interest. The BHF synthetic model used a response y = sample segment area (ha) growing corn, in a linear regression model with an intercept and two predictors x_1 = Landsat pixels classified as corn and x_2 = pixels classified as soybeans, with $\sum_{d \in D} n_d = 37$ observed segments.

Using their synthetic model, similar to Equation (4), BHF were able to predict domain means for corn area per segment in counties as

$$\hat{\mu}_d = \bar{\boldsymbol{x}}'_d\hat{\boldsymbol{\beta}} + \hat{v}_d \tag{6}$$

with $\bar{\boldsymbol{x}}_d$ calculated from all segments in county d, using Landsat pixel class counts of a given crop in each segment (see Table 1, Battese et al., 1988). By estimating $\hat{\boldsymbol{\beta}}$ as fixed effects and $\hat{v}_d$ as random effects using mixed linear regression modeling, BHF obtained the empirical best linear unbiased predictor (EBLUP) of domain-specific means $\hat{\mu}_d^{EBLUP} = \bar{\boldsymbol{x}}'_d\hat{\boldsymbol{\beta}} + \hat{v}_d$, a key contribution of their work because of the favorable properties of the EBLUP.

Goerndt et al. (2013, Equation 5) and Costa et al. (2009, Equation 13) presented the BHF unit-level nested error EBLUP in the form of a composite estimator with domain-level, i.e., county-level estimates obtained from direct and synthetic regression estimates weighted as

$$\hat{\mu}_d^{EBLUP} = \left(\frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \frac{\hat{\sigma}_e^2}{n_d}}\right)\hat{\mu}_d^{DIR} + \left(1 - \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \frac{\hat{\sigma}_e^2}{n_d}}\right)\bar{\boldsymbol{x}}'_d\hat{\boldsymbol{\beta}} \tag{7}$$

where $\hat{\mu}_d^{DIR}$ is a sample-direct estimate of the mean y for small area domain d and $\hat{\boldsymbol{\beta}}$ is the vector of fixed effects regression coefficients obtained by linear mixed modeling with domain-specific random effects. The term σ_v^2 in Equation (7) denotes the variance among domain random effects $v_d \sim \mathcal{N}(0, \sigma_v^2)$, and σ_e^2 as the residual variance of population units—the variance unaccounted for by the EBLUP—within domains, i.e., $e_{di} \sim \mathcal{N}(0, \sigma_e^2)$. This assumption assigns a single residual variance to all domains, with σ_e^2 estimated from the full set of sample residuals. As when using optimal shrinkage weights $\hat{\gamma}_d$ in Equation (5), the weights in Equation (7) ensure that as the domain-specific direct estimator variance gets small, such as when n_d is large, the EBLUP tends toward $\hat{\mu}_d^{DIR}$.

Unit-level paired $(\boldsymbol{x}, y)$ data typically provide an information-rich means of linking indirect data from broad and extensive populations to specific domains of interest. Further, when y is observed by non-random sampling, the model-based properties of the composite estimator support approximate and asymptotic inference bases where direct estimation alone would not. Efforts should be made to verify the veracity of the underlying statistical model and stochastic processes, especially when data collection employs stratification, clustering, or disproportionate sampling among some elements of the population (Sterba, 2009). In applications involving large data sets the computational requirements of unit-level analyses can be demanding, especially when objectives include the validation of variance estimates approximated by Taylor series linearization or when using resampling procedures to estimate variance components (Prasad and Rao, 1990; González-Manteiga and Morales, 2008). Computationally demanding synthetic models can also pose challenges for SAE, but software advances have made steady gains in providing tools to address such challenges (McRoberts et al., 2007). As with any model-based inference approach, customary steps involving model selection, goodness-of-fit, and checking other assumptions are important in unit-level SAE. In cases where influential points, heteroscedastic error variances, or non-independence of residuals present problems, developments have been made to help overcome limitations of the BHF approach (Babcock et al., 2015; Breidenbach et al., 2018).

3.3. Area Level SAE

Unlike unit-level approaches, area-level SAE employs synthetic models that operate at the scale of subpopulation domains, rather than individual sample or observational units. As a consequence, auxiliary data do not need to be paired one-to-one with individual sample observations. Instead, domain-direct estimates (e.g., $\hat{\tau}_d^{DIR}$ or $\hat{\mu}_d^{DIR}$) are paired with domain-specific observations, such as domain means or totals of $\boldsymbol{x} \in d$, which we denote as $\boldsymbol{x}_d$. The paired domain data $(\hat{y}_d^{DIR}, \boldsymbol{x}_d)$ are then used to develop regression models or train other types of synthetic models for use in SAE. Rao and Molina (2015, model 4.2.5) presented a regression-based area-level model

$$\hat{y}_d^{DIR} = \boldsymbol{x}'_d\boldsymbol{\beta} + b_d\psi_d + \epsilon_d \tag{8}$$

where ψ_d are domain-specific model errors and ϵ_d are errors due to sampling on the domain-direct estimates that appear on the left-hand side in Equation (8). In the basic area-level model, Rao and Molina (2015) explain that b_d are domain specific positive constants set to $b_d = 1$ in the model presented by Fay and Herriot (1979), which, for direct estimates of domain totals can be expressed as

$$\hat{\tau}_d^{FH} = \boldsymbol{x}'_d\boldsymbol{\beta} + \psi_d + \epsilon_d \tag{9}$$

Fay-Herriot (F-H) models (Fay and Herriot, 1979) like the one shown in Equation (9) are often synonymous with area-level SAE, as they have seen considerable use in area-level applications

since their introduction. An EBLUP derived from the F-H model, expressed as a composite estimator having a form similar to Equation (5) shows the roles of direct and synthetic estimators

$$\hat{\tau}_d^{FH} = \left(\hat{\gamma}_d\right) \hat{\tau}_d^{DIR} + \left(1 - \hat{\gamma}_d\right) \boldsymbol{x}_d{}' \hat{\boldsymbol{\beta}} \tag{10}$$

with weights $\hat{\gamma}_d = \hat{\sigma}^2_{\psi_d} / (\hat{\sigma}^2_{\psi_d} + \sigma^2_{\epsilon_d})$ composed as the model variance relative to the total variance. A difference in the shrinkage weights' formulations for Equations (7) and (10) is that $\sigma^2_{\epsilon_d}$ in Equation (10) is assumed to be known, but typically specified as the variance estimated from sample observations $y \in d$. Wang and Fuller (2003) developed a modified area-level estimator that accounts for the empirical estimation of domain-direct variances, which has been demonstrated in forest inventory applications (Magnussen et al., 2017). Similar to the unit-level model, $\hat{\boldsymbol{\beta}}$ in Equation (10) is a vector of fixed-effects coefficients from a linear mixed model having domain-specific random effects. As in Equation (7) the F-H EBLUP tends toward the direct estimator when the direct variance is low, and toward the synthetic estimator when the direct variance is high. Because the synthetic estimator in the F-H model operates on subpopulation domains, the same aggregated measures $\boldsymbol{x}_d$ from the auxiliary data that were used in estimating the regression coefficients are also used in predicting $\hat{\tau}_d^{FH}$ in Equation (10). In contrast, the unit-level EBLUP uses an aggregated measure such as $\bar{\boldsymbol{x}}_d$ in Equation (7), despite the model coefficients having been estimated using observed $\boldsymbol{x}$ (and y) values from individual sample units.

In applications framed in a geographic context, such as forest inventories, where field plot observations are often paired to auxiliary data from remote sensing, area level modeling obviates the need for highly accurate plot coordinates, and can be used when there is a degree of misalignment between plots and auxiliary data (Goerndt et al., 2011). This concern is highly relevant when the protection of confidential information prevents the release of exact coordinates of sampled locations to unauthorized personnel. Instead, only direct linkage to a specific domain is needed for each plot. This also facilitates using sample units that possess indistinct sampling boundaries, such as where linear transects or variable-radius plots are used in field sampling (Ver Planck et al., 2018). This method also tends to require less processing time which lends itself to analysis involving very large datasets.

3.4. Other SAE Methods

In addition to the aforementioned model forms, there are other methods which serve as alterations or variations of the above model types and as such are not mutually exclusive from them. Such methods include Bayes methods, and nearest neighbor models which are worthy of particular mention due to their use in forestry SAE (McRoberts, 2012; Babcock et al., 2018; Ver Planck et al., 2018). Bayesian methods, both empirical Bayes (EB) and hierarchical Bayes (HB) offer some advantages over their frequentest counterparts such as being able to model different target variable types such as binary or count data (He and Sun, 2000). HB also gives posterior distributions of the small area parameters, and can therefore avoid relying on unrealistic asymptotic assumptions (Pfeffermann, 2013). Bayesian methods offer flexibility in specifying spatial structures such as spatially correlated random effects (Wang et al., 2018).

Nearest neighbor techniques offer a similar set of benefits for SAE, for example, where estimators for categorical, binary, or count variables are involved. They are non-parametric in that no distributional assumptions regarding response or predictor variables are necessary, which has proven useful when multivariate auxiliary information is used to construct synthetic models. Nearest neighbor models can also accommodate correlated sample and auxiliary data that may arise in spatial or temporal domains (McRoberts et al., 2007; McRoberts, 2012).

4. APPLICATIONS IN FOREST INVENTORY

We now turn to selected examples related to SAE in forest inventory from published research (**Table 1**). Note that in **Table 1**, we exclude a large body of literature employing post-stratification, a widely-used model-assisted estimation technique. Instead we aim to focus on estimators less-commonly used in existing forest inventory production processes. Post-stratification notwithstanding, model-assisted estimators including GREG were among the most widely-used and earliest-adopted methods for increasing estimator precision using models to link auxiliary information from remote-sensing with field sample data. As such, we elected to include a number of model-assisted applications in our example references, even though some of the selected examples do not involve "small area" domains as the term is often used in SAE (**Table 1**).

Among the reasons authors have given for adopting model-assisted estimators was the need to ensure design-based inference in large, multi-resource sample designs (Reich and Aguirre-Bravo, 2009; Næsset et al., 2011). Others cited the need for precise estimation in small area domains nested within broader population surveys as a motivating factor (Goerndt et al., 2011; McRoberts, 2012; Magnussen et al., 2014). A few noted that consistency and additivity of estimates from small areas nested within larger domains were motivating concerns (Reich and Aguirre-Bravo, 2009; Nagle et al., 2019). Others noted the need for sample survey organizations to conduct generic inference, i.e., to make compatible estimates of all forest attributes simultaneously by using the same model to define survey weights (Opsomer et al., 2007; Johnson et al., 2008; McConville et al., 2020). Nearly all research referenced in **Table 1** reported the potential for increased efficiency in estimating forest inventory attributes as a reason for pursuing the work, with the high cost of increasing field sample sizes frequently noted as an operational constraint.

Lidar and digital aerial photogrammetry (DAP) were major data sources used as auxiliary information, with some authors using lidar or DAP point cloud metrics, e.g., height percentiles or pulse return densities aggregated to unit or area levels, and others using canopy height models (CHM) processed from point cloud data (Steinmann et al., 2013; Babcock et al., 2015). Field plots in forest inventories were the primary source of directly-sampled observations, with more than half of the selected studies using NFI or other land-management agency field sample observations,

TABLE 1 | Table of selected studies using SAE-related methods in forest inventory with methods and data used.

References	Estimation method(s)	Direct data	Auxiliary data
Affleck and Gregoire (2015)	GREG	Tree crowns	Tree Branch Attributes
Anderson and Breidenbach (2007)	GREG	FI	lidar
Babcock et al. (2015)	HB	FI	lidar
Babcock et al. (2018)	HB	FI	lidar, Landsat
Breidenbach and Astrup (2012)	GREG, U-EBLUP	NFI	DAP
Breidenbach et al. (2018)	U-EBLUP, A-EBLUP	NFI	DAP
Coulston et al. (2021)	A-EBLUP	NFI	Landsat, TPO
Frank et al. (2020)	U-EBLUP	FI	lidar
Goerndt et al. (2011)	GREG, A-EBLUP, NN	FI	lidar
Goerndt et al. (2013)	GREG, U-EBLUP, NN	NFI	Landsat, NLCD
Goerndt et al. (2019)	A-EBLUP, NN	NFI	Landsat, NLCD, MODIS, EDNA
Green et al. (2020)	A-EBLUP, U-EBLUP	FI	lidar, management records
Magnussen et al. (2014)	GREG, U-EBLUP	NFI	lidar, DAP
Magnussen et al. (2017)	A-EBLUP, HB	FI	lidar, DAP
Mauro et al. (2017)	A-EBLUP, U-EBLUP	FI	lidar
Mauro et al. (2019)	U-EBLUP	FI	lidar, SRM
McConville et al. (2020)	GREG, Other	NFI	Landsat
McRoberts (2012)	NN	NFI	Landsat
McRoberts et al. (2007)	NN	NFI	Landsat
McRoberts et al. (2013)	GREG, A-EBLUP	NFI	lidar
Næsset et al. (2011)	GREG	FI	lidar, InSAR
Næsset et al. (2013)	GREG	NFI	lidar
Nagle et al. (2019)	GREG	NFI	NLCD
Pascual et al. (2018)	U-EBLUP	FI	lidar
Reich and Aguirre-Bravo (2009)	GREG	FI	Landsat
Steinmann et al. (2013)	GREG	NFI	lidar, DAP
Ver Planck et al. (2018)	A-EBLUP, HB	FI	lidar

Methods included are generalized regression estimators (GREG), nearest neighbor (NN), unit-level empirical best linear unbiased predictor (U-EBLUP), area-level best linear unbiased predictor (A-EBLUP), and hierarchical Bayes (HB). Direct data sources include national forest inventories (NFI), non-national based forest inventories (FI), plus directly sampled tree data. Auxiliary data used include tree branch data, light detection and ranging (lidar), digital aerial photogrammetry (DAP), Landsat, timber products outputs (TPO) surveys, national land cover database (NLCD), management records, solar radiation models (SRM), elevation derivatives for national applications (EDNA), moderate resolution imaging spectroradiometer (MODIS), and interferometric synthetic aperture radar (InSAR).

and most others using plots installed for management or research purposes, such as on state, university, private, or public experimental and working forests (Anderson and Breidenbach, 2007; Mauro et al., 2017; Green et al., 2020).

Several studies aimed to examine and augment existing estimation frameworks in forest inventory settings to address potential violations in underlying assumptions. Examples include accounting for heteroscedasticity of variance in synthetic model residuals and spatial or temporal autocorrelation among measurements on observational units or areas of interest (Babcock et al., 2018; Breidenbach et al., 2018; Ver Planck et al., 2018; Mauro et al., 2019). Estimating the change of forest attributes over time with SAE methods was demonstrated successfully by Mauro et al. (2019) and by Coulston et al. (2021), both of which used repeated measurements from forest field plots in their work.

Multiple studies applying one or both unit-level and area-level EBLUP-based SAE appear in **Table 1**. The choice of adopting unit- or area-level SAE in forest inventory applications can depend on limitations of sample or auxiliary data sets, for instance, where georeferencing inaccuracies introduce significant errors in pairing field observations to remotely-sensed or other geospatial data sets, e.g., maps (Næsset et al., 2011; Green et al., 2020). Similar challenges in pairing field data to remote-sensing can occur where plot designs and raster layers are incompatible, such as when variable radius field plots (i.e., angle gauge sampling) are used, or when field plot sizes are small compared to the pixel size in available auxiliary data sources (Goerndt et al., 2011; Ver Planck et al., 2018; Temesgen et al., 2021). Although area-level estimators are flexible to accommodate situations where precise georeferencing is impractical, a trade-off may arise due to the loss of information from aggregating unit-level observations to domain- or area-level scales; further, aggregating sample observations also reduces effective numbers of observations available for synthetic model development (Magnussen et al., 2017).

When georeferencing enables pairing of auxiliary and sample data, sampled units can all serve as $(\boldsymbol{x}, y)$ observations for fitting or training synthetic models. The models can then be used to make predictions on raster cells or other spatial units that correspond to unobserved population units (Babcock et al., 2018; Frank et al., 2020). It follows that unit-level SAE estimators can be used to map predictions or estimates at finer spatial resolutions than area-level methods, the latter of which allow for mapping SAE predictions only at domain levels. Another potential advantage of the unit-level approach is the ability to define arbitrary spatial domains subsequent to sampling without necessarily losing substantial power of inference (Pascual et al., 2018). Gains in efficiency between unit-level and area-level SAE methods have been compared in multiple studies, generally confirming the potential for greater gains from unit-level approaches (Mauro et al., 2017; Breidenbach et al., 2018).

4.1. Gains Over Direct Estimation

4.1.1. Model-Assisted Estimation

Model-assisted estimators have proven to reduce estimator errors considerably, e.g., when measured by the relative efficiency of a model-assisted estimator compared to its simplest direct counterpart—often characterized as simple random sampling—cf. Equations (2) and (3)

$$RE = \frac{Var(\hat{\tau}_y^{MA})}{Var(\hat{\tau}_y^{DIR})} \tag{11}$$

McRoberts et al. (2013) demonstrated substantial variance reduction $100\%(1 - RE) = 84\%$ using a non-linear regression-based model-assisted estimator of growing stock volume per unit area in a 1,300 km^2 study area in southeastern Norway. The gains corresponded to over sixfold increase in apparent sample size, calculated as RE^{-1}. Aerial lidar (0.7 pulses m^{-2}) served as the auxiliary data, with direct observations from n = 145 NFI-type (200 m^2) field plots over the study area. They noted similar variance reduction (82%) using the same method to estimate growing stock volume over a one-half partition of the study area represented by n = 69 field plots, thus demonstrating how sample data from a larger area can be used to increase the precision of estimates in a smaller area (McRoberts et al., 2013).

Model-assisted estimation has been demonstrated to increasing precision in multiple population sub-domains including Breidenbach and Astrup (2012), who tested GREG as a model-assisted estimator in a similar sized study area in Norway divided into 14 municipalities, each having from 1 to 35 NFI sample plots of a total n = 145. They noted gains in precision were smaller and more variable than McRoberts et al. (2013), with RE ranging from 0.35 to 0.87 in eight municipalities having $n_d \geq 6$. They concluded that sample sizes $n_d < 6$ in the other six municipalities gave unstable estimates, a finding similar to Næsset et al. (2011). Their data revealed that despite GREG's limitation in areas with small n_d, its estimates deviated from sample-direct design-based estimates by less than half as much as synthetic model estimates alone, a result consistent with the design-unbiased property of GREG in the model-assisted framework. Næsset et al. (2013) observed consistent improvement in precision of estimates of aboveground forest biomass using GREG with aerial lidar auxiliary data in model-assisted two-stage sampling. Their gains were greatest ($RE = 0.125$) for all cover classes combined ($n = 632$), and only slightly more modest $0.09 \leq RE \leq 0.20$ for individual age and productivity classes, all of which had class-specific sample sizes $n_c \geq 46$.

Reduced RMSE of synthetic model estimates when compared to direct estimator variance has been demonstrated in a number of applications involving the pairing of ALS and NFI-type field sample data (Næsset et al., 2011; Järnstedt et al., 2012; Nord-Larsen and Schumacher, 2012; Kotivuori et al., 2016; Nilsson et al., 2017; Novo-Fernandez et al., 2019). A variety of synthetic modeling approaches have been tested including parametric and non-parametric regression modeling, Random Forests and other ensemble predictive models, and nearest-neighbor imputation, often with a goal of identifying suitable auxiliary data sources for estimating forest biophysical attributes (Latifi et al., 2010; Popescu et al., 2011; Bright et al., 2012; Rahlf et al., 2014). A common theme in these studies is the direct examination of synthetic model prediction errors (e.g., using cross validation) without formulating the models in a design-based or composite modeling framework to mitigate potential synthetic estimator bias (cf. McRoberts et al., 2013; Irulappa-Pillai-Vijayakumar et al., 2019; McConville et al., 2020). In model-assisted applications the usual goal is to increase the precision of population-level estimates, and less often to produce estimates for domains that divide a larger population into small areas where direct estimator instability can be a concern. Where investigated, model-assisted estimators were able to reduce small area uncertainties considerably, within limits of direct-data sampling intensity and the strength of model relationships involving indirect and auxiliary data (Breidenbach and Astrup, 2012).

4.1.2. Unit-Level SAE

Research comparing unit-level SAE to model-assisted estimators has shown that gains in precision are generally greater for unit-level EBLUPS than model-assisted estimates (e.g., GREG) primarily when direct-domain sample sizes are small. In directly comparing unit-level EBLUPs to model-assisted GREG estimates, Breidenbach and Astrup (2012) noted an average $\overline{RE} = 0.86$, meaning unit-level SAE MSEs were lower than direct estimate variances by an additional 14%, on average, compared to GREG. Not all EBLUP MSEs were smaller than those computed for GREG. Comparisons showed greatest gains in five of six municipalities having $6 \leq n_d \leq 17$, but smaller—even negative—gains ($\overline{RE} = 0.93$ and 1.15) were noted in two municipalities having $n_d \geq 29$. Such findings indicate that unit-level EBLUPS may not have as clear of an advantage over GREG in reducing estimator variance when domain sample sizes are relatively large; nonetheless, even when n_d was relatively large, EBLUP performance was not much worse than model-assisted regression estimates (Breidenbach and Astrup, 2012). EBLUPS showed considerable stability across the range from $1 \leq n_d \leq 35$ compared to GREG, with unit-level SAE relative errors ranging from just [7.0, 12.4] % compared to a range of

[0.6, 25.4] % for GREG, even with three municipalities having $n_d = 1$ excluded for the GREG results since their errors could not be calculated (Breidenbach and Astrup, 2012). Magnussen et al. (2014) reported virtually identical gains for model-assisted and unit-level EBLUP estimates (both $\overline{RE} = 0.49$) when compared to direct volume per hectare estimates for forest districts in Switzerland. The near identical results may have been related to comparatively large average sample sizes $\bar{n}_d = 79$ across the $D = 108$ Swiss forest districts.

Of the studies listed in **Table 1** that compared unit-level EBLUPs to domain-direct estimates, the largest variance reductions noted ($\overline{RE} = 0.03$) were in a study of forest volume in Burgos Province, Spain, made up of $D = 54$ stands in an area covering about 13.65 km^2 (Pascual et al., 2018). The same authors reported more modest gains ($\overline{RE} = 0.50$) in a 3 km^2 management area having $D = 6$ stands near Cercedilla, Spain. The authors attributed the greater gains at Burgos as being due to the lower sampling intensity there (about 0.5 %) compared to a 4 % sampling intensity at the Cercedilla site (Pascual et al., 2018). Large gains were reported for unit-level volume EBLUPs ($\overline{RE} = 0.09$) tested by Mauro et al. (2017) in an area roughly 8 km^2. A high degree of positive skewness was evident in the distribution of n_d across domains, with roughly 30% of map unit domains having $n_d \leq 2$ despite $\bar{n}_d = 10.3$ across $D = 64$ map units having any sample observations (Mauro et al., 2017). In estimating volume change in $D = 24$ stands experimentally manipulated for three levels of forest structural diversity, Mauro et al. (2019) reported unit-level gains ($\overline{RE} = 0.71$) for 7-year volume change estimates. The sample design relied on 1 remeasured plot per 8 ha on a systematic grid, for a sparse but narrow range of domain-direct sample sizes, with $3 \leq n_d \leq 10$ and $\bar{n}_d = 6.3$ (Mauro et al., 2019). Breidenbach et al. (2018) reported greater efficiencies $\overline{RE} = 0.43$ and $\overline{RE} = 0.28$ compared to direct estimate variances, with the latter result including a formulation that accounted for heteroskedasticity of variance in synthetic model residuals.

By artificially reducing sample sizes from the available $n = 680$ NFI plots across $D = 12$ county-sized domains for estimating Oregon Coast Range forest volumes, Goerndt et al. (2013) demonstrated diminishing efficiency gains with increasing n_d in unit-level EBLUPS from $\overline{RE} = 0.41$ ($n = 136$), to $\overline{RE} = 0.57$ ($n = 204$), to $\overline{RE} = 0.71$ ($n = 272$). They also found that alternative unit-level composite estimators calculated with smoothed variances performed well in terms of increased precision and low apparent biases in unit-level SAE (Costa et al., 2003; Goerndt et al., 2013). The alternative estimators employed multiple linear regression, as well as nearest neighbor and gradient-nearest-neighbor imputation in synthetic models to achieve balances between bias and precision of SAE (Ohmann and Gregory, 2002).

4.1.3. Area-Level SAE

A number of studies have shown area-level SAE precision gains for timber volume or biomass compared to sample-direct estimates. Breidenbach et al. (2018), for example, reported a reduction in overall standard error from 32.7 $m^3 ha^{-1}$ for direct volume estimates compared to F-H area-level RMSE of 23.1 $m^3 ha^{-1}$ ($\overline{RE}$ = 0.50). A nearly identical gain in efficiency ($\overline{RE}$ = 0.50) was reported by Goerndt et al. (2011) for volume estimates in forest stands treated as small area domains with intentionally-reduced sample sizes between $2 \leq n_d \leq 4$. Relative gains tended to be less when the simulated small n_d were increased by factors of two and three, with no gains for larger sample size increases. Despite finding modest gains when n_d was large, area-level EBLUPS were demonstrably superior—in terms of RE and lack of apparent biases—to two synthetic estimators and two composite James-Stein type estimators tested (James and Stein, 1961; Goerndt et al., 2011). In testing area-level SAE with counties as small-area domains, a composite estimator similar to F-H showed $0.43 \leq \overline{RE} \leq 0.91$ over a 20-state region of the northeastern U.S. (Goerndt et al., 2019). In the same study a composite estimator based on a non-parametric nearest-neighbor (NN) synthetic model showed slightly less gain in efficiency than the F-H type approach. Despite this, Goerndt et al. (2019) indicated the NN approach may have lower model bias and they cautioned against potential biases in model-based estimators, pointing out the need for thorough checking of model assumptions to ensure validity of model-based inferences.

The pattern of decreasing gains in efficiency with increasing n_d was also noted by Mauro et al. (2017), who reported an average $\overline{RE}$ = 0.48 over $D = 84$ management areas (domains), while no gain ($\overline{RE}$ = 1.13) was noted in 14 of the domains having $n_d \geq 25$ sample plots (cf. Goerndt et al., 2011, 2019). Green et al. (2020) noted average $\overline{RE}$ = 0.79 in F-H estimates of timber volume across $D = 40$ stands, with little or no efficiency gains in stands where direct estimates were already quite precise (relative standard errors $<$ 10%). Greater gains ($\overline{RE} \approx 0.35$) were noted in stands having direct relative standard errors $>$ 25%. Findings such as these indicate that where domain-direct estimates are already quite precise, as in cases where n_d is large or variation within domains is inherently low, F-H type estimators may exhibit a limited ability to further increase precision over domain-direct estimates.

Magnussen et al. (2017) reported $\overline{RE}$ ranging from 0.44 to 0.77 in four study areas in Spain, Norway, Switzerland, and Germany, using a modification of F-H that treats domain-specific variances as estimates rather than known constants (Wang and Fuller, 2003). In the same study they found greater gains $0.28 \leq \overline{RE} \leq 0.34$ by including a non-stationary spatial correlation process in an area-level composite estimator which accounted for the spatial covariance structure in model residuals (Chandra et al., 2012, 2015). In a third approach Magnussen et al. (2017) used the HB approach of Datta and Mandal (2015) to obtain efficiency gains intermediate ($0.27 \leq \overline{RE} \leq 0.81$) compared to their baseline—specifying empirically estimated variances—and non-stationary spatial F-H approaches. The Bayesian approach demonstrated several advantages related to estimated posterior distributions for specific domains, especially when the random-effect variance was largely attributable to a small number of domains from the larger population (Magnussen et al., 2017). Coulston et al. (2021) evaluated the performance of area-level F-H models including a simultaneous autoregressive (SAR) model of residual spatial correlation among domains in estimating forest removals, noting the spatial model improved efficiency of estimates at scales of individual counties, but not at the scale of larger survey regions

encompassing groups of 12-20 counties each in the southeastearn United States. Ver Planck et al. (2018) compared F-H gains with no accounting for spatial autocorrelation ($\overline{RE} = 0.23$) to a conditionally-autoregressive (CAR) F-H model that further reduced estimator variance ($\overline{RE} = 0.19$), noting that the CAR model performance was generally greater in domains (stands) sharing boundaries with high numbers (> 10) of neighboring stands. The non-stationary spatial process (Chandra et al., 2012) may improve upon simpler CAR and SAR modeling approaches, as it employs a distance-weighted measure of correlation among domains rather than a simple binary model based on domain adjacency. Spatial models provide a promising tool for future applications of area-level SAE that account for non-trivial spatial correlations among small area domains likely to be realistic in many forest inventory applications (Finley et al., 2011; Magnussen et al., 2017).

4.1.4. Unit Level vs. Area Level SAE

Although area-level EBLUPS have shown clear gains in efficiency over direct estimates in forest inventory SAE applications, still greater gains have been demonstrated using unit-level approaches when suitable data are available and model assumptions are met (Breidenbach et al., 2016, 2018). In comparing both approaches to direct estimates of forest volume, Mauro et al. (2017) observed halving of variance ($\overline{RE}$ = 0.48) and ten-fold reduction ($\overline{RE}$ = 0.09) for area-level and unit-level estimates, respectively. A notable feature in the study was the large proportion—slightly more than half—of the 84 stand groupings defined as small area domains containing $n_d \leq 6$ field plots, linking greatest gains in efficiency to domains having relatively small n_d. Their unit-level results achieved variance reductions among the largest of any reported in forest inventory literature (Mauro et al., 2017; Pascual et al., 2018). By comparison, Breidenbach et al. (2018) reported $\overline{RE} = 0.50$ (see Section 4.1.3) for area-level and $\overline{RE} = 0.28$ (see Section 4.1.2) for unit-level estimators applied to a common data set. Their domain-direct sample sizes were also small [$4 \leq n_d \leq 7$], so the source of differential gains between the two studies' unit-level estimators may be related to other factors including the strength of the synthetic model relationships, which we weren't able to compare between the two studies (Mauro et al., 2017; Breidenbach et al., 2018). A distinct barrier to achieving large gains in efficiency with unit-level SAE, especially compared to area-level approaches, is the ability to accurately georeference field plots and spatial auxiliary data sources. Green et al. (2020) noted this as a possible reason for the lack of gains in their unit-level models compared to area-level SAE.

4.2. Variance Estimation

Variances for model coefficients and estimates of finite-population parameters for design-based direct or model-assisted estimates (Sections 2.1 and 2.2) are in most cases calculable using commercially available statistical software packages (Molina and Marhuenda, 2015; Breidenbach, 2018; McConville et al., 2018; Hill et al., 2021). Other variance estimators are documented in research literature in sufficient detail to facilitate calculation with scientific programming software (McRoberts, 2012; Mandallaz et al., 2013; Babcock et al., 2015; Magnussen et al., 2017; Mauro et al., 2017; Frank et al., 2020). In design-based model-assisted estimators, variance calculations exist in closed form for some estimators such as GREG, or as approximations for others including ratio estimators (Särndal et al., 1992; Breidenbach and Astrup, 2012; Mandallaz, 2013; Magnussen et al., 2018). Because variance calculations often require accounting for non-independence of observations or heteroskedasticity of residuals, or where algorithmic synthetic models such as non-parametric or nearest-neighbor modeling are employed, iterative methods can be used to estimate approximate variances, even where closed-form solutions exist when model assumptions allow for them (McRoberts et al., 2007). Numerical approaches, such as leave-one-out cross validation or parametric bootstrap estimation have proven useful in variance estimation, although care must be taken to ensure that bootstrap data-generating mechanisms are aligned with error correlation structures and distributional assumptions.

Standard errors of domain-direct estimates are required inputs for FH and other area-based SAE, which can pose a problem when n_d is insufficient to give stable variance estimates (Särndal, 1984; Breidenbach and Astrup, 2012). A solution for such applications is the use of generalized variance estimators (Valliant, 1987; Wolter, 2007; Goerndt et al., 2013; Coulston et al., 2021). Generalized variance functions tend to give variance estimates that are highly dependent on n_d, e.g., Coulston et al. (2021), which may differ from direct variance estimates in domains having sufficient sampling intensity to produce stable standard errors.

Closed form solutions typically do not exist for variance estimation for SAE composite estimators, e.g., when domain-level estimates are obtained as EBLUPs (Fay and Herriot, 1979; Battese et al., 1988; Rao and Molina, 2015). Variance of EBLUPs is assessed by mean squared errors (MSE) rather than standard errors to distinguish EBLUPS from design-unbiased estimators. The MSEs are routinely calculated in SAE software as additive combinations of terms representing uncertainties associated with (a) prediction of random effects, (b) estimation of regression model coefficients, and (c) estimation of the random-effect variance, i.e., the variance among small-area domains (Prasad and Rao, 1990). Alternatives involving parametric bootstrap variance estimation are employed in some applications, as are variances determined from posterior distributions in Bayesian analyses (Prasad and Rao, 1990; Babcock et al., 2015; Molina and Marhuenda, 2015).

4.3. Emerging Applications

Bayesian methods have been applied to SAE across disciplines for several decades (Morris, 1983; Ghosh and Rao, 1994), with recent applications in forest inventory settings as well (e.g., Finley et al., 2011). Babcock et al. (2018) used HB to estimate aboveground biomass with coupled auxiliary data from Landsat and lidar, to resolve incomplete coverage of remote sensing data. Of the HB models they tested, one incorporating spatial random effects with lidar as auxiliary data led to $0.33 \leq \overline{RE} \leq 0.51$ across their 4 areas of interest. By incorporating coregionalization and adding tree cover derived from Landsat to complement incomplete lidar coverage, the range of $\overline{RE}$ decreased to $0.16 \leq$

$\overline{RE} \leq 0.35$. Ver Planck et al. (2018) also saw success in formulating the F-H approach in a HB framework to increase precision of aboveground biomass SAE over direct estimates ($\overline{RE} = 0.23$). Further improvement was demonstrated by adding conditional autoregressive random effects to account for spatial correlation ($\overline{RE} = 0.80$), with the autoregressive model giving greater gains in precision in domains that shared boundaries with larger numbers of neighbors (Ver Planck et al., 2018). Adding a spatial random effect did not always lead to greater predictive performance, however, as shown by Babcock et al. (2015), who attributed modest to negative gains ($0.83 \leq \overline{RE} \leq 1.27$) to possible overfitting.

Estimating population parameters for multiple attributes of interest is an important concern in forest inventory applications ranging from local-scale stand assessments to regional and national multi-resource inventories (Babcock et al., 2013; Lochhead et al., 2018). Multivariate F-H and spatial F-H approaches have been developed and incorporated into statistical software, but their application to forest inventory has been limited (Molina and Marhuenda, 2015; Benavent and Morales, 2016). While multiple domain-specific estimates can be obtained using SAE in separate modeling procedures for each attribute, doing so fails to maintain logical consistencies among estimates, and overlooks potential gains in efficiency that may otherwise be realized by accounting for cross-attribute correlations (Mauro et al., 2017; Coulston et al., 2021). Generic inference in model-assisted design-based estimation affords consistency in multivariate estimates, but may come at a cost of increased standard errors in attributes uncorrelated or weakly correlated with selected auxiliary variables (McConville et al., 2020). Nearest-neighbor approaches have been used successfully in multivariate forest inventory settings, including multivariate model-assisted estimation to improve estimator efficiencies while preserving consistency among estimates (Chirici et al., 2016; McRoberts et al., 2017). Hierarchical Bayesian multivariate methods for SAE have been demonstrated for both unit-level and area-level settings suitable for forest inventory applications (Datta et al., 1998; Arima et al., 2017). Recent advances have also increased computational efficiencies for Bayesian analyses that can be applied to multivariate SAE involving very large data sets, expanding opportunities for further advances in this area (Finley et al., 2015, 2017; Datta et al., 2016; Babcock et al., 2018).

Another development in SAE applications for forest inventory aims at modeling sample and observational units at a level approaching that of individual trees rather than forest plots or larger subpopulation domains (Næsset, 2002; Mauro et al., 2016). Frank et al. (2020) tested a semi-individual tree (s-ITC) model approach—analogous to a unit-level approach—by segmenting tree crowns in lidar point clouds and delineating s-ITC units around them. Compared to a unit-level approach the s-ITC model showed a $\overline{RE} = 1.04$ for volume, $\overline{RE} = 0.71$ for basal area, $\overline{RE} = 1.38$ for stem density, and $\overline{RE} = 0.48$ for quadratic mean diameter. Despite attaining similar precision as plot-based unit-level EBLUPs for volume estimates at the population level, the s-ITC approach showed potential for its increased spatial resolution and ability to estimate population parameters more closely related to individual trees, such as mean diameter (Frank et al., 2020).

Applications of enhanced estimators such as model-assisted and SAE are not limited to cases where estimates per unit of land area are needed. Affleck and Gregoire (2015) compared estimation of crown biomass using randomized branch sampling to provide sample data for GREG. In their examination of a univariate estimator, model-assisted estimation led to improvement in the precision of estimates across a range of simulated sample sizes in randomized branch sampling ($n = 5$, 10, 20 and $\overline{RE} = 0.54$, 0.71, and 0.97, respectively), although the authors cautioned against the possible trade-off between precision and design-unbiasedness (Affleck and Gregoire, 2015). Gains in precision were not limited to one sampling scheme. Increased precision was also seen for $n = 5$ based on other branch sampling methods: probability proportional to size sampling ($\overline{RE} = 0.94$), simple random sampling ($\overline{RE} = 0.28$), and stratified random sampling ($\overline{RE} = 0.28$). The potential gains in biomass estimation at the tree level demonstrates how SAE may prove a useful tool in other contexts than estimating forest inventory attributes for small geographic areas.

5. CONCLUSIONS

Small area estimation (SAE) is a growing area of research in forest inventory owing largely to its ability to support model-based inference in small area domains lacking sufficient sample data to provide stable estimates using purely design-based estimation. A variety of modeling techniques can be employed in the SAE framework with linear mixed modeling among the most widely used to-date. A unifying requirement is the use of auxiliary data that allows estimators to both borrow strength from indirect sample data that co-vary with auxiliary observations and to provide auxiliary population parameters as predictors in synthetic models for composite estimation. The availability of large data sets like those collected in NFIs, along with increasingly available auxiliary data such as DAP, ALS, satellite remote-sensing, or other digital map products has made SAE of particular interest to forest inventory specialists. Increases in precision from SAE can provide efficient alternatives to sample intensification when insufficient sample data are available to meet needed tolerances for estimator error.

The examples summarized here demonstrate potential benefits of SAE, along with some limitations researchers have encountered in applying evolving model-assisted design-based estimators or composite estimators that lie within unit-level or area-level SAE frameworks. As with any model-based approaches considerable attention should be paid to model assumptions including distributional assumptions for residuals, correct model form (e.g., linear vs. non-linear), careful selection of model predictors, and accounting for correlation structures among synthetic model residuals. Challenges may arise from a lack of correspondence between sample and auxiliary data such as when precise pairing of (x, y) observations is impractical, or when auxiliary data are unavailable for specific areas or time periods of interest. Consideration of alternative approaches including design-based and model-assisted estimation, area-level, or unit-level (model-based) SAE is needed to ensure suitability of methods given inferential needs. Many topics for further research have been identified in the literature reviewed here, pointing

out opportunities for future improvement in forest inventory applications of SAE. Although no tool can address needs in all circumstances, methods like those reviewed here provide flexible, efficient alternatives to reduce the need for sample intensification in many cases and to meet tolerance specifications in others at reduced cost by increasing estimator efficiency.

AUTHOR CONTRIBUTIONS

GD, PR, and JC: conception, design, and outline. GD and PR: literature review and document drafts. JC, PG, BW, and GM: reviewed drafts, provided corrections and revisions, discussion of needed changes, additions, and redesign. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by USDA Forest Service Southern Research Station and Virginia Tech Department of Forest Resources and Environmental Conservation, Joint Venture Agreement 20-JV-11330145-074. Sponsorship of research into improving estimation of timber damage in tropical storms in the southeastern United States. Funds from the Virginia Tech Library paid for open-access fees.

REFERENCES

Affleck, D. L. R., and Gregoire, T. G. (2015). Generalized and synthetic regression estimators for randomized branch sampling. *Forestry* 88, 599–611. doi: 10.1093/forestry/cpv027

Anderson, H. E., and Breidenbach, J. (2007). "Statistical properties of mean stand biomass estimators in a LIDAR-based double sampling forest survey design," in *Proceedings of the ISPRS Workshop on Laser Scanning and SilviLaser* (Espoo), 8–13.

Arima, S., Bell, W. R., Datta, G. S., Franco, C., and Liseo, B. (2017). Multivariate fay-herriot bayesian estimation of small area means under functional measurement error. *J. R. Stat. Soc. Ser. A Stat Soc.* 180, 1191–1209. doi: 10.1111/rssa.12321

Astrup, R., Rahlf, J., Bjorkelo, K., Debella-Gilo, M., Gjertsen, A.-K., and Breidenbach, J. (2019). Forest information at multiple scales: development, evaluation and application of the Norwegian forest resources map SR16. *Scand. J. Forest Res.* 34, 484–496. doi: 10.1080/02827581.2019.1588989

Babcock, C., Finley, A. O., Andersen, H.-E., Pattison, R., Cook, B. D., Morton, D. C., et al. (2018). Geostatistical estimation of forest biomass in interior Alaska combining Landsat-derived tree cover, sampled airborne lidar and field observations. *Remote Sens. Environ.* 212, 212–230. doi: 10.1016/j.rse.2018.04.044

Babcock, C., Finley, A. O., Bradford, J. B., Kolka, R., Birdsey, R., and Ryan, M. G. (2015). LiDAR based prediction of forest biomass using hierarchical models with spatially varying coefficients. *Remote Sens. Environ.* 169, 113–127. doi: 10.1016/j.rse.2015.07.028

Babcock, C., Matney, J., Finley, A. O., Weiskittel, A., and Cook, B. D. (2013). Multivariate spatial regression models for predicting individual tree structure variables using lidar data. *IEEE J. Select. Top. Appl. Earth Observat. Remote Sens.* 6, 6–14. doi: 10.1109/JSTARS.2012.2215582

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, 10. doi: 10.1080/01621459.1988.10478561

Bechtold, W. A., and Patterson, P. L. (2005). *The Enhanced Forest Inventory and Analysis Program - National Sampling Design and Estimation Procedures.* Technical report. U.S. Department of Agriculture, Forest Service, Southern Research Station.

Benavent, R., and Morales, D. (2016). Multivariate Fay-Herriot models for small area estimation. *Comput. Stat. Data Anal.* 94, 372–390. doi: 10.1016/j.csda.2015.07.013

Breidenbach, J. (2018). *JoSAE: Unit-Level and Area-Level Small Area Estimation.* Comprehensive R Archive Network. Available online at: https://CRAN.R-project.org/package=JoSAE (accessed February 28, 2022).

Breidenbach, J., and Astrup, R. (2012). Small area estimation of forest attributes in the Norwegian National Forest Inventory. *Eur. J. Forest Res.* 131, 1255–1267. doi: 10.1007/s10342-012-0596-7

Breidenbach, J., Granhus, A., Hylen, G., Eriksen, R., and Astrup, R. (2020). A century of National Forest Inventory in Norway-informing past, present, and future decisions. *Forest Ecosyst.* 7, 46. doi: 10.1186/s40663-020-00261-0

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

Breidenbach, J., McRoberts, R. E., and Astrup, R. (2016). Empirical coverage of model-based variance estimators for remote sensing assisted estimation of stand-level timber volume. *Remote Sens. Environ.* 173, 274–281. doi: 10.1016/j.rse.2015.07.026

Breidenbach, J., Nothdurft, A., and Kandler, G. (2010). Comparison of nearest neighbour approaches for small area estimation of tree species-specific forest inventory attributes in central Europe using airborne laser scanner data. *Eur. J. Forest Res.* 129, 833–846. doi: 10.1007/s10342-010-0384-1

Bright, B. C., Hicke, J. A., and Hudak, A. T. (2012). Estimating aboveground carbon stocks of a forest affected by mountain pine beetle in Idaho using lidar and multispectral imagery. *Remote Sens. Environ.* 124, 270–281. doi: 10.1016/j.rse.2012.05.016

Burk, T. E., and Ek, A. R. (1982). Application of empirical Bayes/James-Stein procedures to simultaneous estimation problems in forest inventory. *Forest Sci.* 28, 753–771.

Chandra, H., Salvati, N., and Chambers, R. (2015). A spatially nonstationary Fay-Herriot model for small area estimation. *J. Surv. Stat. Methodol.* 3, 109–135. doi: 10.1093/jssam/smu026

Chandra, H., Salvati, N., Chambers, R., and Tzavidis, N. (2012). Small area estimation under spatial nonstationarity. *Comput. Stat. Data Anal.* 56, 2875–2888. doi: 10.1016/j.csda.2012.02.006

Chirici, G., Mura, M., McInerney, D., Py, N., Tomppo, E. O., Waser, L. T., et al. (2016). A meta-analysis and review of the literature on the k-nearest neighbors technique for forestry applications that use remotely sensed data. *Remote Sens. Environ.* 176, 282–294. doi: 10.1016/j.rse.2016.02.001

Costa, A., Satorra, A., and Ventura, E. (2003). "An empirical evaluation of small area estimators," in *Economics Working Papers 674* (Barcelona: Department of Economics and Business, Pompeu Fabra University). Available online at: https://ideas.repec.org/p/upf/upfgen/674.html

Costa, A., Satorra, A., and Ventura, E. (2009). On the performance of small-area estimators: fixed vs. random area parameters. *Sort* 33, 85–104. Available online at: http://dmle.icmat.es/pdf/SORT_2009_33_01_04.pdf (accessed February 28, 2022).

Coulston, J. W., Green, P. C., Radtke, P. J., Prisley, S. P., Brooks, E. B., Thomas, V. A., et al. (2021). Enhancing the precision of broad-scale forestland removals estimates with small area estimation techniques. *Forestry Int. J. Forest Res.* 94, 427–441. doi: 10.1093/forestry/cpaa045

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *J. Am. Stat. Assoc.* 111, 800–812. doi: 10.1080/01621459.2015.1044091

Datta, G. S., Day, B., and Maiti, T. (1998). Multivariate bayesian small area estimation: an application to survey and satellite data. *Sankhya Indian J. Stat. Ser. A* 60, 344–362.

Datta, G. S., and Mandal, A. (2015). Small area estimation with uncertain random effects. *J. Am. Stat. Assoc.* 110, 1735–1744. doi: 10.1080/01621459.2015.1016526

Fay, R. E., 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, 269–277. doi: 10.1080/01621459.1979.10482505

Federal Committee on Statistical Methodology (1993). *Indirect Estimators in Federal Programs*. Technical report, Office of Management and Budget, Washington, DC.

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, 31–48. doi: 10.1198/jasa.2011.ap09653

Finley, A. O., Banerjee0, S., Zhou, Y., Cook, B. D., and Babcock, C. (2017). Joint hierarchical models for sparsely sampled high-dimensional lidar and forest variables. *Remote Sens. Environ*. 190, 149–161. doi: 10.1016/j.rse.2016.12.004

Finley, A. O., Bsnerjee, S., and Gelfand, A. E. (2015). spBayes for large univariate and multivariate point-referenced spatio-temporal data models. *J. Stat. Softw*. 63, 1–28. doi: 10.18637/jss.v063.i13

Frank, B., Mauro, F., and Temesgen, H. (2020). Model-based estimation of forest inventory attributes using Lidar: a comparison of the area-based and semi-individual tree crown approaches. *Remote Sens*. 12, 2525. doi: 10.3390/rs12162525

Fuller, W. A. (1999). Environmental surveys over time. *J. Agric. Biol. Environ. Stat*. 4, 331–354. doi: 10.2307/1400493

Ghosh, M., and Rao, J. N. K. (1994). Small area estimation: an appraisal. *Stat. Sci*. 9, 55–76. doi: 10.1214/ss/1177010647

Goerndt, M. E., Monleon, V. J., and Temesgen, H. (2011). A comparison of small-area estimation techniques to estimate selected stand attributes using LiDAR-derived auxiliary variables. *Can. J. Forest Res*. 41, 1189–1201. doi: 10.1139/x11-033

Goerndt, M. E., Monleon, V. J., and Temesgen, H. (2013). Small-area estimation of county-level forest attributes using ground data and remote sensed auxiliary information. *Forest Sci*. 59, 536–548. doi: 10.5849/forsci.12-073

Goerndt, M. E., Wilson, B. T., and Aguilar, F. X. (2019). Comparison of small area estimation methods applied to biopower feedstock supply in the Northern U.S. region. *Biomass Bioenergy* 121, 64–77. doi: 10.1016/j.biombioe.2018.12.008

González -Manteiga, W., and Morales, D. (2008). Bootstrap mean squared error of small-area EBLUP. *J. Stat. Comput. Simul*. 78, 443–462. doi: 10.1080/00949650601141811

Green, P. C., Burkhart, H. E., Coulston, J. W., and Radtke, P. J. (2020). A novel application of small area estimation in loblolly pine forest inventory. *Forestry Int. J. Forest Res*. 93, 444–457. doi: 10.1093/forestry/cpz073

Gregoire, T., and Valentine, H. (2007). *Sampling Strategies for Natural Resources and the Environment*. Boca Raton, FL: Chapman and Hall/CRC. doi: 10.1201/9780203498880

Gregoire, T. G. (1998). Design-based and model-based inference in survey sampling: appreciating the difference. *Can. J. Forest Res*. 28, 1429–1447. doi: 10.1139/x98-166

Guldin, R. W. (2021). A systematic review of small domain estimation research in forestry during the twenty-first century from outside the United States. *Front. Forests Glob. Change* 4, 695929. doi: 10.3389/ffgc.2021.695929

He, Z., and Sun, D. (2000). Hierarchical Bayes estimation of hunting success rates with spatial correlations. *Biometrics* 56, 360–367. doi: 10.1111/j.0006-341X.2000.00360.x

Hill, A., Massey, A., and Mandallaz, D. (2021). The R package forestinventory: design-based global and small area estimations for multiphase forest inventories. *J. Stat. Softw*. 97, 1–40. doi: 10.18637/jss.v097.i04

Irulappa-Pillai-Vijayakumar, D. B., Renaud, J.-P., Morneau, F., McRoberts, R. E., and Vega, C. (2019). Increasing precision for french forest inventory estimates using the k-nn technique with optical and photogrammetric data and model-assisted estimators. *Remote Sens*. 11, 991. doi: 10.3390/rs11080991

James, W., and Stein, C. (1961). "Estimation with quadratic loss," in *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, ed J. Neyman (Berkeley, CA: University of California Press), 361–379.

Järnstedt, J., Pekkarinen, A., Tuominen, S., Ginzler, C., Holopainen, M., and Viitala, R. (2012). Forest variable estimation using a high-resolution digital surface model. *ISPRS J. Photogrammetry Remote Sens*. 74, 78–84. doi: 10.1016/j.isprsjprs.2012.08.006

Johnson, A. A., Breidt, F. J., and Opsomer, J. D. (2008). Estimating distribution functions from survey data using nonparametric regression. *J. Stat. Theory Pract*. 2, 419–431. doi: 10.1080/15598608.2008.10411884

Kangas, A., Myllymäki, M., Gobakken, T., and Næsset, E. (2016). Model-assisted forest inventory with parametric, semiparametric, and nonparametric models. *Can. J. Forest Res*. 46, 855–868. doi: 10.1139/cjfr-2015-0504

Kotivuori, E., Korhonen, L., and Packalen, P. (2016). Nationwide airborne laser scanning based models for volume, biomass and dominant height in Finland. *Silva Fennica* 50, 1567. doi: 10.14214/sf.1567

Latifi, H., Nothdurft, A., and Koch, B. (2010). Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: application of multiple optical/LiDAR-derived predictors. *Forestry* 83, 395–407. doi: 10.1093/forestry/cpq022

Lehtonen, R., and Veijanen, A. (2009). "Chapter 31: Design-based methods of estimation for domains and small areas," in *Handbook of Statistics*, ed C. R. Rao (Amsterdam: Elsevier), 219–249. doi: 10.1016/S0169-7161(09)00231-4

Lochhead, K., LeMay, V., Bull, G., Schwab, O., and Halperin, J. (2018). Multivariate estimation for accurate and logically consistent forest-attributes maps at macroscales. *Can. J. Forest Res*. 48, 345–359. doi: 10.1139/cjfr-2017-0221

Magnussen, S., Mandallaz, D., Breidenbach, J., Lanz, A., and Ginzler, C. (2014). National forest inventories in the service of small area estimation of stem volume. *Can. J. Forest Res*. 44, 1079–1090. doi: 10.1139/cjfr-2013-0448

Magnussen, S., Mauro, F., Breidenbach, J., Lanz, A., and Kändler, G. (2017). Area-level analysis of forest inventory variables. *Eur. J. Forest Res*. 136, 839–855. doi: 10.1007/s10342-017-1074-z

Magnussen, S., Nord-Larsen, T., and Riis-Nielsen, T. (2018). Lidar supported estimators of wood volume and aboveground biomass from the Danish National Forest Inventory (2012–2016). *Remote Sens. Environ*. 211, 146–153. doi: 10.1016/j.rse.2018.04.015

Mandallaz, D. (2013). Design-based properties of some small-area estimators in forest inventory with two-phase sampling. *Can. J. Forest Res*. 43, 441–449. doi: 10.1139/cjfr-2012-0381

Mandallaz, D., Breschan, J., and Hill, A. (2013). New regression estimators in forest inventories with two-phase sampling and partially exhaustive information: a design-based Monte Carlo approach with applications to small-area estimation. *Can. J. Forest Res*. 43, 1023–1031. doi: 10.1139/cjfr-2013-0181

Mauro, F., Molina, I., Garcia-Abril, A., Valbuena, R., and Ayuga-Tellez, E. (2016). Remote sensing estimates and measures of uncertainty for forest variables at different aggregation levels. *Environmetrics* 27, 225–238. doi: 10.1002/env.2387

Mauro, F., Monleon, V. J., Temesgen, H., and Ford, K. R. (2017). Analysis of area level and unit level models for small area estimation in forest inventories assisted with LiDAR auxiliary information. *PLoS ONE* 12, e0189401. doi: 10.1371/journal.pone.0189401

Mauro, F., Ritchie, M., Wing, B., Frank, B., Monleon, V., Temesgen, H., et al. (2019). Estimation of changes of forest structural attributes at three different spatial aggregation levels in Northern California using multitemporal LiDAR. *Remote Sens*. 11, 923. doi: 10.3390/rs11080923

McConville, K., Tang, B., Zhu, G., Cheung, S., and Li, S. (2018). *mase: Model-Assisted Survey Estimation*. Comprehensive R Archive Network. Available online at: https://cran.r-project.org/package=mase (accessed February 28, 2022).

McConville, K. S., Breidt, F. J., Lee, T. C. M., and Moisen, G. G. (2017). Model-assisted survey regression estimation with the Lasso. *J. Surv. Stat. Methodol*. 5, 131–158. doi: 10.1093/jssam/smw041

McConville, K. S., Moisen, G. G., and Frescino, T. S. (2020). A tutorial on model-assisted estimation with application to forest inventory. *Forests* 11, 244. doi: 10.3390/f11020244

McRoberts, R. E. (2012). Estimating forest attribute parameters for small areas using nearest neighbors techniques. *Forest Ecol. Manage*. 272, 3–12. doi: 10.1016/j.foreco.2011.06.039

McRoberts, R. E., Chen, Q., and Walters, B. F. (2017). Multivariate inference for forest inventories using auxiliary airborne laser scanning data. *Forest Ecol. Manage*. 401, 295–303. doi: 10.1016/j.foreco.2017.07.017

McRoberts, R. E., Næsset, E., and Gobakken, T. (2013). Inference for lidar-assisted estimation of forest growing stock volume. *Remote Sens. Environ*. 128, 268–275. doi: 10.1016/j.rse.2012.10.007

McRoberts, R. E., Tomppo, E. O., Finley, A. O., and Heikkinen, J. (2007). Estimating areal means and variances of forest attributes using the k-Nearest Neighbors technique and satellite imagery. *Remote Sens. Environ*. 111, 466–480. doi: 10.1016/j.rse.2007.04.002

Molina, I., and Marhuenda, Y. (2015). sae: An R package for small area estimation. *R J*. 7, 81–98. doi: 10.32614/RJ-2015-007

Morris, C. N. (1983). Parametric empirical bayes inference: theory and applications. *J. Am. Stat. Assoc.* 78, 47–55. doi: 10.1080/01621459.1983.10477920

Næsset, E. (2002). Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. *Remote Sens. Environ.* 80, 88–99. doi: 10.1016/S0034-4257(01)00290-5

Næsset, E., Gobakken, T., Bollandss, O. M., Gregoire, T. G., Nelson, R., and Sthl, G. (2013). Comparison of precision of biomass estimates in regional field sample surveys and airborne LiDAR-assisted surveys in Hedmark County, Norway. *Remote Sens. Environ.* 130, 108–120. doi: 10.1016/j.rse.2012.11.010

Næsset, E., Gobakken, T., Solberg, S., Gregoire, T. G., Nelson, R., Stahl, G., et al. (2011). Model-assisted regional forest biomass estimation using LiDAR and InSAR as auxiliary data: A case study from a boreal forest area. *Remote Sens. Environ.* 115, 3599–3614. doi: 10.1016/j.rse.2011.08.021

Nagle, N. N., Schroeder, T. A., and Rose, B. (2019). A regularized raking estimator for small-area mapping from forest inventory surveys. *Forests* 10, 1045. doi: 10.3390/f10111045

Nilsson, M., Nordkvist, K., Jonzan, J., Lindgren, N., Axensten, P., Wallerman, J., et al. (2017). A nationwide forest attribute map of Sweden predicted using airborne laser scanning data and field data from the national forest inventory. *Remote Sens. Environ.* 194, 447–454. doi: 10.1016/j.rse.2016.10.022

Nord-Larsen, T., and Schumacher, J. (2012). Estimation of forest resources from a country wide laser scanning survey and national forest inventory data. *Remote Sens. Environ.* 119, 148–157. doi: 10.1016/j.rse.2011.12.022

Novo-Fernandez, A., Barrio-Anta, M., Recondo, C., Camara-Obregon, A., and Lopez-Sanchez, C. A. (2019). Integration of national forest inventory and nationwide airborne laser scanning data to improve forest yield predictions in north-western Spain. *Remote Sens.* 11, 1693. doi: 10.3390/rs11141693

Ohmann, J. L., and Gregory, M. J. (2002). Predictive mapping of forest composition and structure with direct gradient analysis and nearest- neighbor imputation in coastal Oregon, USA. *Can. J. Forest Res.* 32, 725–741. doi: 10.1139/x02-011

Opsomer, J. D., Breidt, F. J., Moisen, G. G., and Kauermann, G. (2007). Model-assisted estimation of forest resources with generalized additive models. *J. Am. Stat. Assoc.* 102, 400–409. doi: 10.1198/016214506000001491

Pascual, C., Mauro, F., Abril, A. G., and Manzanera, J. A. (2018). "Applications of ALS (Airborne Laser Scanning) data to forest inventory. Experiences with pine stands from mountainous environments in Spain," in *IOP Conference Series-Earth and Environmental Science* (Voronezh), 1–12. doi: 10.1088/1755-1315/226/1/012001

Patterson, P. L., Healey, S. P., Stahl, G., Saarela, S., Holm, S., Andersen, H. E., et al. (2019). Statistical properties of hybrid estimators proposed for GEDI-NASA's global ecosystem dynamics investigation. *Environ. Res. Lett.* 14, 065007. doi: 10.1088/1748-9326/ab18df

Pfeffermann, D. (2002). Small area estimation–New developments and directions. *Int. Stat. Rev.* 70, 125–143. doi: 10.1111/j.1751-5823.2002.tb00352.x

Pfeffermann, D. (2013). New important developments in small area estimation. *Stat. Sci.* 28, 40–68. doi: 10.1214/12-STS395

Popescu, S. C., Zhao, K., Neuenschwander, A., and Lin, C. (2011). Satellite lidar vs. small footprint airborne lidar: comparing the accuracy of aboveground biomass estimates and forest structure metrics at footprint level. *Remote Sens. Environ.* 115, 2786–2797. doi: 10.1016/j.rse.2011.01.026

Prasad, N., and Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *J. Am. Stat. Assoc.* 85, 163–171. doi: 10.1080/01621459.1990.10475320

Radtke, P., and Bolstad, P. (2001). Laser point-quadrat sampling for estimating foliage height profiles in broad-leaved forests. *Can. J. Forest Res.* 31, 410–418. doi: 10.1139/x00-182

Rahlf, J., Breidenbach, J., Solberg, S., Næsset, E., and Astrup, R. (2014). Comparison of four types of 3d data for timber volume estimation. *Remote Sens. Environ.* 155, 325–333. doi: 10.1016/j.rse.2014.08.036

Rahman, A., and Harding, A. (2017). *Small Area Estimation and Microsimulation Modeling*. Boca Raton, FL: CRC Press. doi: 10.1201/9781315372143

Rao, J. N. K. (2008). Some methods for small area estimation. *Rivista Internazionale di Scienze Sociali* 116, 387–405. Available online at: http://www.jstor.org/stable/41625216 (accessed February 28, 2022).

Rao, J. N. K., and Molina, I. (2015). *Small Area Estimation, 2nd Edn*. Hoboken, NJ: John Wiley & Sons, Inc. doi: 10.1002/9781118735855

Reich, R. M., and Aguirre-Bravo, C. (2009). Small-area estimation of forest stand structure in Jalisco, Mexico. *J. Forest. Res.* 20, 285–292. doi: 10.1007/s11676-009-0050-y

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York, NY: Springer Science & Business Media. doi: 10.1007/978-1-4612-4378-6

Särndal, C. E. (1984). Design-consistent versus model-dependent estimation for small domains. *J. Am. Stat. Assoc.* 79, 624–631. doi: 10.1080/01621459.1984.10478089

Schaible, W. L. (1993). "Indirect estimators: definition, characteristics, and recommendations," in *Proceedings of the Survey Research Methods Section* (San Francisco, CA: American Statistical Association), 1–10.

Schreuder, H. T., Gregoire, T. G., and Wood, G. B. (1993). *Sampling Methods for Multiresource Forest Inventory*. New York, NY: John Wiley & Sons.

Schumacher, F. X. (1945). Statistical method in forestry. *Biometr. Bull.* 1, 29–32. doi: 10.2307/3001954

Shiver, B., and Borders, B. (1996). *Sampling Techniques for Forest Resource Inventory*. New York, NY: Wiley.

Skinner, C., and Wakefield, J. (2017). Introduction to the design and analysis of complex survey data. *Stat. Sci.* 32, 165–175. doi: 10.1214/17-STS614

Stahl, G., Saarela, S., Schnell, S., Holm, S., Breidenbach, J., Healey, S. P., et al. (2016). Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation. *Forest Ecosyst.* 3, 5. doi: 10.1186/s40663-016-0064-9

Stehman, S. V. (2009). Model-assisted estimation as a unifying framework for estimating the area of land cover and land-cover change from remote sensing. *Remote Sens. Environ.* 113, 2455–2462. doi: 10.1016/j.rse.2009.07.006

Steinmann, K., Mandallaz, D., Ginzler, C., and Lanz, A. (2013). Small area estimations of proportion of forest and timber volume combining Lidar data and stereo aerial images with terrestrial data. *Scand. J. Forest Res.* 28, 373–385. doi: 10.1080/02827581.2012.754936

Sterba, S. K. (2009). Alternative model-based and design-based frameworks for inference from samples to populations: from polarization to integration. *Multivariate Behav. Res.* 44, 711–740. doi: 10.1080/00273170903333574

Sugasawa, S., and Kubokawa, T. (2020). Small area estimation with mixed models: a review. *Jpn. J. Stat. Data Sci.* 3, 693–720. doi: 10.1007/s42081-020-00076-x

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. Forests Glob. Change* 4:745916. doi: 10.3389/ffgc.2021.745916

Thompson, S. K. (2012). *Sampling, 3rd Edn*. Hoboken, NJ: John Wiley &Sons. doi: 10.1002/9781118162934

Tipton, J., Opsomer, J., and Moisen, G. (2013). Properties of endogenous post-stratified estimation using remote sensing data. *Remote Sens. Environ.* 139, 130–137. doi: 10.1016/j.rse.2013.07.035

Valliant, R. (1987). Generalized variance functions in stratified two-stage sampling. *J. Am. Stat. Assoc.* 82, 499–508. doi: 10.1080/01621459.1987.10478454

Ver Planck, N. R., Finley, A. O., Kershaw, J. A., 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

Wang, J., and Fuller, W. A. (2003). The mean squared error of small area predictors constructed with estimated area variances. *J. Am. Stat. Assoc.* 98, 716–723. doi: 10.1198/016214503000000620

Wang, X., Berg, E., Zhu, Z., Sun, D., and Demuth, G. (2018). Small area estimation of proportions with constraint for national resources inventory survey. *J. Agric. Biol. Environ. Stat.* 23, 509–528. doi: 10.1007/s13253-018-0329-6

Wolter, K. M. (2007). *Introduction to Variance Estimation. Statistics for Social and Behavioral Sciences, 2nd Edn.* New York, NY: Springer. doi: 10.1007/978-0-387-35099-8_15

Conflict of Interest: The 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.

this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Forests and Global Change

ORIGINAL RESEARCH
published: 12 April 2022
doi: 10.3389/ffgc.2022.745874

Simplifying Small Area Estimation With rFIA: A Demonstration of Tools and Techniques

Hunter Stanke[1,2]*, *Andrew O. Finley*[1] *and Grant M. Domke*[3]

[1] *Department of Forestry, Michigan State University, East Lansing, MI, United States,* [2] *School of Environmental and Forest Sciences, University of Washington, Seattle, WA, United States,* [3] *Forest Service, Northern Research Station, US Department of Agriculture, St. Paul, MN, United States*

OPEN ACCESS

Edited by:
Gretchen Moisen, United States Forest Service (USDA), United States

Reviewed by:
Kelly McConville, Reed College, United States
Andrew Lister, United States Forest Service (USDA), United States

***Correspondence:**
Hunter Stanke
stankehu@msu.edu

Specialty section:
This article was submitted to Forest Management, a section of the journal Frontiers in Forests and Global Change

Received: *22 July 2021*
Accepted: *23 February 2022*
Published: *12 April 2022*

Citation:
Stanke H, Finley AO and Domke GM (2022) Simplifying Small Area Estimation With rFIA: A Demonstration of Tools and Techniques. Front. For. Glob. Change 5:745874. doi: 10.3389/ffgc.2022.745874

The United States (US) Department of Agriculture Forest Service Forest Inventory and Analysis (FIA) program operates the national forest inventory of the US. Traditionally, the FIA program has relied on sample-based approaches—permanent plot networks and associated design-based estimators—to estimate forest variables across large geographic areas and long periods of time. These approaches generally offer unbiased inference on large domains but fail to provide reliable estimates for small domains due to low sample sizes. Rising demand for small domain estimates will thus require the FIA program to adopt non-traditional estimation approaches that are capable of delivering defensible estimates of forest variables at increased spatial and temporal resolution, without the expense of collecting additional field data. In light of this challenge, the development of small area estimation (SAE) methods—estimation techniques that support inference on small domains—for FIA data has become an active and highly productive area of research. Yet, SAE methods remain difficult to apply to FIA data, due in part to the complex data structures and survey design used by the FIA program. Herein, we present the potential of rFIA, an open-source R package designed to increase the accessibility of FIA data, to simplify the application of a broad suite of SAE methods to FIA data. We demonstrate this potential *via* two case studies: (1) estimation of contemporary county-level forest carbon stocks across the conterminous US using a spatial Fay-Herriot model; and (2) temporally-explicit estimation of multi-decadal trends in merchantable wood volume in Washington County, Maine using a Bayesian multi-level model. In both cases, we show the application of SAE techniques offers considerable improvements in precision over FIA's traditional, post-stratified estimators. Finally, we offer a discussion of the potential role that rFIA and other open-source tools might play in accelerating the adoption of SAE techniques among users of FIA data.

Keywords: forest inventory and analysis (FIA), R package, forest carbon, merchantable wood volume, Bayesian mixed-effects models, spatial Fay-Herriot models, area-level models, unit-level models

INTRODUCTION

The United States (US) Department of Agriculture Forest Inventory and Analysis (FIA) program conducts the US national forest inventory (NFI), collecting data describing the condition of forest ecosystems on a large network of permanent inventory plots distributed across all lands in the nation (Smith, 2002). These data offer a unique and powerful resource for determining the extent, magnitude, and causes of long-term changes in forest health, timber resources, and forest landowner characteristics across large regions in the US (Wurtzebach et al., 2020). The FIA program has traditionally relied on post-stratification to improve precision of point and change estimates (Bechtold and Patterson, 2005; Westfall et al., 2011). Like other NFIs (Köhl et al., 2006; Breidenbach and Astrup, 2012), FIA has experienced increased demand for estimates within smaller spatial, temporal, and biophysical domains than post-stratification can reasonably deliver (e.g., annual, stand-level estimates). The development of estimation techniques that support inference on small domains—referred to as small area estimation (SAE) methods—using FIA data is an active area of research, with considerable progress made in the last decade (Schroeder et al., 2014; Lister et al., 2020; Coulston et al., 2021; Hou et al., 2021). SAE methods are numerous and diverse, though most seek to improve inference on small domains by making use of statistical models and auxiliary information that is correlated with target variables (Rao and Molina, 2015).

Despite recent progress in SAE method development, many FIA data users are likely to find such techniques difficult to implement due to limitations in data accessibility and complexity in survey design. Here, we demonstrate the potential of rFIA (Stanke et al., 2020), an open-source R package (R Core Team, 2021), to reduce barriers in data access that arise from complexity in data coding, database structure, and Structured Query Language used by the FIA program. Using a simple yet powerful design, rFIA implements the post-stratified, design-based estimation procedures described in Bechtold and Patterson (2005) for over 60 forest variables and allows users to return intermediate summaries of all variables for use in modeling studies (i.e., plot, condition, and/or tree-level). Further, target variables can be easily estimated for domains defined by any combination of spatial zones (i.e., spatial polygons), temporal extents (e.g., most recent measurements), and/or biophysical attributes (e.g., species, site classifications).

Model-based SAE techniques offer a valuable alternative to the design-based, post-stratified estimators implemented in rFIA. Model-based SAE methods often seek to borrow information from non-target domains (e.g., from neighboring spatial zones if domains are defined by spatial boundaries) and auxiliary data (e.g., remote sensing data) to improve precision of estimated quantities for a domain of interest, and can generally be classified into two distinct groups: unit-level and domain-level (also referred to as area-level) models. Unit-level models are constructed at the level of population units, where population units are defined as the minimal units that can be sampled from a target population. With respect to FIA's survey design, field plots represent population units (in the finite population sense) and target populations are defined by any spatial and/or temporal region with known extent. Unit-level models relate target variables measured on sampled population units to auxiliary data that is available for all population units (e.g., wall-to-wall remote sensing data) in order to predict quantities of the target variables for a domain of interest (i.e., where domains are defined by some combination of population units; Rao and Molina, 2015). In contrast, domain-level models are constructed at the level of domains. Here, domain-specific auxiliary information (e.g., county-level census data, where counties represent domains) is related to post-stratified or direct estimates within corresponding domains (Rao and Molina, 2015). Hence, domain-level models effectively "adjust" direct domain estimates in light of auxiliary information.

By design, rFIA does not implement model-based SAE techniques directly, owing to their exceptional variety and requirements for thorough model checking and validation. Rather, rFIA automates the process of summarizing FIA data to a form that is appropriate for input to a wide variety of unit- and domain-level SAE models. Hence, rFIA allows the user to focus their attention on model development and data output, as opposed to the intricacies of FIA's data structure and sampling design.

Here we present two case studies chosen to demonstrate some aspects of rFIA's potential to simplify model-based SAE applications using FIA data. First, we use the post-stratified estimators implemented in rFIA to estimate current forest carbon stocks within counties across the conterminous US (CONUS), and develop a domain-level spatial Fay-Herriot SAE model to couple these direct estimates with auxiliary climate variables and improve precision of estimated carbon stocks. Second, we derive a temporally-explicit unit-level estimator of total merchantable volume for a small spatial domain in Maine (i.e., Washington County), and compare precision of the model-based estimator to that of a design-based, post-stratified estimator of merchantable volume for the domains of interest (Washington County, all years over the period 1999–2025). Specifically, we use rFIA to extract survey design information associated with current volume inventories in the State, and produce plot-level summaries of merchantable volume for all plot visits since 1999. We then develop a Bayesian multi-level model to estimate merchantable volume at annual time-steps, and use the approach presented in Little (2004) to derive a robust model-based estimator of total merchantable volume for all domains of interest. All code and data used in these case studies are available in **Appendices A**, **B**, on GitHub (https://github.com/hunter-stanke/FGC_rFIA_SAE), and at our official website (https://rfia.netlify.app).

METHODS

FIA Data

Data Collection

Since 1999 FIA has operated an extensive nationally-consistent annual forest survey designed to monitor changes in forests across all lands in the US (Smith, 2002). The program measures forest variables on a network of permanent ground plots that are systematically distributed at a base intensity of ∼1 plot per 2,428

hectares across the US (Smith, 2002). Data collected on ground plots are stored in a large, public database (i.e., the FIA Database), however the true locations of ground plots are not released in order to protect the ecological integrity of plots and the privacy rights of private landowners (Shaw, 2008).

For trees 12.7 cm diameter at breast height (d.b.h.) and larger, tree attributes (e.g., species, live/dead, mortality agent) and variables (e.g., d.b.h., height, volume) are measured on a cluster of four 168 m^2 subplots at each plot location (Bechtold and Patterson, 2005). Trees 2.54–12.7 cm d.b.h. are measured on a microplot (13.5 m^2) contained within each subplot, and rare events such as very large trees are measured on an optional macroplot (1,012 m^2) surrounding each subplot (Bechtold and Patterson, 2005). Importantly, some variables in the FIA database, like tree biomass and carbon, are modeled from variables measured on field plots and auxiliary variables, such as mean annual temperature, that are joined with the plots based on their spatial location.

Survey Design

Traditionally, the FIA program has used post-stratification to improve precision of point and change estimates, account for variability in non-response rates, and to allow sample intensity to vary across regions (Smith, 2002; Bechtold and Patterson, 2005; Tinkham et al., 2018). Importantly, post-stratification is applied to populations defined by a set of exhaustive and mutually exclusive geographic units with known areas—known as estimation units using FIA's terminology. Estimation units are often formed from administrative boundaries, for example counties, county groups, or large ownerships and are constrained by State boundaries (i.e., estimation units can only fall within one State). FIA implements post-stratification by dividing each estimation unit into relatively homogeneous strata using wall-to-wall remotely-sensed imagery. Strata are designed to minimize within-strata sample variances, while ensuring constant within-strata sample intensity. In short, FIA's survey design is hierarchical and area-based: States are comprised of multiple estimation units, estimation units are divided into multiple strata, and strata contain multiple inventory plots. We refer readers to Bechtold and Patterson (2005) for a complete description of FIA's post-stratified survey design.

FIA uses an annual panel system to estimate current inventories and change. Inventory cycles—the period of time required to measure all ground plots with at least one forest condition within an estimation unit—are generally 5–7 years in length in the eastern US, and 10 years in length in the western US (Bechtold and Patterson, 2005). A mutually exclusive and spatially-balanced subset of ground plots with at least one forest condition are measured in each year of an inventory cycle, forming a series of independent annual panels. For example in an ideal 5-year inventory cycle, 20% of ground plots are measured annually, such that 100% of plots are measured once between Year 1 and Year 5. In Year 6, the subset of plots measured in Year 1 are remeasured, and a second inventory cycle emerges consisting of all plots measured between Year 2 and Year 6 (not independent of the previous cycle, as 80% of measurements are shared).

Precision of point and change estimates can often be improved by combining annual panels within an inventory cycle (i.e., by augmenting current data with data collected previously). While FIA does not prescribe a core procedure for combining panels (Bechtold and Patterson, 2005), the temporally-indifferent approach, which effectively pools data from annual panels into a single periodic inventory, is the most widely known and used. From our example 5-year inventory cycle above, the temporally-indifferent approach pools all data collected between Years 1 and 5 and computes point estimates from the aggregated sample, assuming all plots are measured simultaneously at the end of the inventory cycle. Estimates of change could first be computed in Year 6 in our example (consisting of a single annual panel, 20% of remeasured plots), and change estimates for a full inventory cycle could first be computed following Year 10. In the case studies that follow, we use the periodic, or temporally-indifferent, approach to estimate contemporary carbon stocks across the CONUS, and the post-stratified estimator applied to individual annual panels to characterize temporal trends in merchantable volume in Maine. Importantly, both approaches rely on the same direct post-stratified estimator, differing only in their treatment of time as dimension of the survey design (i.e., the temporal subset of data that the estimators are applied to).

The rFIA R Package

rFIA is an open source package for the statistical computing environment R (R Core Team, 2021), and was designed to simplify the process of working with FIA data. Specifically, rFIA alleviates hurdles arising from FIA's complex survey design and database structure by offering a simple and highly flexible toolset for data acquisition and management (e.g., downloading and storing FIA data), population estimation (e.g., estimation of totals and ratios for domains of interest), and alternative summary of FIA data (e.g., plot-level summaries of forest variables). We provide a brief description of the key features of rFIA here, and refer readers to Stanke et al. (2020) for a detailed description of the package and our official website (https://rfia.netlify.app/) for example code and details regarding package installation.

Core functions in the rFIA R package can be divided into three categories: (1) utility functions designed to acquire, load, and save modifications to FIA data; (2) subset functions designed to help users navigate FIA's survey design and subset inventories of interest in their applications; and (3) estimator functions that ingest raw FIA data and produce population estimates (e.g., totals, ratios, and associated variances) or intermediate-level summaries (e.g., plot- or tree-level summaries) of forest variables within user-defined populations of interest. **Table 1** provides a brief description of the rFIA functions used in the case studies presented herein, and **Appendices A**, **B** provide all associated code required to reproduce these case studies.

By default, rFIA implements standard estimation routines used by the FIA program—post-stratified estimators and a temporally-indifferent (i.e., periodic) approach to combining annual panels within inventory cycles—to produce population estimates for more than 60 forest variables. These estimation routines have been tested extensively across Forest Service regions and potential domains of interest (e.g., defined by species,

TABLE 1 | Descriptions of core rFIA functions used in case studies presented herein.

rFIA function	Description
Utility functions	
getFIA()	Download FIA data, load into R, and optionally save to disk
readFIA()	Load FIA database into R environment from disk
Subset functions	
clipFIA()	Spatial and temporal queries for FIA data
getDesignInfo()	Extract survey design information for post-stratified inventories
Estimator functions	
carbon()	Estimate carbon stocks by IPCC forest carbon pools
volume()	Estimate merchantable volume on standing trees

land types) to ensure national-consistency and appropriate behavior of the estimators under a broad range of user-inputs. Furthermore, resulting estimates have been validated against official FIA estimation tools (i.e., EVALIDator; Miles, 2021), and found to be accurate to two decimal places for all forest variables (Stanke et al., 2020). In addition to standard estimation approaches, rFIA offers users the ability to produce population estimates for individual annual panels or combine annual panels within an inventory cycle using a moving-average approach with potentially time-decaying weights (simple, linear, or exponential moving averages). We refer readers to section 2.2 of Stanke et al. (2020) for additional details on the estimation routines implemented in rFIA.

Domain-Level Model for Forest Carbon Stocks

To demonstrate rFIA's capacity to simplify development of domain-level small area estimators, we estimate contemporary forest carbon stocks by county across the CONUS using a spatial Fay-Herriot model (Fay and Herriot, 1979; Petrucci and Salvati, 2006). This process consists of two primary stages: (1) produce post-stratified estimates of carbon stocks and associated variances for all forestland in each county (i.e., domain), and (2) "smooth" post-stratified estimates using a model constructed from domain-average climate variables and spatial random effects to improve precision of estimated quantities within each domain. **Figure 1** provides a conceptual diagram that illustrates key steps in our general estimation approach.

FIA measures/models forest carbon variables on all forested portions of inventory plots (Domke, 2022). Here, forestland is defined as land with at least 10% tree canopy cover (or had previously, or is expected to have in the future) that occurs in a patch of at least 0.4 ha in extent and that is not narrower than 37 m. The carbon() function in rFIA draws from forest carbon variables to produce population estimates of forest carbon stocks, where carbon stocks include the following ecosystem components: live overstory, live understory, standing dead wood, down dead wood, litter, and soil organic material. Here live overstory, live understory, and standing dead wood encompass both aboveground and belowground carbon stocks.

We used rFIA to download an appropriate subset of the FIA Database from the FIA DataMart (FIA DataMart, 2021), and select the most recent subset of current volume inventories within each State across the CONUS. We then used the carbon() function to estimate total carbon stocks within counties using the periodic, temporally-indifferent approach (i.e., the same methods implemented by EVALIDator; Miles, 2021). Here, total carbon stocks are a sum of all ecosystem components across public and private forestland, and are expressed as a population total. We convert estimates of population totals (tons CO2e) to population means (tons CO2e $\cdot$ ha^{-1}) by dividing population totals by the areal extent of each county (known quantities). Similarly, we convert the variance of the population total to the variance of the population mean by dividing by the square of the areal extent of each domain.

We next fit a spatial Fay-Herriot model to the post-stratified estimates of population means, using the sae R package (Molina and Marhuenda, 2015). Fay-Herriot models are widely used in small area estimation and generally use domain-level auxiliary data in an attempt to improve the precision of domain estimates for a target variable. These models are often defined in two stages, in which variability arising from imperfect observation of the target variable within a domain (e.g., variability arising from sampling) is modeled separately from variability arising from functional processes (e.g., processes represented in the auxiliary data). This framework is particularly useful as it allows estimation of relationships between auxiliary variables and the true state of a target variable, without requiring that the true state of the target variable be known. Instead, the probabilistic linkage between imperfect observations of the target variable (e.g., sample-based estimates with known error) and its true state are used to estimate these relationships, thereby allowing information to be "borrowed" across domains (e.g., *via* shared regression coefficients) and often improving the precision of domain estimates for the target variable (Molina and Marhuenda, 2015).

Let $\bar{Y}_d$ denote the estimated population mean of county d obtained *via* the post-stratified estimators from rFIA, and $\mathrm{v}(\bar{Y}_d)$ the estimated variance of $\bar{Y}_d$. Importantly, the estimators of $\bar{Y}_d$ and $\mathrm{v}(\bar{Y}_d)$ are derived under a design-based framework, and hence can be assumed unbiased for large samples (an assumption that is potentially violated for domains with few observations). The spatial Fay-Herriot model for county d in $1, 2, \ldots, D$, where D is the number of counties ($D = 3,107$), is then defined as

$$\bar{Y}_d = Z_d + \epsilon_d, \tag{1}$$

$$Z_d = \mathbf{x}_d^\top \boldsymbol{\beta} + \nu_d, \tag{2}$$

where Z_d denotes the true, but unobserved value of the population mean in county d, and ϵ_d is a normally distributed error term with zero mean and variance $\mathrm{v}(\bar{Y}_d)$. Equation (1) represents post-stratified estimates of county-level population means from rFIA as imperfect observations of true (unobserved) county-level population means. In other words, we represent the post-stratified estimate for domain d as being drawn from a normal distribution with mean Z_d (unobserved, and to be

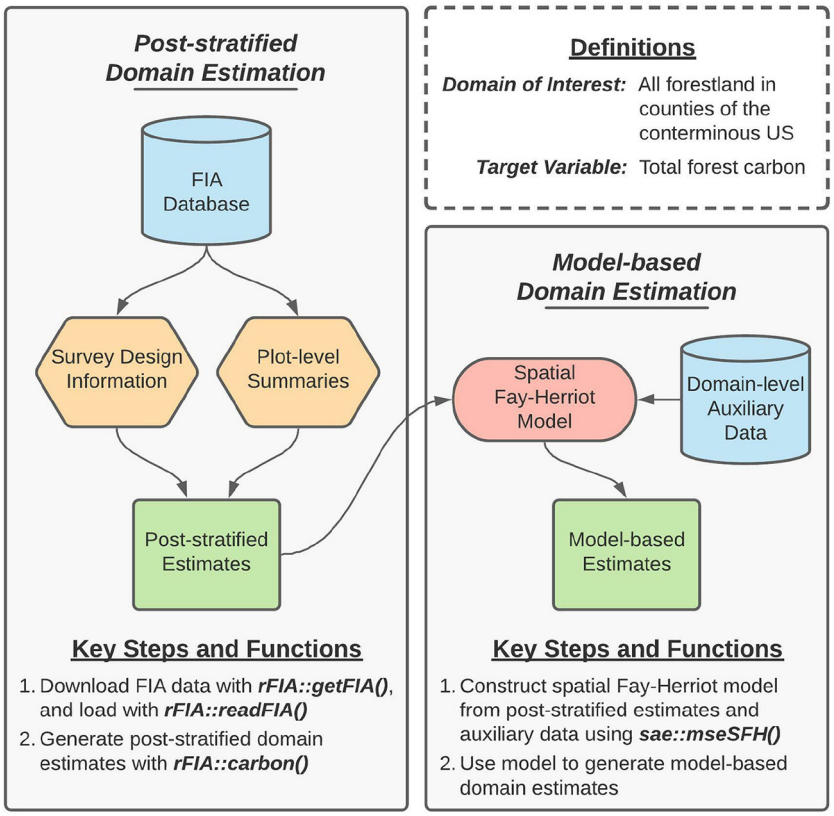


FIGURE 1 | Concept map illustrating key steps, functions, and workflows used in the development of our spatial Fay-Herriot model of forest carbon stocks in conterminous US. Here, blue cylinders represent data inputs, orange hexagons represent intermediate data products, red ovals represent models, and green rectangles represent domain estimates.

estimated) and variance $\mathrm{v}(\bar{Y}_d)$ (estimated directly from FIA data, assumed to be known).

In Equation (2), $\mathbf{x}_d$ is a vector of length three comprising an intercept and two climate predictors for county d, and $\boldsymbol{\beta}$ is an associated vector of regression coefficients. Climate predictors include mean annual temperature and precipitation, and were obtained from the long-term (30-year) climate normals hosted in the PRISM climate dataset (PRISM Climate Group, 2010). Climate normals were distributed on a 800 m^2 grid spanning the CONUS, and we took an average of grid cells within each county to produce domain-level climate predictors. The collection of county random effects $\mathbf{v} = (v_1, v_2, \ldots, v_D)^\top$ is assumed to follow a first order simultaneous autoregressive (SAR) process

$$\mathbf{v} = \rho \mathbf{W}\mathbf{v} + \boldsymbol{\tau}, \tag{3}$$

where ρ is the autocorrelation parameter defined on the range $(-1, 1)$, and each element of the vector $\boldsymbol{\tau}$ is a normally distributed error term with mean zero and variance σ_v^2. Finally, $\mathbf{W}$ is a $D \times D$ row-standardized county proximity matrix. In words, Equations (2)–(3) represent the true county-level population means (Z_d, unobserved) as a linear function of our climate predictors and a first-order spatial process which accounts for all variation in Z_d unexplained by $\mathbf{x}_d^\top \boldsymbol{\beta}$ (i.e., linear relationship between population means and climate variables).

Petrucci and Salvati (2006) present an empirical best linear unbiased predictor (EBLUP) under the Fay-Herriot model with spatially correlated random effects, and an analytic estimator of the mean squared error (MSE) of the EBLUP is described in Singh et al. (2005). We use the sae R package (Molina and Marhuenda, 2015) to fit the model described in Equations (1)–(3), and obtain the EBLUP of population means $\bar{Y}_d^{\mathrm{EBLUP}}$ and associated mean squared error $\mathrm{MSE}(\bar{Y}_d^{\mathrm{EBLUP}})$ for all domains *via* restricted maximum likelihood.

We use the relative standard error (RSE, expressed as a percentage) as a standardized measure of precision of the estimators of forest carbon stocks

$$\mathrm{RSE}_d^{PS} = \frac{100\,[\mathrm{v}(\bar{Y}_d)]^{0.5}}{\bar{Y}_d}, \tag{4}$$

$$\mathrm{RSE}_d^{EBLUP} = \frac{100\,[\mathrm{MSE}(\bar{Y}_d^{\mathrm{EBLUP}})]^{0.5}}{\bar{Y}_d^{\mathrm{EBLUP}}}. \tag{5}$$

Here, a lower RSE indicates higher precision. Following Coulston et al. (2021), we compare the precision of post-stratified (design-based) and model-based estimators of forest carbon stocks using

the ratio of their respective standard errors for each domain

$$\mathrm{SER}_d = \frac{[\mathrm{MSE}(\bar{Y}_d^{\mathrm{EBLUP}})]^{0.5}}{[\mathrm{v}(\bar{Y}_d)]^{0.5}}, \quad (6)$$

where SER_d denotes the ratio of the standard error of the post-stratified estimator (assumed unbiased) to that of the EBLUP for domain d (derived from MSE, cannot be assured to be unbiased). Hence, a SER less than one indicates the EBLUP yields more precise estimates of forest carbon stocks than the post-stratified estimator.

Unit-Level Model for Merchantable Wood Volume Trends

To demonstrate rFIA's capacity to simplify development of unit-level small domain estimators of forest variables, we use a Bayesian multi-level model to estimate multi-decadal trends in merchantable wood volume in Washington County, Maine. This process consists of four primary stages: (1) extract survey design information associated with the most recent "current volume" inventory in Maine; (2) produce plot-level summaries of merchantable volume for all FIA plot visits within our target population; (3) fit a Bayesian multi-level linear model to estimate plot- and stratum-level trends in mean merchantable volume, accounting for repeated inventory plot observations; and (4) summarize regression model coefficients using post-stratified design weights, yielding a robust model-based estimator of temporal trends in total merchantable wood volume across Washington County. Note that in this case study, domains are defined by spatial, temporal, and biophysical boundaries, i.e., by the spatial boundary of Washington County, by individual years over the period 1999–2025, and by the unknown extent of timberland (defined below) in the region. **Figure 2** provides a conceptual diagram that illustrates key steps in our general estimation approach.

FIA records merchantable wood volume of all trees (d.b.h. $\geq$ 12.7 cm) on forested inventory plots. The volume() function in rFIA uses these observations to produce population estimates and plot-level summaries of merchantable wood volume in the bole and sawtimber portions of trees. We consider net merchantable bole volume herein, defined as the volume of wood in the central stem of trees (d.b.h. $\geq$ 12.7 cm), from a 30.5 cm stump to a minimum 10.2 cm top diameter, or to where the central stem breaks into limbs all of which are $\leq$ 10.2 cm in diameter (Burrill et al., 2021). Volume loss due to rot and form defect are deducted. Further, FIA defines timberland as the subset of forestland that is capable of producing crops of industrial wood and is not withdrawn from timber utilization by legal statute or administrative regulation (i.e., it excludes wilderness areas; Burrill et al., 2021).

We used rFIA to download the Maine subset of the FIA Database from the FIA DataMart (FIA DataMart, 2021), extract survey design information (i.e., stratum and population areas) for the most recent current volume inventory in the State (2019 inventory), and summarize plot-level net merchantable bole volume for all plot-visits in the State since the onset of the annual FIA program (i.e., first plots measured in 1999). Here, plot-level summaries of merchantable volume are simply a sum of merchantable volume on all trees within our domain of interest—timberland in Washington County—at each inventory plot, expressed on a per-area basis ($m^3 \cdot ha^{-1}$). All plots outside our domain of interest (e.g., non-forested) receive a value of zero.

In the 2019 inventory, Washington County is split into three distinct estimation units (split into private and public ownerships, and inland census water). As FIA's estimation units are geographically distinct (i.e., independent populations), we combine these estimation units into a single target population representing Washington County. Importantly, FIA's estimation units should not be confused with population units in a finite sampling framework. Estimation units can be seen as minimum target populations for estimation using FIA's survey design. These populations are comprised of many population units, some of which may be sampled (i.e., plot locations).

We next formulate a multi-level linear model to characterize plot-, stratum-, and domain-level trends in merchantable wood volume from our visit-level summaries. By explicitly acknowledging the nested, hierarchical nature of FIA's survey design in our multi-level model, we can derive inference at multiple scales simultaneously (e.g., estimation of both plot- and stratum-level trends), partition estimated variance (i.e., uncertainty) across scales, and improve parameter estimates by allowing partial-pooling of information within groups (e.g., when few observations are available on a plot, estimated trends are "pulled" toward the stratum-level mean). This is in contrast to conventional approaches that may perform independent linear regressions for each plot (i.e., no pooling of information) or combine data from all plots within a stratum and perform a single linear regression (i.e., complete pooling of information) to estimate trends across scales.

Let y_{hij} denote the merchantable bole volume within our domain of interest that was observed at visit j, on plot i, belonging to stratum h. Further, let t_{hij} denote the year of visit j on plot i, relative to onset of the annual FIA program (i.e., $t = 0, 1, 2$ for plots visited in 1999, 2000, 2001, etc.). Our model is then defined as

$$y_{hij} = \alpha_{hi} + \beta_{hi} \cdot t_{hij} + \epsilon_{hij}, \quad (7)$$

where α_{hi} is a plot-level intercept term describing the mean merchantable volume at plot i, belonging to stratum h, in 1999 (i.e., onset of the annual FIA program, $t = 0$), and β_{hi} is a plot-level slope term describing the average annual change in mean merchantable volume at plot i, belonging to stratum h, over the period 1999–2019. The error term ϵ_{hij} is assumed normally-distributed with zero mean and constant variance.

Trends in merchantable volume are expected to vary both among plots (e.g., growth rates vary by forest type, and some plots may be harvested) and among strata (e.g., predominately forested vs. non-forested strata). We model this variability by treating plot-level parameters (α_{hi} and β_{hi}) as random effects that follow distributions defined by associated stratum-level parameters (α_h and β_h), and similarly treating stratum-level parameters as random effects that follow distributions defined by

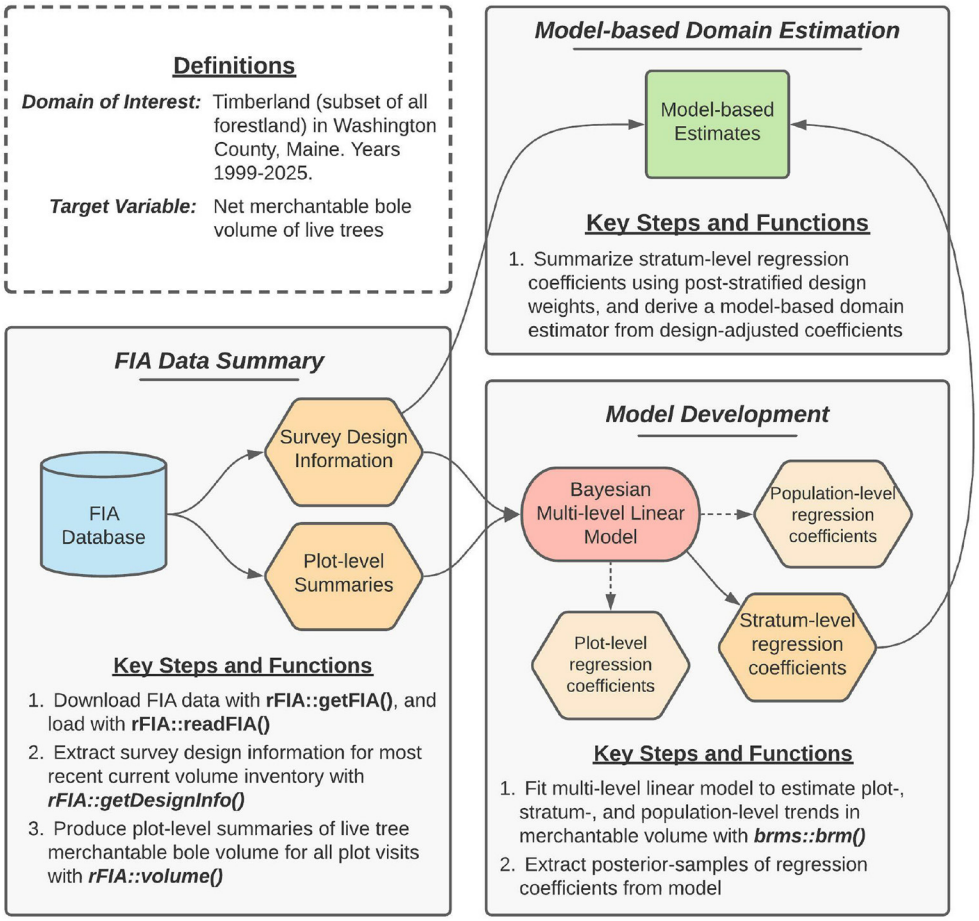


FIGURE 2 | Concept map illustrating key steps, functions, and workflows used in to estimate multi-decadal trends in merchantable wood volume in Washington County, Maine. Here, blue cylinders represent data inputs, orange hexagons represent intermediate data products, red ovals represent models, and green rectangles represent domain estimates.

a set of population-level parameters (α and β):

$$\alpha_{hi} \sim \text{normal}(\alpha_h, \sigma^2_{\alpha_{hi}}), \quad (8)$$

$$\beta_{hi} \sim \text{normal}(\beta_h, \sigma^2_{\beta_{hi}}), \quad (9)$$

$$\alpha_h \sim \text{normal}(\alpha, \sigma^2_{\alpha_h}), \quad (10)$$

$$\beta_h \sim \text{normal}(\beta, \sigma^2_{\beta_h}), \quad (11)$$

where $\sigma^2_{\alpha_{hi}}$ and $\sigma^2_{\beta_{hi}}$ are the stratum-level (among plot) variances of the regression coefficients, and $\sigma^2_{\alpha_h}$ and $\sigma^2_{\beta_h}$ are the associated population-level (among stratum) variances. In words, Equation (7) states that plot-level trends (defined by α_{hi} and β_{hi}, for each plot in $i = \{1, \ldots, 310\}$) are estimated from data collected at each visit of an FIA plot (y_{hij}), Equations (8)–(9) state that stratum-level trends (defined by α_h and β_h, defined for each stratum in $h = \{1, \ldots, 6\}$) represent an "average" of plot-level trends for all plots within a particular stratum, and Equations (10)–(11) state that the domain-level trend represents an "average" of overall population-level trends.

To complete the Bayesian specification of Equation (7) we assigned prior distributions to all parameters. We choose weakly informative normal priors for α (i.e., mean 50, standard deviation 250) and β (i.e., mean 0, standard deviation 100), and weakly informative half student-t priors for all variance terms (i.e., mean 0, scale 100, 3 degrees of freedom; Gelman, 2006). The mean and standard deviation assigned to priors for α and β differ, as α represents a point-in-time estimate while β represents an estimate of average annual change. Hence, assigning a prior to α with a positive mean reflects our knowledge of the non-negativity of the target variable (ideally would be addressed *via* specification of a non-negative likelihood function, but is not here due to computational constraints), and assigning a prior with zero mean to β represents an assumption of no change in the population over time. Further, as β represents an annual rate, we expect it's absolute value to be considerably less than the population total at a point-in-time (e.g., α), and our assignment of a lower standard deviation to the prior on β reflects this belief. Using these priors, we estimated the model using Hamiltonian Monte Carlo (HMC) algorithms implemented in

the probabilistic programming language, Stan (Carpenter et al., 2017), and affiliated R package, brms (Bürkner, 2017). We simulated three Markov chains, for a total of 4,000 iterations per chain. We assessed convergence *via* visual inspection of traceplots, and ensured proper model specification *via* posterior predictive checks.

While the set of parameters estimated in Equations (10)–(11) (α and β) allow us to derive an estimator of population-level trends in merchantable volume, such estimators ignore variation in the size of strata (i.e., which is known from FIA's survey design) and thus may be biased toward stratum that contain a large number of plots (as sampling intensity may vary across strata, a constant relationship between plot number and stratum size cannot be assumed). This bias may be addressed, however, by adjusting population-level parameters using a product of model- and design-weights (Little, 2004). Let α_h^* and β_h^* denote a set of posterior samples of stratum-level regression coefficients observed at a single iteration of the HMC algorithm. We then compute design-adjusted estimates of population-level regression coefficients, denoted as $\hat{\alpha}^*$ and $\hat{\beta}^*$, for each set of posterior samples as

$$\hat{\alpha}^* = A^{-1} \sum_{h=1}^{H} A_h \cdot \alpha_h^*, \tag{12}$$

$$\hat{\beta}^* = A^{-1} \sum_{h=1}^{H} A_h \cdot \beta_h^*, \tag{13}$$

where A_h is the known area of stratum h, and A is the combined area of all H strata (i.e., $A = \sum_{h=1}^{H} A_h$, equivalent to the combined area of estimation units). Here, model-weights are implicit in estimates of stratum-level parameters, arising from the hierarchical nature of the model described in Equations (8)–(9). In contrast, design weights are explicit, with large strata receiving more weight than small strata. In essence, we take an area-weighted mean of regression coefficients across strata to estimate trends at the population-level, thereby explicitly acknowledging features of FIA's survey design in the construction of our model-based estimator of population parameters.

Using our adjusted population-level regression coefficients, we derive a robust model-based estimator (Little, 2004) of the population mean and total for our domains of interest, denoted as $\bar{Y}(t)^*$ and $\acute{Y}(t)^*$, respectively:

$$\bar{Y}(t)^* = \hat{\alpha}^* + \hat{\beta}^* \cdot t, \tag{14}$$

$$\acute{Y}(t)^* = A \cdot \bar{Y}_t^*. \tag{15}$$

Here, variability in $\bar{Y}(t)$ and $\acute{Y}(t)$ across posterior samples reflects uncertainty in the model-based estimator of the population parameters. We produce point estimates of population parameters and their associated variances from the posterior mean and variance, and obtain 95% interval estimates from the 2.5 to 97.5% percentiles of the posterior samples for each population parameter. Similarly, we compute the relative standard error for each estimator as the ratio of the posterior standard deviation to the posterior mean.

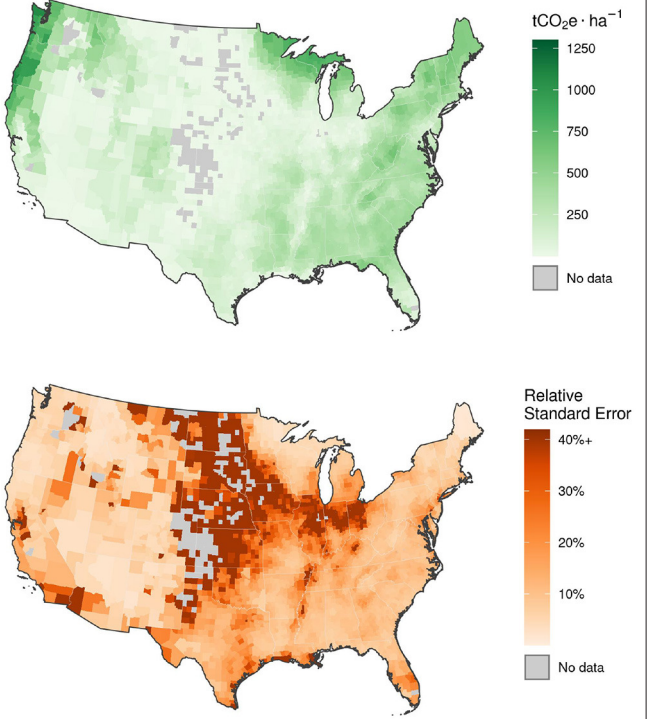


FIGURE 3 | County-level estimates of mean forest carbon density (tons CO_2 equivalent per hectare, $tCO_2e \cdot ha^{-1}$) produced by the spatial Fay-Herriot model with climate predictors **(top)**, and associated relative standard error (%; **bottom**). Gray shaded counties indicate no forested FIA plots were encountered in the county during the most recent current volume inventory, i.e., post-stratified estimator of total forest carbon and associated variance for the county are equal to zero.

Finally, we evaluate the performance of the model-based estimator of trends in total merchantable volume by comparing model-based population estimates to post-stratified annual estimates for the same population of interest over the period 1999–2019. All post-stratified estimates were computed using the annual approach implemented in the volume function in rFIA, and hence represent estimates of individual annual panels. We have elected to use estimates for annual panels because our domains are partially defined by individual years. A direct estimator then, by definition, should draw only from data collected within a particular year to produce domain estimates. Importantly, this approach differs from standard FIA estimation procedures, which pool data from multiple (up to 10) annual panels within an inventory cycle to generate domain estimates.

RESULTS

Results from design-based and model-based estimators are often not strictly comparable due to fundamental differences in their underlying inferential paradigms (see, e.g., Little, 2004). Of particular importance, design-based estimators can be reasonably assumed unbiased for large samples, whereas model-based estimators cannot be assured to be

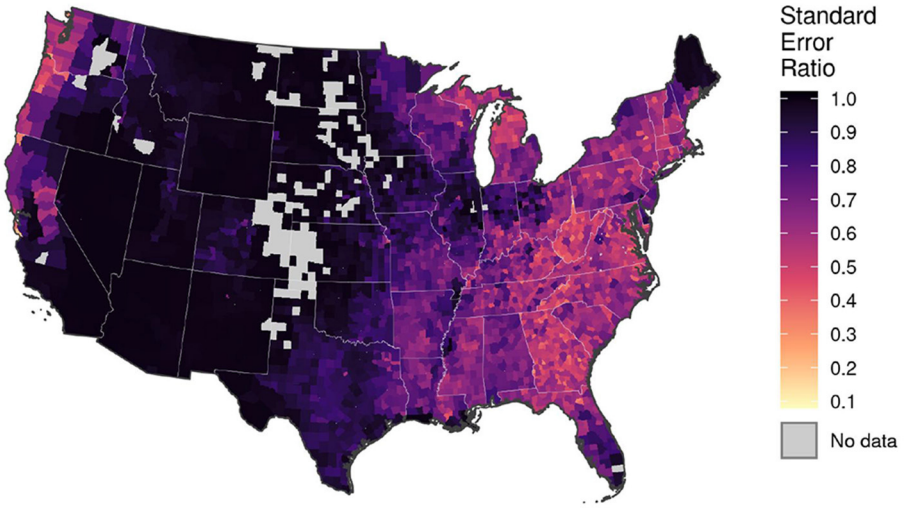


FIGURE 4 | County-level ratios of the standard error of the spatial Fay-Herriot model-based estimator of mean forest carbon density, relative to that of the post-stratified estimator. Ratios <1 indicate the model-based approach yields a more precise estimator of forest carbon stocks than the traditional design-based approach. Gray shaded counties indicate no forested FIA plots were encountered in the county during the most recent current volume inventory, i.e., direct estimator of total forest carbon and associated variance for the county are equal to zero.

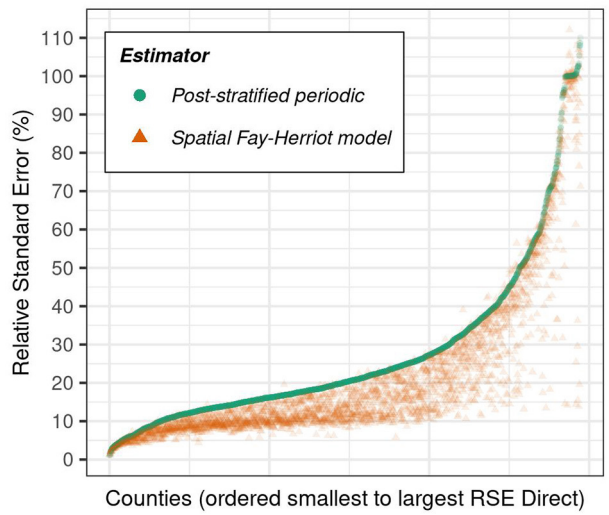


FIGURE 5 | Relative standard error (%) of model-based (i.e., spatial Fay-Herriot model) and post-stratified estimator estimators of mean forest carbon density by county, ordered by increasing relative standard error of the direct estimator.

unbiased (Lohr, 2019), and in the event of model mis-specification, adverse effects on inference can be substantial (Little, 2004). Even among model-based estimators, frequentist and Bayesian inferences yield different interpretation in some cases (see, e.g., Gelman et al., 2004). Therefore, comparing results derived from these different paradigms, presented in subsequent sections, should be received with an understanding about the respective modes of inference. For example, in some cases we compare design-based estimate derived confidence intervals to Bayesian model-based credible intervals. While it can be convincingly argued such comparisons are not appropriate, we present comparative results to explore general patterns in estimates and highlight estimators' qualities.

County-Level Forest Carbon Stocks

Our results indicate the EBLUP derived from the spatial Fay-Herriot model (described in Equations 1–3) offers considerable improvements in precision relative to the post-stratified estimator of county-level forest carbon stocks across much of the CONUS. We present model-based estimates of mean forest carbon density, along with associated estimates of precision, in **Figure 3**. Similarly, we map the spatial distribution of the SER in **Figure 4**. Finally, we illustrate improvements in relative precision offered by the model-based estimator (i.e., measured by the relative standard error), along a gradient of relative precision in the post-stratified estimator, in **Figure 5**.

The spatial Fay-Herriot model yields spatially smooth estimates of county-level forest carbon stocks, that generally reflects the distribution of forestland across the CONUS (**Figure 3**). The largest estimated forest carbon densities are in the coastal Pacific Northwest, Northern Lake States, and Appalachian regions. In contrast, the smallest estimated forest carbon densities appear in the Southwest, Great Basin, and Northern Plains. We show the relative precision of the model-based estimator generally decreases with estimated mean forest carbon density (**Figure 3**; lower precision in counties with low carbon density relative to high carbon density) and with county size (**Figure 5**; lower precision in small counties relative to large counties). Notably, we show the relative precision of the model-based estimator was generally smallest in the

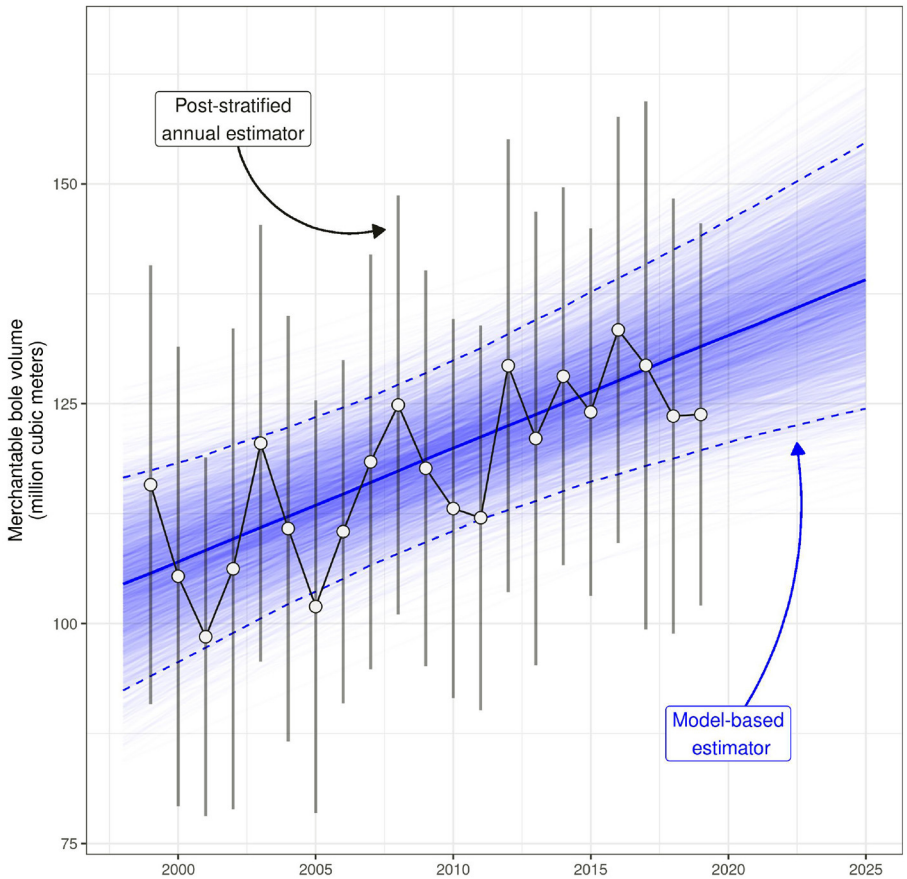


FIGURE 6 | Annual model-based and design-based estimates of total merchantable wood volume (million m^3) on timberland in Washington County, Maine. Model-based point estimates are derived from the posterior median of Hamiltonian Monte Carlo (HMC) samples of parameters presented in Equations (14)–(15), and are represented by the solid, dark blue line. Similarly, model-based interval estimates (i.e., Bayesian 95% credible intervals) are derived from the 2.5 to 97.5% quantiles of HMC samples, and are represented by dashed, dark blue lines. Further, realizations of parameters presented in Equations (14)–(15) from each HMC sample are represented as thin, semi-transparent blue lines. Hence the posterior predictive distribution of the model-based estimator of total merchantable wood volume can be inferred from the relative density of thin blue lines in a given region of the graph (i.e., higher density of lines indicates higher posterior probability). Annual, design-based point estimates are represented by white circles, and are connected by a solid black line. Design-based interval estimates (95% confidence intervals) associated with each annual point estimate are presented as vertical gray bars. All design-based estimates were produced using the annual, post-stratified estimation approach implemented in rFIA (Stanke et al., 2020).

Northern Plains and Southern Lake States regions, likely arising from a combination of small county sizes and relatively low forestland area.

We show the model-based estimator of forest carbon stocks offered the greatest improvements in precision in the coastal Pacific Northwest and eastern US, relative to the post-stratified estimator (**Figure 4**). In these regions, the SER commonly fell below 0.5, indicating the standard error of the model-based estimator was less than half that of the post-stratified estimator for a given county. Across the Interior West, in contrast, we show the model-based estimator rarely improved precision by more than 10% (i.e., SER commonly exceeded 0.9). Further, results presented in **Figure 5** indicate the model-based estimator generally offered consistent improvements in relative precision over the post-stratified estimator, regardless of the absolute magnitude of the post-stratified estimator's relative precision.

Trends in Merchantable Wood Volume in Washington County

Our results indicate the model-based estimator of total merchantable wood volume in Washington County, Maine (approach described in Equations 7–15) offers substantial improvements in precision relative to the post-stratified estimator (**Figures 6**, **7**). Specifically, we show 95% credible intervals associated with model-based point estimates are consistently narrower than 95% confidence intervals associated with the post-stratified estimator (**Figure 6**). On average over the period 1999–2019, the relative standard error of the model-based estimator was 55.9% lower than that of the post-stratified estimator (ranging from 48.9 to 62.4% lower across all years; **Figure 7**), indicating the model-based estimator is more than twice as precise as the post-stratified estimator for our domain of interest. Further, consistent alignment of post-stratified and

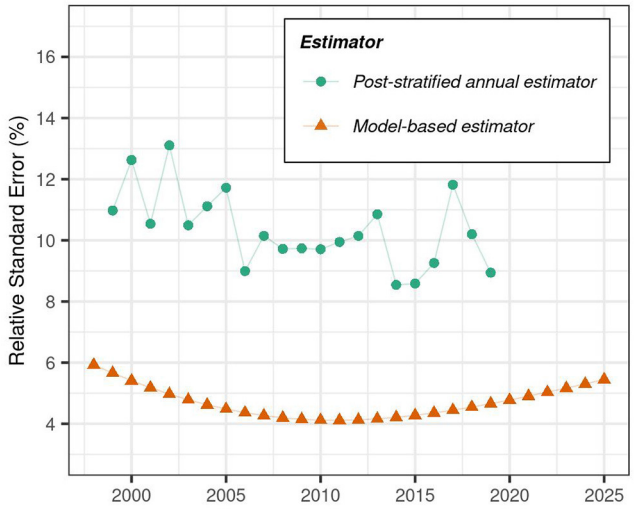


FIGURE 7 | Relative standard error (%) of model-based and post-stratified estimators of trends total merchantable wood volume on timberland in Washington County, Maine.

model-based point estimates suggests the model-based estimator is generally unbiased for the domain of interest (**Figure 6**).

Both approaches indicate that total merchantable wood volume in Washington County has increased considerably over the period 1999–2019 (**Figure 6**). Notably however, the model-based approach yields a smooth, linear trend in total merchantable volume. The post-stratified estimator, in contrast, exhibits large inter-annual variability (±5–10% per year, arising from sampling) and pronounced cyclical patterns over the same period (arising from remeasurement of annual panels). Further, the model-based estimator offers an intuitive approach to characterize the magnitude, direction, and statistical significance of temporal trends in our target variable—a feature the post-stratified estimator lacks (absent estimating change from remeasured plots). Specifically, the posterior distribution of the adjusted population-level regression coefficient, $\hat{\beta}$, yields an estimator of average annual change in total merchantable wood volume across our domain of interest. The posterior median of $\hat{\beta}$ was 1,293,900 $m^3 \cdot yr^{-1}$ (95% credible interval: 588,000–1,989,000 $m^3 \cdot yr^{-1}$), indicating a relatively rapid increase in total merchantable wood volume over the last two decades. Further, we show the probability that $\hat{\beta}$ exceeds 0 is > 0.999, indicating very high certainty in the observed upward trend.

Finally, the model-based approach offers the ability to forecast changes in our variable of interest, along with associated estimates of uncertainty. We highlight this unique capacity in **Figures 6**, **7** by predicting total merchantable wood volume, along with estimates of relative precision, over the period 2020–2025—years for which no FIA data has yet been collected/released for our target population. By the year 2025, we estimate, with 95% probability, that total merchantable wood volume on timberland in Washington County, Maine will range between 124.4 and 154.7 m^3.

DISCUSSION

The FIA program operates the largest network of permanent forest inventory plots in the world, making it well suited to provide critical information on US forests over large geographic and temporal domains (e.g., periodic, state-level estimates). However, the program has experienced increased demand for estimates of forest variables for smaller spatial and temporal domains than traditional sample-based estimation approaches can deliver. Providing such estimates without additional investments in field sampling requires adopting alternative estimation approaches. Here, we presented two case studies that demonstrated some aspects of rFIA's potential to simplify application of SAE to data collected by the FIA program, and thus accelerate adoption of such techniques by FIA data users.

First, we estimate contemporary county-level forest carbon stocks across the CONUS using a domain-level spatial Fay-Herriot model (**Figure 3**), and show the model-based approach offers considerable gains in precision across the predominately forested regions of the CONUS (**Figure 4**). Previous efforts have applied spatial Fay-Herriot models to FIA data to improve precision of estimators of forest density variables (Goerndt et al., 2011), private landowner characteristics (Ver Planck et al., 2017), and forestland removals (Coulston et al., 2021). Domain-level models are particularly useful when inventory plot locations are unknown or measured imperfectly, as spatial auxiliary data need not be associated with plot locations, but rather with domains (Rao and Molina, 2015; Mauro et al., 2017). That is, spatial predictors can be used in domain-level models without requiring the release of actual FIA plot locations. We provide all code and data used to develop the domain-level model presented herein in **Appendix A**, on GitHub (https://github.com/hunter-stanke/FGC_rFIA_SAE) and at our official website (https://rfia.netlify.app). Our procedures can be easily adapted for use with alternative target variables, spatial regions, and/or auxiliary data, and we encourage interested users to adapt our code for use in their own applications of domain-level SAE models.

Second, we follow the approach presented in Little (2004) to develop a temporally-explicit unit-level estimator of multi-decadal trends in merchantable wood volume in Washington County, Maine, using a Bayesian multi-level model. We show the model-based approach offered substantial improvements in precision of annual estimates, relative to the traditional, post-stratified approach (**Figures 6**, **7**). Further, we show the model-based estimator offers an intuitive approach to characterizing the magnitude, direction, and statistical significance of temporal trends, and allows predictions of the target variable to be made for unobserved domains, with associated uncertainty (e.g., forecast change). Unit-level SAE models have been widely applied to FIA data in recent decades (Ohmann and Gregory, 2002; Goerndt et al., 2011; McRoberts et al., 2017; Babcock et al., 2018), and frequently draw from remotely-sensed auxiliary variables to support domain estimation. However, extending the approach presented herein to incorporate spatial auxiliary data will present challenges for most users of FIA data, as neither the true locations of inventory plots, nor the spatial boundaries of strata used for post-stratification are available in the public version of

the FIA Database. Nevertheless, the unit-level model presented can be easily adapted for applications involving alternative populations of interest, and might be useful in the detection and characterization of long-term change in forest ecosystems. Further, such models can be used to characterize the status and change in forest variables at spatial and/or temporal domains that are not currently possible using sample-based approaches (e.g., stand-level estimates).

Role of rFIA in Accelerating the Adoption of SAE Techniques for FIA Data

We posit that rFIA has the potential to simplify the application of model-based SAE techniques to FIA data in three key ways. First, rFIA implements standard, periodic post-stratified estimators—consistent with the estimators implemented by FIA's popular online estimation tool, EVALIDator (Miles, 2021)—within highly flexible, user-defined domains. These direct estimators, along with their associated variances, form the basis for construction of domain-level estimators, as demonstrated by our spatial Fay-Herriot model (Fay and Herriot, 1979; Petrucci and Salvati, 2006; Pratesi and Salvati, 2008) of county-level forest carbon stocks. Second, rFIA implements post-stratified estimators for individual annual panels, offering increased temporal specificity over standard periodic estimation approaches (i.e., the temporally-indifferent estimator), and supporting the development of small area estimators that require direct annual estimates of forest variables at aggregate scales. Examples of such temporally-explicit, domain-level estimators include mixed-estimators (Van Deusen, 1999) and the spatial-temporal Fay-Herriot model (Marhuenda et al., 2013). Finally, rFIA allows summaries of forest variables to be returned for individual response units (i.e., plot-level) and provides utility functions for extracting design information relevant to particular inventory cycles (e.g., stratum assignments and weights). Together, these data can be used to construct a wide variety of unit-level estimators that acknowledge features of FIA's survey design, as demonstrated in our multi-level model of trends in merchantable wood volume in Washington County, Maine.

Adoption of SAE methods by FIA data users (particularly new users) is limited more by FIA's complex data structure and survey design than by the availability of tools that implement SAE methods. Thus, we argue the primary benefit of rFIA in accelerating SAE method adoption is its ability to simplify the process of summarizing and formatting FIA data to serve as input to a wide variety of SAE models. There is a large suite of existing, open-source tools that provide generalized implementations of many domain-level and unit-level SAE models. For example, the sae R package (Molina and Marhuenda, 2015) is specifically designed to implement domain-level SAE models, and we draw from this functionality to develop the domain-level model of forest carbon stocks presented herein. Our intention is not to duplicate efforts of others by implementing common SAE models natively in rFIA, but rather to reduce barriers to the application of such SAE models to FIA data that arise from the complexity of FIA's data structure and sampling design.

Future Extensions of rFIA

Current efforts to extend rFIA include the implementation of a suite of model-based time-series estimators that aim to improve the precision of annual estimates of forest variables, thereby increasing the relevance of FIA data for change detection, characterization, and attribution. Specifically, we aim to provide an intuitive implementation of Van Deusen's mixed-estimator (Van Deusen, 1999), which was recently shown by Hou et al. (2021) to offer considerable improvements in the precision of annual FIA-based forest land area estimates, at both the state- and county-levels. Further, we aim to provide an alternative Bayesian estimator of annual trends in forest variables based on a measurement error model (e.g., similar to Bayesian meta-analysis; Sutton and Abrams, 2001). Notably, both approaches effectively smooth annual, post-stratified estimates of forest variables, and hence are compatible with FIA's existing survey design and database structure.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

HS and AF designed the research. HS performed the research and analyzed the data. HS, AF, and GD wrote and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by National Science Foundation grants DMS-1916395, EF-1253225, and EF-1241874; USDA Forest Service, Region 9, Forest Health Protection, Northern Research Station; US National Park Service; and Michigan State University AgBioResearch.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/ffgc.2022.745874/full#supplementary-material

REFERENCES

Babcock, C., Finley, A. O., Andersen, H.-E., Pattison, R., Cook, B. D., Morton, D. C., et al. (2018). Geostatistical estimation of forest biomass in interior Alaska combining landsat-derived tree cover, sampled airborne Lidar and field observations. *Remote Sensing Environ.* 212, 212–230. doi: 10.1016/j.rse.2018.04.044

Bechtold, W. A., and Patterson, P. L. (2005). *The Enhanced Forest Inventory and Analysis PROGRAM-National Sampling Design and Estimation Procedures*. Gen. Tech. Rep. SRS-80. US Department of Agriculture, Forest Service, Southern Research Station, Asheville, NC, 85.

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

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *J. Stat. Softw*. 80, 1–28. doi: 10.18637/jss.v080.i01

Burrill, E. A., DiTommaso, A. M., Turner, J. A., Pugh, S. A., Menlove, J., Christiansen, G., et al. (2021). *The Forest Inventory and Analysis Database: Database Description and User Guide Version 9.0 for Phase 2*. U.S. Department of Agriculture, Forest Service. Available online at: http://www.fia.fs.fed.us/library/database-documentation/

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: a probabilistic programming language. *J. Stat. Softw*. 76, 1–32. doi: 10.18637/jss.v076.i01

Coulston, J. W., Green, P. C., Radtke, P. J., Prisley, S. P., Brooks, E. B., Thomas, V. A., et al. (2021). Enhancing the precision of broad-scale forestland removals estimates with small area estimation techniques. *Forestry* 94, 427–441. doi: 10.1093/forestry/cpaa045

Domke, G. M., Walters, B. F., Smith, J. E., and Woodall, C. W. (2022). "FIA carbon attributes," in *Sampling and Estimation Documentation for the Enhanced Forest Inventory and Analysis Program: 2020*, eds J. A. Westfall, J. W. Coulston, G. G. Moisen, and H. Erik-Andersen (Madison, WI: U.S Department of Agriculture, Forest Service).

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, 269–277. doi: 10.1080/01621459.1979.10482505

FIA DataMart (2021). *Forest Inventory and Analysis Database*. St. Paul, MN: FIA DataMart.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal*. 1, 515–534. doi: 10.1214/06-BA117A

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis, 2nd Edn*. New York, NY: Chapman and Hall; CRC Press. doi: 10.1201/9780429258480

Goerndt, M. E., Monleon, V. J., and Temesgen, H. (2011). A comparison of small-area estimation techniques to estimate selected stand attributes using Lidar-derived auxiliary variables. *Can. J. For. Res*. 41, 1189–1201. doi: 10.1139/x11-033

Hou, Z., Domke, G. M., Russell, M. B., Coulston, J. W., Nelson, M. D., Xu, Q., et al. (2021). Updating annual state-and county-level forest inventory estimates with data assimilation and FIA data. *For. Ecol. Manage*. 483, 118777. doi: 10.1016/j.foreco.2020.118777

Köhl, M., Magnussen, S., and Marchetti, M. (2006). *Sampling Methods, Remote Sensing and GIS Multiresource Forest Inventory*. Berlin: Springer. doi: 10.1007/978-3-540-32572-7

Lister, A. J., Andersen, H., Frescino, T., Gatziolis, D., Healey, S., Heath, L. S., et al. (2020). Use of remote sensing data to improve the efficiency of national forest inventories: a case study from the United States National Forest Inventory. *Forests* 11, 1364. doi: 10.3390/f11121364

Little, R. J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *J. Am. Stat. Assoc*. 99, 546–556. doi: 10.1198/016214504000000467

Lohr, S. L. (2019). *Sampling: Design and Analysis*. New York, NY: Chapman and Hall; CRC Press. doi: 10.1201/9780429296284

Marhuenda, Y., Molina, I., and Morales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Comput. Stat. Data Anal*. 58, 308–325. doi: 10.1016/j.csda.2012.09.002

Mauro, F., Monleon, V. J., Temesgen, H., and Ford, K. R. (2017). Analysis of area level and unit level models for small area estimation in forest inventories assisted with Lidar auxiliary information. *PLoS ONE* 12, e0189401. doi: 10.1371/journal.pone.0189401

McRoberts, R. E., Chen, Q., and Walters, B. F. (2017). Multivariate inference for forest inventories using auxiliary airborne laser scanning data. *For. Ecol. Manage*. 401, 295–303. doi: 10.1016/j.foreco.2017.07.017

Miles, P. D. (2021). Forest inventory EVALIDator web-application version 1.8.0.1. Saint Paul, MN.

Molina, I., and Marhuenda, Y. (2015). sae: an R package for small area estimation. *R J*. 7, 81. doi: 10.32614/RJ-2015-007

Ohmann, J. L., and Gregory, M. J. (2002). Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, USA. *Can. J. For. Res*. 32, 725–741. doi: 10.1139/x02-011

Petrucci, A., and Salvati, N. (2006). Small area estimation for spatial correlation in watershed erosion assessment. *J. Agric. Biol. Environ. Stat*. 11, 169–182. doi: 10.1198/108571106X110531

Pratesi, M., and Salvati, N. (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Stat. Methods Appl*. 17, 113–141. doi: 10.1007/s10260-007-0061-9

PRISM Climate Group (2010). *PRISM Climate Group*.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rao, J. N., and Molina, I. (2015). *Small Area Estimation*. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9781118735855

Schroeder, T. A., Healey, S. P., Moisen, G. G., Frescino, T. S., Cohen, W. B., Huang, C., et al. (2014). Improving estimates of forest disturbance by combining observations from Landsat time series with US Forest Service Forest Inventory and Analysis data. *Remote Sensing Environ*. 154, 61–73. doi: 10.1016/j.rse.2014.08.005

Shaw, J. D. (2008). Benefits of a strategic national forest inventory to science and society: the USDA Forest Service Forest Inventory and Analysis program. *Iforest-Biogeosci. For*. 1, 81. doi: 10.3832/ifor0345-0010081

Singh, B. B., Shukla, G. K., and Kundu, D. (2005). Spatio-temporal models in small area estimation. *Survey Methodol*. 31, 183.

Smith, W. B. (2002). Forest inventory and analysis: a national inventory and monitoring program. *Environ. Pollut*. 116, S233–S242. doi: 10.1016/S0269-7491(01)00255-X

Stanke, H., Finley, A. O., Weed, A. S., Walters, B. F., and Domke, G. M. (2020). rFIA: An R package for estimation of forest attributes with the US Forest Inventory and Analysis database. *Environ. Model. Softw*. 127, 104664. doi: 10.1016/j.envsoft.2020.104664

Sutton, A. J., and Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Stat. Methods Med. Res*. 10, 277–303. doi: 10.1177/096228020101000404

Tinkham, W. T., Mahoney, P. R., Hudak, A. T., Domke, G. M., Falkowski, M. J., et al. (2018). Applications of the United States Forest Inventory and Analysis dataset: a review and future directions. *Can. J. For. Res*. 48, 1251–1268. doi: 10.1139/cjfr-2018-0196

Van Deusen, P. C. (1999). Modeling trends with annual survey data. *Can. J. For. Res*. 29, 1824–1828. doi: 10.1139/x99-142

Ver Planck, N. R., Finley, A. O., and Huff, E. S. (2017). Hierarchical Bayesian models for small area estimation of county-level private forest landowner population. *Can. J. For. Res*. 47, 1577–1589. doi: 10.1139/cjfr-2017-0154

Westfall, J. A., Patterson, P. L., and Coulston, J. W. (2011). Post-stratified estimation: within-strata and total sample size recommendations. *Can. J. For. Res*. 41, 1130–1139. doi: 10.1139/x11-031

Wurtzebach, Z., DeRose, R. J., Bush, R. R., Goeking, S. A., Healey, S., Menlove, J., et al. (2020). Supporting national forest system planning with forest inventory and analysis data. *J. For*. 118, 289–306. doi: 10.1093/jofore/fvz061

Conflict of Interest: The 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Forests and Global Change

ORIGINAL RESEARCH
published: 26 April 2022
doi: 10.3389/ffgc.2022.769917

Increased Precision in County-Level Volume Estimates in the United States National Forest Inventory With Area-Level Small Area Estimation

Qianqian Cao[1], Garret T. Dettmann[1], Philip J. Radtke[1], John W. Coulston[2], Jill Derwin[1], Valerie A. Thomas[1], Harold E. Burkhart[1] and Randolph H. Wynne[1]*

[1] Department of Forest Resources and Environmental Conservation, Virginia Tech, Blacksburg, VA, United States, [2] Southern Research Station, Forest Service, United States Forest Service, Asheville, NC, United States

Many National Forest Inventory (NFI) stakeholders would benefit from accurate estimates at finer geographic scales than most currently implemented in operational estimates using NFI sample data. In the past decade small area estimation techniques have been shown to increase precision in forest inventory estimates by combining field observations and remote-sensing. We sought to demonstrate the potential for improving the precision of forest inventory growing stock volume estimates for counties in United States of North Carolina, Tennessee, and Virginia, by pairing canopy height models from digital aerial photogrammetry (DAP) and field plot data from the United States NFI. Area-level Fay-Herriot estimators were used to avoid the need for precise (GPS) coordinates of field plots. Reductions in standard errors averaging 30% for North Carolina county estimates were observed, with 19% average reductions in standard errors in both Tennessee and Virginia. Accounting for spatial autocorrelation among adjacent counties provided further gains in precision when the three states were treated as a single forest land population; however, analyses conducted one state at a time showed that good results could be achieved without accounting for spatial autocorrelation. Apparent gains in sample sizes ranged from about 65% in Virginia to 128% in North Carolina, compared to the current number of inventory plots. Results should allow for determining whether acquisition of statewide DAP would be cost-effective as a means for increasing the accuracy of county-level forest volume estimates in the United States NFI.

Keywords: spatial Fay-Herriot models, model-assisted analysis, model-based estimation, composite estimators, forest inventory

OPEN ACCESS

Edited by:
Isabel Cañellas, Centro de Investigación Forestal (INIA), Spain

Reviewed by:
Fernando Montes, Instituto Nacional de Investigación y Tecnología Agroalimentaria (INIA), Spain
César Pérez Cruzado, University of Santiago de Compostela, Spain

****Correspondence:***
Philip J. Radtke
pradtke@vt.edu

Specialty section:
This article was submitted to Forest Management, a section of the journal Frontiers in Forests and Global Change

Received: *02 September 2021*
Accepted: *30 March 2022*
Published: *26 April 2022*

Citation:
Cao Q, Dettmann GT, Radtke PJ, Coulston JW, Derwin J, Thomas VA, Burkhart HE and Wynne RH (2022) Increased Precision in County-Level Volume Estimates in the United States National Forest Inventory With Area-Level Small Area Estimation. Front. For. Glob. Change 5:769917. doi: 10.3389/ffgc.2022.769917

INTRODUCTION

National Forest Inventories (NFI) are designed to produce estimates of forest attributes on regional to national scales; however, many stakeholders would benefit from accurate estimates at finer geographic scales using NFI sample data in a cost-effective manner (Reams et al., 1999; Coulston et al., 2021). Even though the spatial extent of NFI surveys can be very large, sampling intensities may be insufficient to reliably estimate attributes on small areas carved out from what are often expansive target forest populations (Breidenbach and Astrup, 2012). As an example, in the United States NFI, coordinated through the United States Department of Agriculture, Forest

Service, Forest Inventory and Analysis (FIA) program, sampling is conducted on a network of field plots with an average intensity of one plot per 2,430 ha (6,000 acres) and a remeasurement interval of 5 years (Bechtold and Patterson, 2005). The FIA survey design provides sufficient sampling intensity to meet precision requirements specified as ±5% (one standard error), in terms of growing stock volume per billion cubic feet (28.3 million m^3) on commercial forest land (USDA Forest Service National Headquarters, 2008). For estimates on large areas of forestland, this standard can be readily met; however, individual county subdivisions within states rarely contain timber volumes this large. As such, the 5% precision standard for individual county forest volume estimates is generally unattainable from the survey as designed.

Survey designers have recognized this limitation for decades, even to the degree of anticipating the example presented above, as did W.A. Fuller in the following comment made over two decades ago,

"the client will always require more than is specified at the design stage. For example, the client will explain that they require estimates only at the regional or national level and then, when data are available, ask for county estimates," (Fuller, 1999).

Estimating forest attributes for areas smaller than a single state is typically done by making direct estimates from plots sampled within an area of interest; however, it is difficult to obtain acceptably reliable direct estimates within relatively small areas that may contain few inventory plots, or when precise coordinates of a sample unit may be unavailable to analysts due to regulatory restrictions (Reich and Aguirre-Bravo, 2009; Mauro et al., 2016; Magnussen et al., 2017). In solving the problem of insufficient sample sizes, an inventory supported by auxiliary information is an effective approach to predicting forest attributes at unsampled locations. Remote sensing data sets often serve as sources of auxiliary information (McRoberts, 2012; Brosofske et al., 2014). Statistical methods, such as regression or generalized regression, imputation, interpolation, and machine-learning algorithms have been used to link remote sensing data to field observations from NFI samples.

As a class of statistical estimators, small area estimation (SAE) techniques combine survey sample data with auxiliary information – often statistically-related to sample attributes of interest – that will improve the precision of direct estimates. SAE methods are often categorized as (1) domain-direct estimation, (2) domain-indirect estimation, and (3) composite estimators (Rao and Molina, 2015). In domain-direct estimation, parameters for an area or domain of interest are estimated primarily from sample data observed inside that domain. Domain-indirect, estimation also makes use of sample information from outside a domain of interest, with a goal of reducing the standard error of the estimate within the small area domain. Indirect estimators borrow strength from observed sample data (y) outside the domain of interest by linking them to auxiliary data ($\boldsymbol{x}$) using a model $y \sim \boldsymbol{x}$, sometimes called a synthetic estimator or shrinkage estimator, to increase the precision of parameter estimates (Lehtonen and Veijanen, 2009). While direct estimators are often unbiased based on their sampling design, they may become unstable or subject to very large variances when sample sizes are small. Further, while indirect estimators, such as regression models, may be capable of generating precise predictions, they may be subject to large biases when model assumptions are violated. To preserve the unbiasedness of direct estimators, while achieving greater precision afforded by indirect (model-based) estimators, the two can be combined in a weighted average. The resulting composite estimator can be optimized to balance the unbiased property of their direct component and the minimum variance property of the synthetic model. In two common approaches, the optimization is achieved using a mixed modeling framework that accounts for both the variation of sample estimates within each small area domain and the variation among domains not explained by the model (Fay and Herriot, 1979; Battese et al., 1988).

Depending on the structural resolution of data available for developing synthetic estimators, SAE models are commonly distinguished as taking either an area-level or unit-level approach (Rao and Molina, 2015). Area-level models operate by treating each small area domain in a population as a single datum ($\boldsymbol{x}$, y) to be used for fitting a synthetic model. The models are then useful for generating estimates on small area domains within the same population. The domain-direct estimate and its sample variance serve as the source of direct information, while the indirect information – sample estimates and variances from other domains – is then linked to the direct domain *via* the model $y \sim \boldsymbol{x}$, which leverages the relationship between sample and auxiliary data sets. Area-based approaches are often synonymous with the Fay and Herriot (1979) estimator, a well-recognized and widely adopted model used in area-based SAE, including a number of forest inventory applications (Green et al., 2020b).

Like area-based SAE, the unit-level approach also aims to make composite estimates for small area domains; however, data used to formulate the model relationship involve the population observational units themselves, which, in NFI applications are usually the field plot observations that comprise a sampling frame (Battese et al., 1988; Breidenbach et al., 2018; Mauro et al., 2019). While area-level SAE requires that sample units can be explicitly tied to the domains on which they were sampled, unit-level approaches require that each sample unit is paired with corresponding data from the auxiliary source (Rao and Molina, 2015). Unit-level analyses that pair field sample observations with geospatial auxiliary information or digital maps require precise coordinates of field sample plots, most often obtained using global navigation satellite systems, e.g., GPS, to facilitate geospatial pairing of field plot data with co-occurring observations from remote sensing data layers. In the forest inventory literature, both area-level and unit-level SAE have been demonstrated to improve the precision of estimates on small area domains while largely preserving their unbiasedness (Wang et al., 2011; Breidenbach and Astrup, 2012; Goerndt et al., 2013; Magnussen et al., 2017; Mauro et al., 2017; Green et al., 2020a).

Our focus here is the Fay-Herriot (FH) area-level approach to demonstrate an application of SAE that can be used with publicly available observations from the FIA database, which do not include precise coordinates for field sample plots that would otherwise be required in unit-level analyses (Goerndt et al., 2011). The application builds on the situation exemplified by Fuller (1999), where our objective was to make use of sample

data from several eastern states surveyed in the national FIA inventory to produce county-level estimates having enhanced precision over what can be achieved by direct estimation alone. Working with the publicly available FIA data for the three states—North Carolina, Tennessee, and Virginia—the FH approach was chosen for SAE. Magnussen et al. (2017) found that accounting for spatial autocorrelations in an area-level model augmented for that purpose increased SAE precision over non-spatial FH. We also examined the potential for adopting the spatial Fay-Herriot (SFH) approach to further increase precision of area-level SAE (Petrucci and Salvati, 2006). A somewhat novel aspect of the work involves the source of the auxiliary information, namely, digital canopy height models (CHM) derived from 3D digital aerial photogrammetry (DAP) acquired for entire states through the USDA National Agriculture Imagery Program (NAIP; Strunk et al., 2020). To demonstrate the potential for using area-level SAE techniques in estimating county-level total forest volume from FIA in single or multiple state settings, we identified the following four research questions to be addressed: (1) To what degree does area-level SAE using NAIP photogrammetrically-derived CHMs improve the precision of FIA direct county-level estimates of forest volume in the three states; (2) Does accounting for spatial correlation among neighboring counties provide additional gains in precision; (3) What effective gains in sampling intensity can be achieved using CHM auxiliary information in the FH modeling framework; and (4) To what degree are SAE results dependent on treating the states as distinct populations, compared to the alternative of treating them as a single multistate population for purpose of estimating county-level volumes?

MATERIALS AND METHODS

Study Area and Forest Inventory and Analysis Data

The study area included states of North Carolina (NC), Tennessee (TN), and Virginia (VA) in the southeastern United States (**Table 1**). States were chosen based on the availability of remote sensing data from USDA NAIP enhanced aerial acquisitions that produced 3d surface point clouds from DAP. The states possess significant forest resources, ranking second (NC), fifth (VA), and eleventh (TN) in total forestland volume of 37 states that lie entirely east of the North American Rocky Mountain Range. Political subdivisions within the states divide them into 95 counties in TN, and 100 counties in each of the other two states, for an average county land area of 1,220 km^2 for the 295 counties in the study area.

Direct estimates of forest volume were obtained using the USDA Forest Service FIA program's database of field plot measurements for the United States' NFI (USDA Forest Service FIA, 2021). Estimates were based on the FIA calendar year 2017 forest evaluation, so that full-panel estimates—those based on complete sample sets of field plot measurements collected over 5 years (2015–2019)—were available for estimation (Bechtold and Patterson, 2005). Sampled plot-level data were processed to produce timber volume estimates and standard errors for each county in the three states. The estimated volumes and their standard errors provided direct estimates of forest inventory to be tested for possible precision gains using SAE techniques. The attribute of interest estimated for each county was the total (net) wood volume in live tree main stems having diameters at breast height ≥12.7 cm (i.e., 5.0 inches). The volume attribute excluded wood contents in stumps below a 30.5 cm (1 foot) height, and topwood above a 10.2 cm (4 inch) upper stem diameter.

TABLE 1 | Forest volume and land area statistics for states in the study area, with the number of forested FIA sample plots (*n*) in each state.

	Net volume		Land area (total)		Forest area %	
State	$ft^3 \times 10^6$	$m^3 \times 10^6$	mi^2	km^2		***n* plots**
North Carolina	43,691	1,237	53,800	139,400	54	3,662
Tennessee	32,072	908	42,100	109,200	52	2,932
Virginia	40,573	1,149	42,800	110,800	62	3,298

Auxiliary Data

Digital data in the form of CHM acquired through USDA NAIP served as source of auxiliary information for the SAE analyses. The data were delivered as digital surface model (DSM) raster files having grid cells of approximately 1 × 1 m (TN) and 5 × 5 m (NC and VA) produced from aerial imagery and 3d DAP methods (Strunk et al., 2020). The DSM data were resampled at 10 m × 10 m resolution for overlay with United States National Elevation Database Digital Elevation Model (DEM) data. CHMs were calculated by subtracting DEM values from the NAIP DSMs, setting any negative values to zero.

Land cover data from the National Land Cover Database were used to remove CHM data for raster cells classified as open water. No accounting was made for other land cover types, as the goal was to use auxiliary information derived as much as possible from the NAIP imagery, rather than depending on land cover or vegetation type classifiers derived from other auxiliary sources. The CHM raster layers were clipped to each county boundary and aggregated into seven 5-m interval height classes spanning, {(0, 5), (5, 10), . . ., (30–35) m}, with values >35 m omitted to remove possible outliers, anomalies, or atmospheric interference, assuming nearly all forest canopies in the study area were ≤35 m in height. County areas covered by each of the seven height classes were calculated as the product of grid cell counts and the cell area in km^2.

Area-Level Fay-Herriot Model

The area-level approach introduced by Fay and Herriot (1979) and implemented in the R package "sae" (Molina and Marhuenda, 2015) was adopted for conducting SAE analysis. The FH approach uses a composite of two estimates that results in empirically best linear unbiased predictions (EBLUP) of an attribute of interest in each spatial domain. Counties were the domains, and total volumes the attributes of interest here, with the population attribute for a given county (d) denoted as (τ_d).

The composite estimator averaged direct and synthetic estimates to produce EBLUPS

$$\widehat{\tau}_d^{EBLUP} = \hat{\gamma}_d \widehat{\tau}_d^{\mathrm{DIR}} + \left(1 - \hat{\gamma}_d\right) X_d^T \widehat{\beta} \quad (1)$$

where $\widehat{\tau}_d^{\mathrm{DIR}}$ is the direct inventory estimate of total volume for the d^{th} county, X_d is a vector of canopy height class areas for the county, $\widehat{\beta}$ is a set of linear regression coefficients estimated from all county-level observations, and the composite weighting factor, i.e., shrinkage $\widehat{\gamma}_d$ is defined as

$$\widehat{\gamma}_d = \frac{\widehat{\sigma}_a^2}{(\widehat{\sigma}_a^2 + \hat{v}(\widehat{\tau}_d^{\mathrm{DIR}}))} \quad (2)$$

where v denotes the direct estimator variance and $\widehat{\sigma}_a^2$ is the estimated variance among the county totals, formulated as mixed effects in the FH model. A basic idea of the FH estimator and composite estimators used in SAE in general is that the weighting factor provides a way to balance the information between the direct $\widehat{\tau}_d^{\mathrm{DIR}}$ and regression-synthetic $X_d^T\widehat{\beta}$ estimators. In (1) the weighting factor (2) accounts for relative sizes of the domain-direct variance $\hat{v}(\widehat{\tau}_d^{\mathrm{DIR}})$ and variance among counties measured by the random effect variance $\widehat{\sigma}_a^2$. Parameter estimation, including mixed-effects coefficients and variances approximated by a polynomial expansion, were estimated using restricted maximum likelihood in the "sae" package. Additional details of the estimation procedure are given by Molina and Marhuenda (2015).

Spatial Fay-Herriot Model

A spatial FH model (SFH) was also used to account for possible spatial correlation among estimates from adjacent pairs of counties (Petrucci and Salvati, 2006). The SFH model is built on a FH model with simultaneously autoregressive spatial correlation structure specified by a single parameter ρ and ($d \times d$) proximity matrix **W** having zero diagonals and ones in off-diagonal elements indexing pairs of adjacent counties, i.e., those that share a common border (Molina and Marhuenda, 2015). The advantage of this spatial model in settings where forest conditions are more alike among adjacent counties than for counties separated by some distance is that it should provide additional precision gains beyond those of the non-spatial FH model.

Model Fitting

In the "sae" package, FH EBLUPS and their mean-squared errors are estimated by functions named eblupFH and mseFH with the response variable specified as the vector of county-level direct estimates of forest volume from FIA sample data, i.e., $\widehat{\tau}_d^{\mathrm{DIR}}$ (10^6 m^3). County-level estimator variances $\hat{v}(\widehat{\tau}_d^{\mathrm{DIR}})$ were given as inputs to the fitting procedure for use in calculating domain weights for use in the composite estimator (1). County-level area coverage (km^2) for each canopy height class were specified as predictors, with the full set of predictors including all seven height classes. A zero intercept was found to be supported for the base fixed-effects model based on increased AIC by >2 units when a global intercept was included, with additive random effects estimated for counties. The same inputs were used in fitting the SFH model, but with the proximity matrix also specified.

Both FH and SFH models were fitted with the response formulated as $\widehat{\mu}_d^{\mathrm{DIR}}$, where μ = county volume per unit land area (m^3 km^{-2}) along with a suitable set of predictors that gave roughly equivalent model results, but on a per area basis rather than for county totals. Scaling the predictors to the same per area basis as μ was necessary to preserve the strength of the relationship between response and predictors. Preliminary analyses indicated no meaningful differences in any of the results or hypotheses tested, including comparison of spatial autocorrelations with the SFH method. As a result, only results for totals are reported here.

As in any multiple regression analysis, the correct specification of predictors was investigated, in part to reduce any effects of co-varying predictors that might inflate the estimated variances of regression coefficients. Using a backward elimination procedure based on the greatest reduction in the Akaike Information Criterion (AIC) predictors were sequentially removed from a reference model having a full set of $p = 7$ predictors, dropping one predictor at a time until AIC was no longer reduced upon removing another predictor. As a further restriction to address a tendency of AIC to select models that are overspecified, we adopted a type-I error rate $\alpha = 0.01$ for model coefficients to be included in final models. Predictors that did not meet this criterion were dropped from final models.

Apparent Sample Size

A simple formulation for the standard error of a population total estimated from sample observations of the total y, assuming random sampling and ignoring any finite population correction, is $\widehat{\sigma}_{\widehat{\tau}} = \widehat{\sigma}_y/\sqrt{n}$. Using this formula and comparing domain-direct estimated standard errors to RMSEs of EBLUPS obtained from a FH or SFH estimation gave the apparent sample size formula

$$n_{app} = n_{for} \times \left[\frac{\widehat{\sigma}_{\widehat{\tau}}^{DIR}}{\widehat{\sigma}_{\widehat{\tau}}^{EBLUP}}\right]^2 \quad (3)$$

where n_{for} is the sample size for the number of forested plots involved in the direct estimate of the total, $\widehat{\sigma}_{\widehat{\tau}}^{\mathrm{DIR}}$ is the standard error of the direct estimate, and $\widehat{\sigma}_{\widehat{\tau}}^{\mathrm{EBLUP}}$ is the RMSE of the EBLUP estimated FH or SHF total. Although FIA volume estimates include observed zeros on non-forested plots, the same formulation as (3) can be used to calculate apparent sample sizes for both forest and non-forest plots. To avoid redundancy, only the results for n_{app} of forested plots are presented here.

Single vs. Multiple State Analyses

In both FH and SFH approaches, composite estimators were first developed using data from NC, TN, and VA fitted separately as "individual states," effectively treating them as distinct populations for purposes of county-level forest volume estimation. Second, estimators were developed using data from all three states in a single model-fitting procedure, i.e, treating the three "combined states" as a single population divided into county domains. Both approaches produced EBLUP estimates for

all counties having at least two FIA sample plots—the minimum needed to estimate the direct estimator variance; however, nothing in the two analyses explicitly constrained the resulting estimates to agree between individual and combined state approaches. Performing separate analyses for each state would lead to three sets of fixed effects coefficients, and three separate random effects variances $\widehat{\sigma}_a^2$, one for each state, compared to just one for the analysis combining the three states. Similarly, the SFH approach gave separate state estimates of ρ for each state, while a single value was estimated in the combined approach.

Evaluation of Spatial Autocorrelation

To compare the FH models to the SFH models we used Likelihood ratio tests (LRT)

$$\text{LRT} = 2(\ln L_1 - \ln L_2)$$

where L_1 and L_2 are the restricted likelihoods calculated for SFH (L_1) and FH (L_2) models, identical except for the correlation parameter ρ. Under the null hypothesis SFH does not improve on the fit of the non-spatial FH model, and large values of the test statistic $LRT \sim \chi^2(1)$ provide evidence for concluding that the SFH model improves on the fit over the FH model.

Estimator Errors

Uncertainty of area-level FH EBLUPs is assessed by mean square error (MSE), calculated in the mseFH R function as an additive combination of terms representing uncertainties associated with (a) prediction of random effects, (b) estimation of regression model coefficients, and (c) estimation of the random-effect variance, i.e., the variance among small-area domains (Prasad and Rao, 1990; Molina and Marhuenda, 2015). The MSE calculation is an second-order approximation using Taylor linearization, shown to be approximately unbiased for large D. Approximate unbiasedness of the Prasad and Rao (1990) MSE estimator holds when any of three estimation techniques are used, including the package default restricted maximum likelihood (REML) used here. A similar approximation for REML estimation based on the findings of Singh et al. (2005) and implemented in the R package function mseSFH was adopted here for calculating MSE for SFH estimates. Both the FH and SFH MSE estimates are known to be asymptotically unbiased, i.e., any bias approaches zero as $D \to \infty$ (cf. Li and Lahiri, 2010; Coulston et al., 2021).

Precision Gains

To evaluate gains in the precision of FH estimators, we used the relative standard errors (RMSE%) for $\widehat{\tau}_d^{\text{DIR}}$ and $\widehat{\tau}_d^{\text{EBLUP}}$ from FIA sample data and FH models, respectively. The RMSE% to be compared for each county were calculated as

$$\text{RMSE\%}_{\text{EBLUP}} = 100 * \sqrt{\text{MSE}(\widehat{\tau}_d^{\text{EBLUP}})} / \widehat{\tau}_d^{\text{DIR}}$$

$$\text{RMSE\%}_{\text{DIR}} = 100 * \sqrt{\text{Var}(\widehat{\tau}_d^{\text{DIR}})} / \widehat{\tau}_d^{\text{DIR}}$$

and a unitless standard error ratio (SER) for each county was calculated as

$$\text{SER} = \text{RMSE\%}_{\text{EBLUP}} / \text{RMSE\%}_{\text{DIR}}$$

RESULTS

Model Fitting

Model fixed effects coefficients values generally increased from lowest to highest with CHM height class, consistent with higher volumes per unit area as would be expected for taller forests (**Table 2**). A notable exception to this increasing trend was the (20, 25] m height class, which contributed less to predicted volumes per km^2 than lower height classes in the same models. The (30, 35] m CHM height class was a significant predictor for all states and models, including in the combined 3-state model, and had the largest coefficient estimates in all the models tested. The Virginia estimator was the only one to include a significant fixed-effects coefficient for the lowest (0, 5] m height class, and its small magnitude reflected the low potential for forest volumes in such low canopy heights, being less than half the size of the next smallest coefficient value estimated for any model or height class (**Table 2**).

Best models for individual state estimators included between 2 and 4 predictors, with variable selection excluding predictors for at least one 5-m height class between any two consecutive height classes kept in the final models (**Table 2**). In the combined three-state spatial model, predictors for two consecutive height bins (15, 20] and (20, 25] were both found to be significant ($\alpha = 0.01$) despite a greater potential for collinearity due to possible overlapping of information in abutting measurement intervals (see **Supplementary Material** for predictor scatterplot/correlation matrices).

Fixed effects for the FH estimators were generally in line with those of the SFH models for the single-state analyses (**Table 2**).

TABLE 2 | Estimated coefficients (10^6 m^3/km^2) for best SAE area-level models ($p < 0.01$ for reported model coefficients).

State	Predictors	Estimated coefficients	
		FH	SFH
North Carolina	CHM (10-15]	0.03786	0.03623
	CHM (20-25]	0.02357	0.02605
	CHM (30-35]	0.08277	0.07889
Tennessee	CHM (15-20]	0.04241	0.04250
	CHM (30-35]	0.10179	0.10090
Virginia	CHM (0-5]	0.00500	0.00465
	CHM (10-15]	0.03360	0.03290
	CHM (20-25]	0.02864	0.03028
	CHM (30-35]	0.07009	0.06837
Three-state region	CHM (5-10]	0.01230	0.01315
	CHM (15-20]	0.04357	0.03381
	CHM (20-25]	*	0.01261
	CHM (30-35]	0.09240	0.07563

**CHM (20-25] was not statistically significant in FH model for three-state region.*

Differences between FH and SFH estimates for any predictor diverged by less than 3% percent on average (8 of 9 single-state FH to SFH pairwise comparisons), although a discrepancy of about 11% in magnitude was noted for the NC (20, 25] m height class predictor. Larger differences were noted between the FH and SFH coefficients in the three-state analysis, with differences greater than 18% observed for two height class parameters and the (20, 25] m height class being significantly different from zero in the SFH model but not in the three-state FH model (**Table 2**). Positive signs on all coefficients suggested that the synthetic models themselves would not produce negative volume estimates, none of which were observed in any of the results produced for the study data.

Spatial vs. Non-spatial Models

Positive spatial autocorrelation estimates for ρ in SFH models (**Table 3**) indicated that geographically adjacent counties' forest volumes tended to vary together in the same direction (Griffith, 2005). While LRT did not support the need for the SFH formulation in individual state FH models, the spatial model was found to significantly improve upon the non-spatial FH model in the combined 3-state analysis (**Table 3**).

Precision of Estimation

Comparison of county-level direct estimates to FH EBLUPS showed no obvious inconsistencies that might point to biases in any sets of estimates (**Figure 1**). Relative RMSE from SAE estimators compared to direct estimate standard errors for county estimates showed magnitudes of reduction in estimator errors (**Figure 2**). Relative errors for two Virginia counties, Hampton City and Newport News, were noticeably higher (>100%) than the bulk of the state's counties (**Figure 2C**). A similarly, high relative error (>80%) was noted for Crockett county, TN (**Figure 2B**). Volume estimates for all three of these counties were below 28,300 m^3 (1 million m^3), the smallest three volume estimates of any counties sampled by at least one FIA forest plot in the study. The two Virginia counties—defined by the United States Census Bureau as "county equivalents," but by the State of Virginia as "independent cities"—had just one FIA forest plot record each, while Crockett County, TN had just two FIA forest plots. No other counties in the study area had fewer than 5 FIA forest sample plots within their boundaries.

Using SER as a measure of the relative reduction in estimator error for each county, North Carolina had the greatest gains in precision at close to a 30% reduction using FH estimators compared to sample domain-direct estimator errors (**Table 4** and **Figure 2A**). Virginia and Tennessee exhibited comparatively modest reductions in estimator errors at about 19% for FH versus direct county estimates (**Figures 2B,C**).

TABLE 3 | Estimated spatial correlation coefficients in SFH models and likelihood ratio tests (LRT) between FH and SFH models.

State	Spatial autocorrelation coefficient	LRT (*P*-value)
North Carolina	0.4627	1.6622 (0.1973)
Tennessee	0.2880	0.7976 (0.3718)
Virginia	0.2204	0.3332 (0.5638)
Three-state region	0.6195	37.8980 (<0.00001)

Apparent Sample Size Gains

Apparent sample sizes for each county were calculated by applying (3) to county domains in the study states (**Figure 3**). Apparent sample size gains for the counties ($n_{gain} = n_{app} - n_{for}$) attributable to the increased efficiencies of FH estimators were summed for each state and the whole study area to determine roughly how many additional FIA field plots would be required in each state to achieve the county-level precisions afforded by FH and SFH estimators (**Table 4**). In comparison to the FIA forested plot sample sizes (**Table 1**), the gains under FH estimation ranged from about 65% in Virginia to about 128% in NC, i.e., more than an apparent doubling of the FIA sample size for forested plots in that state.

DISCUSSION

Reductions in standard errors of direct estimates (volume or biomass) from published studies of area-level SAE vary considerably, from close to zero reduction (SER = 1) for planted pine stands on which relative standard errors of direct estimates were <10% (Green et al., 2020a) to reductions greater than one-half (SER = 0.4) in small natural stands (average 6.1 ha) studied by Ver Planck et al. (2018). Magnussen et al. (2017) achieved considerable error reduction (SER = 0.53) in 25–37 ha management units around Burgos, Spain and Rastatt, Germany, and nearly as good (SER = 0.57) in forest districts and municipalities in Jura canton, Switzerland (70,800 ha) and Vestfold county, Norway (14,900 ha). Breidenbach et al. (2018) attained error reductions (SER = 0.8) similar to what we achieved for TN and VA working with a larger set of domains in Vestfold county than the population studied by Magnussen et al. (2017). While Green et al. (2020a) noted little reduction in standard errors in stands where direct estimates were already quite precise, they showed strong gains (SER = 0.65) in stands having comparatively large uncertainties in direct estimates (up to 30% relative SE). In the states studied here, most counties having direct relative standard errors >30% for total volume did not achieve precision gains as great as their corresponding state averages, i.e., county SERs were not as small as the state average SER (**Figure 2**). However, such counties comprised a small proportion—about 13 of 295—of the counties studied. Although Goerndt et al. (2013) focused more on domain-specific biases than relative gains in precision, their results showed clearly that gains depended a great deal on the attributes of interest being studied, a consideration not pursued here as we only examined FH model performance in estimating total forest volumes.

Several authors have examined the relationship between sample size and precision gains such as those we measured by SER, as well as the effect of direct estimator precision on SER (Magnussen et al., 2014; Green et al., 2020a). By intentionally reducing domain-specific sample sizes (n_d), Goerndt et al. (2011) demonstrated greatest precision gains (SER = 0.7) for

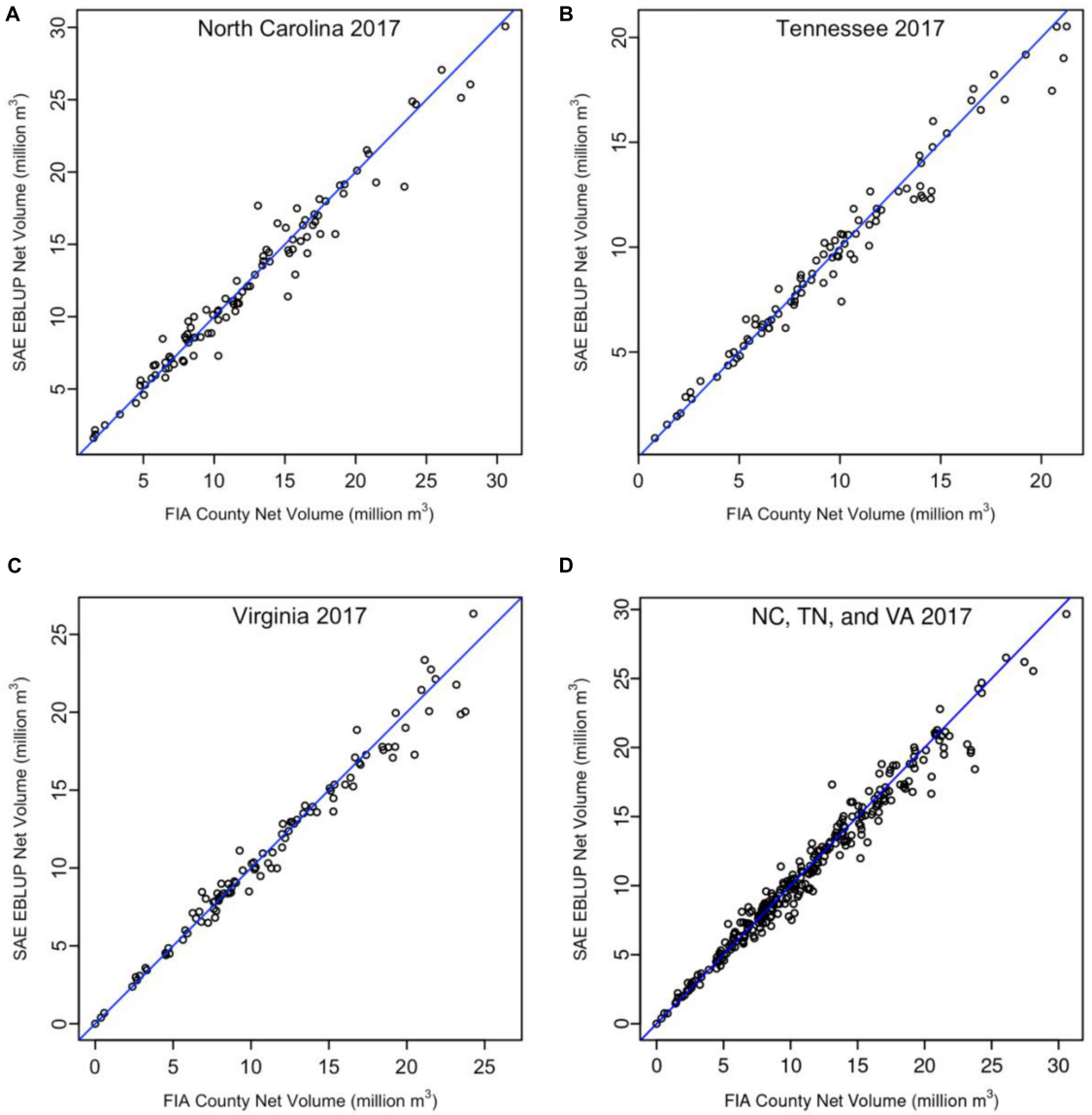


FIGURE 1 | Fay Herriot composite estimates (EBLUP) of county forest volume totals compared to FIA direct estimates for three separate state-level analyses **(A–C)** and the combined 3-state SFH analysis **(D)**.

area-based SAE for $2 \leq n_d \leq 4$. They noted comparatively modest gains (SER $>$ 0.7) for $n_d > 4$, with no appreciable reduction in variances compared to direct estimates for large samples ($n_d > 30$). Area-level SAE over a twenty-state region of the northeastern United States involving counties as small-area domains showed gains comparable to those observed in the three-state region studied here, with $0.65 \leq$ SER ≤ 0.95 (Goerndt et al., 2019). Despite there being many large sample sizes in the three states we studied—median $n_d = 31$ (forested plots) and median $n_d = 50$ (all plots)—SERs for the 295 counties studied here averaged 0.77 (**Figure 2D**) with SER 2.5th and 97.5th percentiles [0.58, 0.94] (**Supplementary Table 1**) showing appreciable gains in precision for FH results across all domain-specific (county) sample sizes. Precision gains observed here across a range of forested plots' sample sizes from $5 \leq n_d \leq 96$ differ from what other authors have shown, possibly because our analysis did not rely on generalized variance functions, which are known to convey mainly information on n_d, rather than directly estimated within-domain variation (Goerndt et al., 2011; Coulston et al., 2021).

While precision gains were notable for estimating total wood volume with FH EBLUPS compared to direct, design-based estimates, a significant concern is that end-users often require volume estimates disaggregated by forest type or species groups (e.g., hardwoods or softwoods), or by product classes such as pulpwood or sawtimber (Coulston et al., 2021). End-users may also wish to preserve consistency among estimates of multiple attributes for which NFI data provides estimates,

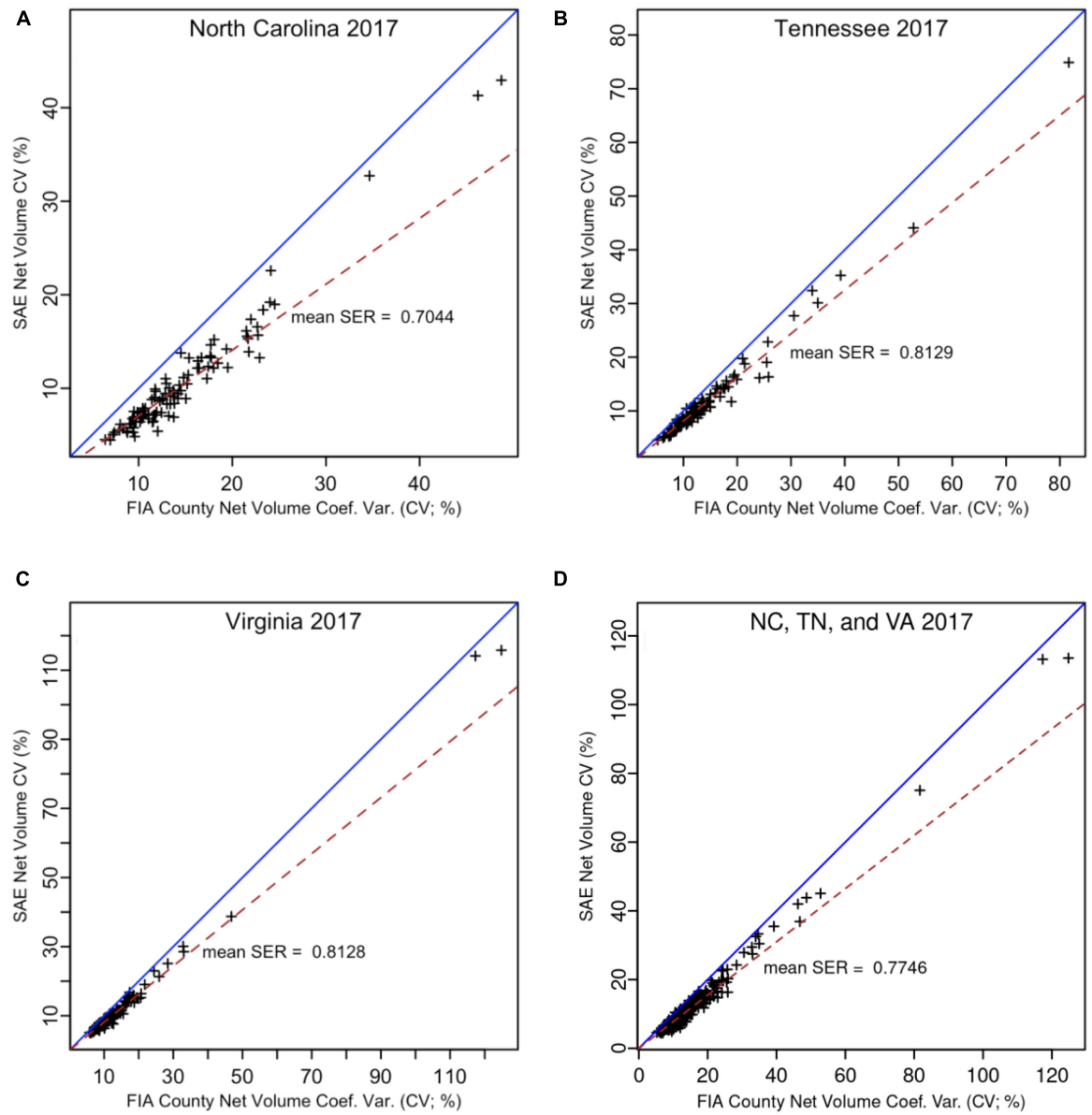


FIGURE 2 | Fay Herriot estimates RMSE% compared to FIA direct estimate standard errors (percent of county total) for three state-level analyses **(A–C)** and the combined 3-state analysis **(D)**. Non-spatial FH estimators were used for data in panels **(A–C)**, while the SFH estimator was used for panel **(D)**. (note: SER = ratio of standard errors).

including biomass, stem density, basal area, or any of dozens to hundreds of other attributes. While not pursued here, multivariate FH models have been developed and applied to situations where multivariate estimates were sought from repeated sampling (Ghosh et al., 1996; Benavent and Morales, 2016). Model-assisted methods may also allow for simultaneous and compatible estimation of multiple attributes, or generic inference, as distinguished from the focus on a single attribute of interest, or specific inference, pursued here (McRoberts et al., 2017; McConville et al., 2020). Expanding the analyses conducted here to multivariate cases would constitute a significant augmentation.

Accounting for spatial correlation among adjacent counties had minimal effect in single-state analyses conducted here and modest effect in the combined model involving all three states, likely due to a mismatch between the scale at which the counties in this study differed from scales of natural and anthropogenic processes affecting total wood volumes in the states and counties studied. We note that the mean land area of 295 counties studied here was 1,218 km^2, with a range of [107, 4,047] km^2 (**Supplementary Table 1**). The lack of spatial correlation differed from other work that found substantial reductions in SER when spatial autocorrelation was accounted for Magnussen et al. (2014); Ver Planck et al. (2018). Observed gains may be

TABLE 4 | Mean standard error ratios (SER) of EBLUP RMSE to direct standard error, averaged over counties within state groupings.

	Single state results		Combined state results		Apparent n_d gains	
State	**FH**	**SFH**	**FH**	**SFH**	**FH**	**SFH†**
North Carolina	0.7044	0.6957	0.783	0.7363	4,692	3,921
Tennessee	0.8129	0.7984	0.8567	0.8166	1,937	1,746
Virginia	0.8128	0.8155	0.817	0.7729	2,137	2,893
Three states	0.7759	0.7694	0.8183	0.7746	8,766	8,560

Unshaded entries correspond to alternative FH or SFH results supported by likelihood ratio tests (cf. ***Table 3****).*
†*results from SFH analysis performed on three-states combined.*

related to the areas of domains studied, such as in relatively small stands studied by Ver Planck et al. (2018), which were necessarily much closer in proximity (e.g., as could be measured by centroid distances) due to their small land areas than the county domains we studied. Closer proximities among domains does not necessarily explain the gains achieved in accounting for spatial correlation noted by Magnussen et al. (2014) in their study of large Swiss forest districts spanning an area of 14,000 km^2. One result from our combined 3-state analysis that agreed with results of Ver Planck et al. (2018) is that accounting for spatial correlation can increase overall precision when averaged across many small area domains. Our results were also consistent with results presented by Ver Planck et al. (2018) that showed while the SFH reduced relative errors in some domains, it led to increased errors in others. When presented as maps of relative standard errors for counties in the study region (**Figures 4**, **5**), it becomes evident that gains in precision for combining states and using SFH reduced the errors in some counties in Virginia while increasing errors in some counties in North Carolina and Tennessee. Such tradeoffs should be considered when choosing an approach for implementation.

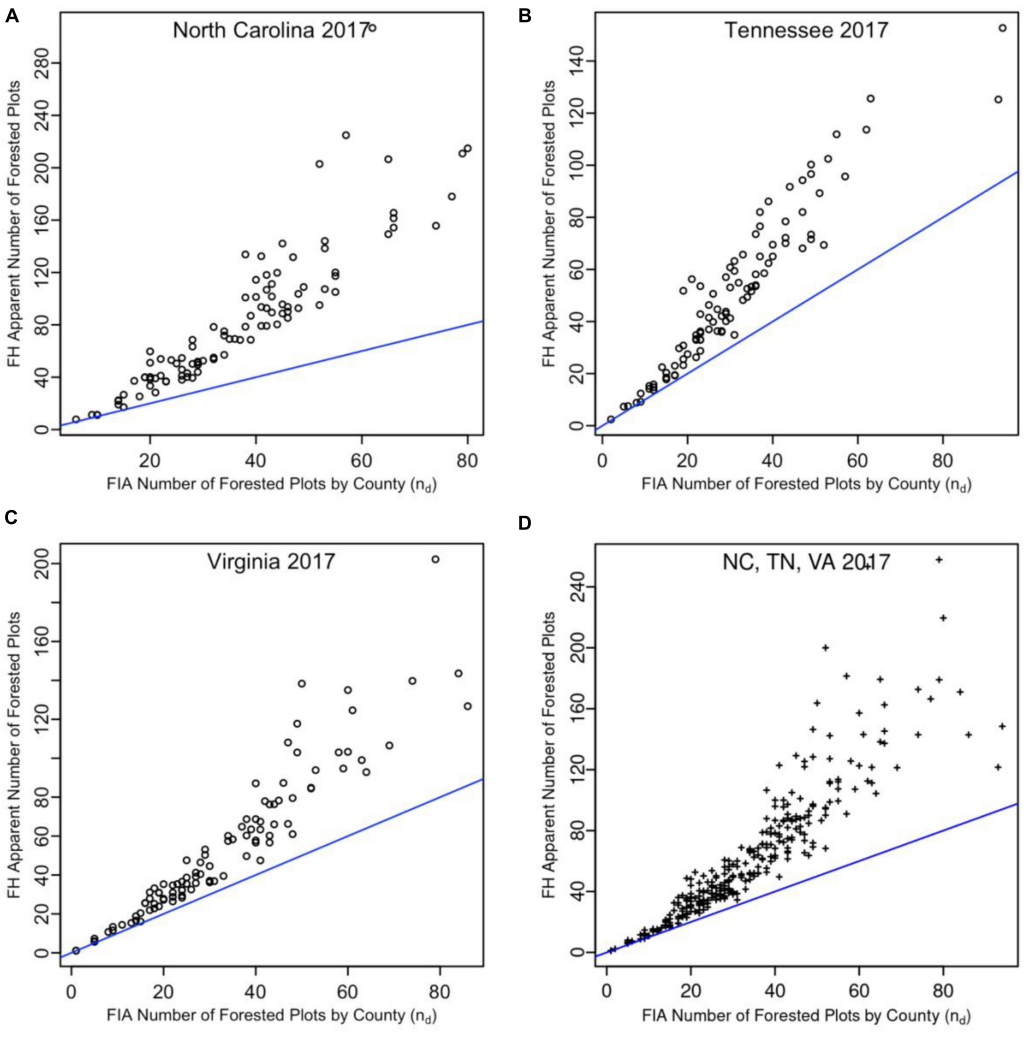


FIGURE 3 | Apparent domain sample sizes ($n_{d,\text{app}}$) for total forest volume of study area counties compared to FIA sample sizes from direct estimates. Non-spatial FH estimators were used for results shown in panel **(A–C)**, while the SFH estimator was used for panel **(D)**.

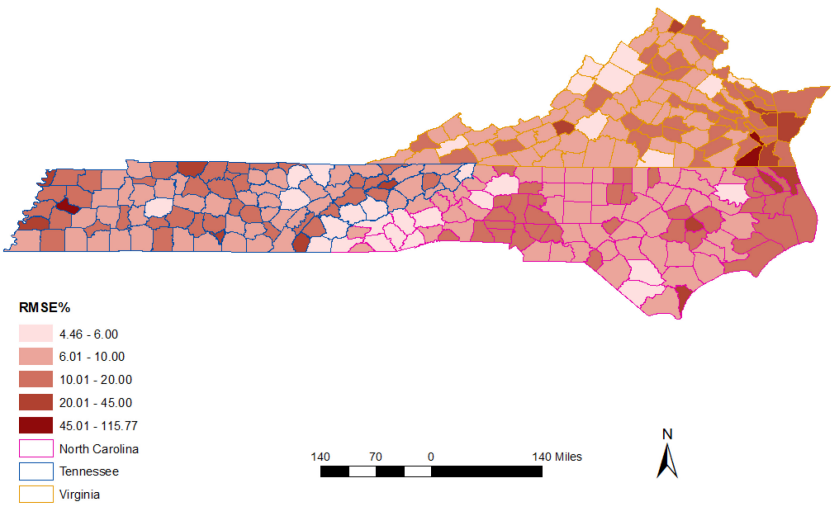


FIGURE 4 | Relative standard errors of non-spatial Fay-Herriot EBLUP estimates of county total forest volume, obtained from separate models for each of three states, North Carolina, Tennessee, and Virginia.

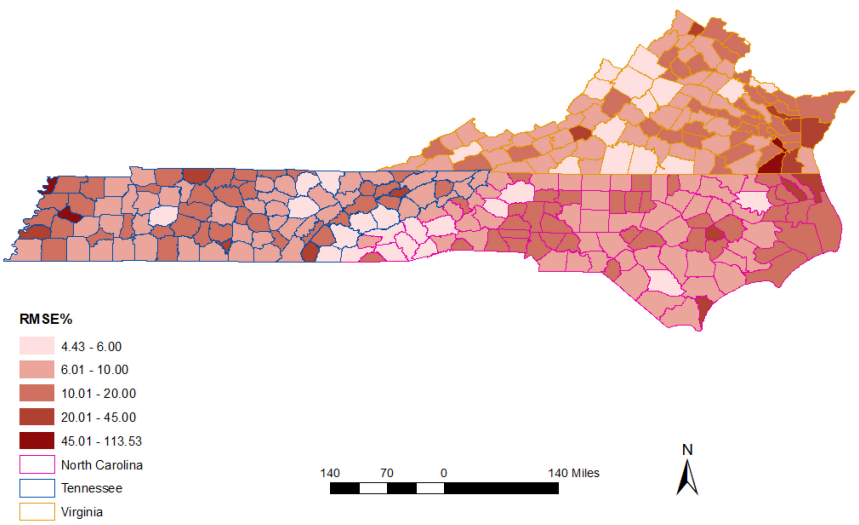


FIGURE 5 | Relative standard errors of simultaneous autoregressive spatial Fay-Herriot EBLUP estimates of county total forest volume obtained from a single model combining data for three states in the southeastern United States.

Since county areas are treated as known, fixed quantities in the FIA estimation framework, scaling of responses, predictors, model coefficients, and estimated RMSEs to a per-area basis can be accomplished *post hoc* by dividing quantities related to totals by the corresponding county areas (**Supplementary Table 1**). No recalculation of model parameter estimates, including the spatial correlation parameters in SFH results is necessary. Future work could look at how results generated per unit of forest land might differ from the area totals (or totals scaled per unit area of all land) presented here. Since forest land area is an estimated quantity in the FIA inventory design, estimates per unit of forest land area would consist of ratio estimates with both numerator (volume) and denominator (forest land area) being estimated quantities. It is possible that the spatial correlation structure for such ratio estimates differs from those noted here.

Further gains in estimator efficiency should be possible by including additional auxiliary information to reduce likely inconsistencies in the largely unmodified NAIP CHM information used here (Hansen et al., 2013; Potapov et al., 2020). Apart from filtering CHMs to omit areas over water, no efforts were made here to eliminate pixels in the CHMs representing heights of vegetation or other structures lying outside of areas that would typically be classified as forestland in the FIA inventory. Features such as buildings or other raised structures, urban forests, small patches of trees growing outside of areas defined as forest (see Glossary of Terms, Bechtold and Patterson, 2005), agricultural crops, or features in other land types undoubtedly contributed to the summed canopy height class metrics derived from CHM metrics used as predictors in the synthetic models developed here (Hansen et al., 2016). Although the current study was focused on gains to be realized with minimal processing of NAIP CHMs as-delivered, exploring the impact of non-forest features on forest volume estimates is a topic of considerable interest (Potapov et al., 2021).

Cost Effectiveness

Results of this work demonstrated the magnitudes of gains in efficiency possible by integrating NAIP-derived CHMs with direct sample data from NFI field plots using the area-level FH and SFH procedures. Such information can be used to evaluate cost-effectiveness of implementing the approach to other states or repeating a NAIP 3d acquisition at a future date, such as in monitoring forest growing stock change over time. In considering this, we defined the costs of collecting sample data from a single survey plot (C_{plot}) and the cost of acquiring NAIP 3d DAP coverage for a state (C_{NAIP}).

The cost of additional forest plots needed in direct estimation to attain the same level of precision that SAE analysis using NAIP CHM data provided is $C_{\text{plot}} \times n_{\text{gain}}$, leading to the following cost effectiveness (C_e) calculation

$$C_e = \frac{C_{\text{NAIP}}}{C_{\text{plot}} \times n_{\text{gain}}}$$

where values of $C_e << 1$ should indicate some degree of cost-effectiveness for acquiring NAIP 3d imagery. In the three states studied here, the inequality $C_e < 1$ requires a cost ratio $C_{\text{NAIP}}/C_{\text{plot}} < n_{\text{gain}}$, with values of n_{gain} ranging from 1,937 to 4,692 (**Table 4**). From this basic calculation, it appears that NAIP acquisition costs below about 2,000 times the cost of adding an additional sample plot measurement would be worth considering for states like North Carolina, Tennessee, or Virginia, at least from the perspective of cost-effectiveness. In actual settings, cost considerations would not likely be this simple. For instance, when evaluating a decision to install new field plots, potential costs would undoubtedly be different than the cost of remeasuring existing plots. Other factors such as the number of plots and forest acreage in states would also merit consideration.

Precision Standards

Direct and SAE estimates of precision in this study can shed light on the degree to which the FIA standard for precision could be met or exceeded using FH type estimators. Because the estimates in this study are for all forest land, rather than limited to commercial forest land, the 5% standard may not be representative of FIA direct estimates' precision, despite approaching the 5% standard (**Figure 6**). Even so, gains achieved using SAE demonstrate that the standard can be met at a smaller volume threshold, perhaps 90% of the current 1 billion ft^3 by incorporating photogrammetric CHM information into estimates (**Figure 6**).

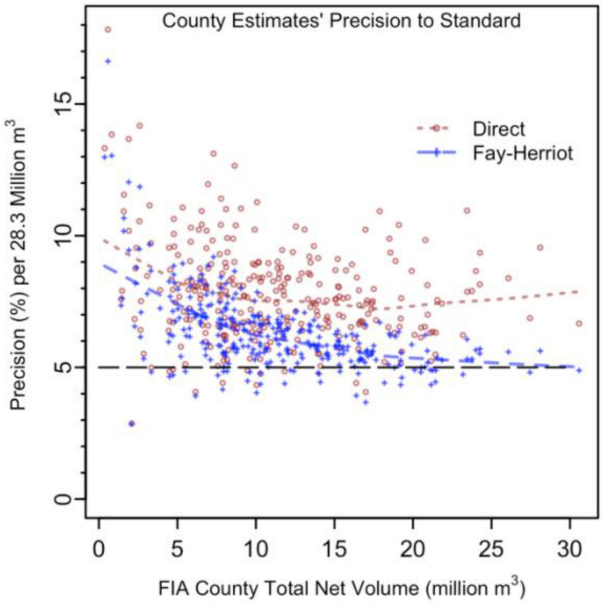


FIGURE 6 | Relative standard errors of FIA direct and FH county forest volume estimates (million m^3) with 5% precision standard per 28.3 million m^3.

CONCLUSION

Area-level SAE models using NAIP 3d DAP canopy heights as auxiliary information provided precision gains averaging between 19 and 30% for estimates of county-level forest volumes in North Carolina, Tennessee, and Virginia, compared to estimates made from FIA sample data alone. Choosing the appropriate populations from which to generate county-level FH estimates, i.e., using single states or combining data from multiple states, should be given due consideration in operational inventory settings. The applied research presented here is the first example we know of that applied SAE techniques to FIA survey data at state and county-level scales, which should make results relevant to stakeholders concerned with increasing efficiencies in FIA inventory estimation. Results suggest that the non-spatial model seemed adequate in generating county-level estimates in single-state settings, while the area-level model accounting for spatial autocorrelation was better suited in the combined three-state setting. Composite FH-type area-level estimators showed high potential for increasing precision in county-level estimates of growing stock volume with considerable gains in apparent

sample sizes, a clear measure of cost-effectiveness, and seemingly little or no added bias. Further examination of potential gains in estimating other forest attributes in the FIA program—including measuring change over time–seem warranted, as do the use of other sources of auxiliary information and the application of these methods to other states and regions in the United States.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

QC and GD: analysis and principal authorship. PR: research planning, analysis, and principal authorship. JC: project coordination, research planning, analysis, and co-authorship. JD: analysis and technical consultation. VT: research planning and scientific consultation. HB: scientific consultation and co-authorship. RW: project coordination and research planning. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by USDA, Forest Service Southern Research Station, 20-JV-11330145-074.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/ffgc.2022.769917/full#supplementary-material

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, 28–36. doi: 10.1080/01621459.1988.10478561

Bechtold, W. A., and Patterson, P. L. (2005). *The Enhanced Forest Inventory and Analysis Program – National Sampling Design and Estimation Procedures. Gen. Tech. Rep. SRS-80*. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station, 85.

Benavent, R., and Morales, D. (2016). Multivariate Fay–Herriot models for small area estimation. *Comput. Stat. Data Anal.* 94, 372–390. doi: 10.1016/j.csda.2015.07.013

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

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

Brosofske, K. D., Froese, R. E., Falkowski, M. J., and Banskota, A. (2014). A review of methods for mapping and prediction of inventory attributes for operational forest management. *For. Sci.* 60, 733–756. doi: 10.5849/forsci.12-134

Coulston, J. W., Green, P. C., Radtke, P. J., Prisley, S. P., Brooks, E. B., Thomas, V. A., et al. (2021). Enhancing the precision of broad-scale forestland removals estimates with small area estimation techniques. *For. Int. J. For. Res.* 94, 427–441. doi: 10.1093/forestry/cpaa045

Fay, R. E., 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, 269–277. doi: 10.2307/2286322

Fuller, W. A. (1999). Environmental surveys over time. *J. Agric. Biol. Environ. Stat.* 4, 331–345. doi: 10.2307/1400493

Ghosh, M., Nangia, N., and Kim, D. H. (1996). Estimation of median income of four-person families: a Bayesian time series approach. *J. Am. Stat. Assoc.* 91, 1423–1431. doi: 10.1080/01621459.1996.10476710

Goerndt, M. E., Monleon, V. J., and Temesgen, H. (2011). A comparison of small-area estimation techniques to estimate selected stand attributes using LiDAR-derived auxiliary variables. *Can. J. For. Res.* 41, 1189–1201. doi: 10.1139/x11-033

Goerndt, M. E., Monleon, V. J., and Temesgen, H. (2013). Small-area estimation of county-level forest attributes using ground data and remote sensed auxiliary information. *For. Sci.* 59, 536–548. doi: 10.5849/forsci.12-073

Goerndt, M. E., Wilson, B. T., and Aguilar, F. X. (2019). Comparison of small area estimation methods applied to biopower feedstock supply in the Northern U.S. region. *Biomass Bioenergy* 121, 64–77. doi: 10.1016/j.biombioe.2018.12.008

Green, P. C., Burkhart, H. E., Coulston, J. W., and Radtke, P. J. (2020a). A novel application of small area estimation in loblolly pine forest inventory. *Forestry* 93, 444–457. doi: 10.1093/forestry/cpz073

Green, P. C., Burkhart, H., Coulston, J., Radtke, P., and Thomas, V. A. (2020b). Auxiliary information resolution effects on small area estimation in plantation forest inventory. *Forestry* 93, 685–693. doi: 10.1093/forestry/cpaa012

Griffith, D. A. (2005). Effective geographic sample size in the presence of spatial autocorrelation. *Ann. Assoc. Am. Geogr.* 95, 740–760. doi: 10.1111/j.1467-8306.2005.00484.x

Hansen, M. C., Potapov, P. V., Goetz, S. J., Turubanova, S., Tyukavina, A., Krylov, A., et al. (2016). Mapping tree height distributions in Sub-Saharan Africa using Landsat 7 and 8 data. *Remote Sens. Environ.* 185, 221–232. doi: 10.1016/j.rse.2016.02.023

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., et al. (2013). High-resolution global maps of 21st-century forest cover change. *Science* 342, 850–853. doi: 10.1126/science.1244693

Lehtonen, R., and Veijanen, A. (2009). "Chapter 31 — Design-based methods for domains and small areas," in *Handbook of Statistics 29B Sample Surveys : Inference and Analysis*, eds D. Pfefferman and C. R. Rao (Amsterdam: Elsevier), 219–249. doi: 10.1016/s0169-7161(09)00231-4

Li, H. L., and Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *J. Multiv. Anal.* 101, 882–892. doi: 10.1016/j.jmva.2009.10.009

Magnussen, S., Mandallaz, D., Breidenbach, J., Lanz, A., and Ginzler, C. (2014). National forest inventories in the service of small area estimation of stem volume. *Can. J. Res.* 44, 1079–1090. doi: 10.1139/cjfr-2013-0448

Magnussen, S., Mauro, F., Breidenbach, J., Lanz, A., and Kändler, G. (2017). Area-level analysis of forest inventory variables. *Eur. J. For. Res.* 136, 839–855. doi: 10.1371/journal.pone.0189401

Mauro, F., Molina, I., García-Abril, A., Valbuena, R., and Ayuga-Téllez, E. (2016). Remote sensing estimates and measures of uncertainty for forest variables at different aggregation levels. *Environmetrics* 27, 225–238. doi: 10.1002/env.2387

Mauro, F., Monleon, V. J., Temesgen, H., and Ford, K. R. (2017). Analysis of area level and unit level models for small area estimation in forest inventories assisted with LiDAR auxiliary information. *PLoS One* 12:e0189401. doi: 10.1371/journal.pone.0189401

Mauro, F., Ritchie, M., Wing, B., Frank, B., Monleon, V., Temesgen, H., et al. (2019). Estimation of changes of forest structural attributes at three different spatial aggregation levels in northern California using multitemporal LiDAR. *Remote Sens.* 11:293.

McConville, K. S., Moisen, G. G., and Frescino, T. S. (2020). A tutorial on model-assisted estimation with application to forest inventory. *Forests* 11:244. doi: 10.3390/f11020244

McRoberts, R. E. (2012). Estimating forest attribute parameters for small areas using nearest neighbors techniques. *For. Ecol. Manag.* 272, 3–12. doi: 10.1016/j.foreco.2011.06.039

McRoberts, R. E., Chen, Q., and Walters, B. F. (2017). Multivariate inference for forest inventories using auxiliary airborne laser scanning data. *For. Ecol. Manag.* 401, 295–303. doi: 10.1016/j.foreco.2017.07.017

Molina, I., and Marhuenda, Y. (2015). sae: an R package for small area estimation. *R J.* 7, 81–98. doi: 10.32614/rj-2015-007

Petrucci, A., and Salvati, N. (2006). Small area estimation for spatial correlation in watershed erosion assessment. *J. Agric. Biol. Environ. Stat.* 11, 169–182. doi: 10.1198/108571106x110531

Potapov, P., Hansen, M. C., Kommareddy, I., Kommareddy, A., Turubanova, S., Pickens, A., et al. (2020). Landsat analysis ready data for global land cover and land cover change mapping. *Remote Sens.* 12:426. doi: 10.3390/rs12030426

Potapov, P., Li, X., Hernández-Serna, A., Tyukavina, A., Hansen, M. C., Kommareddy, A., et al. (2021). Mapping global forest canopy height through integration of GEDI and Landsat data. *Remote Sens. Environ.* 253:112165. doi: 10.1016/j.rse.2020.112165

Prasad, N. G. N., and Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *J. Am. Stat. Assoc.* 85, 163–171. doi: 10.1080/01621459.1990.10475320

Rao, J. N. K., and Molina, I. (2015). *Small Area Estimation*. Hoboken, NJ: John Wiley & Sons, 480.

Reams, G. A., Roesch, F. A., and Cost, N. D. (1999). Annual forest inventory: cornerstone of sustainability in the south. *J. For.* 97, 21–26.

Reich, R. M., and Aguirre-Bravo, C. (2009). Small-area estimation of forest stand structure in Jalisco, Mexico. *J. For. Res.* 20, 285–292. doi: 10.1007/s11676-009-0050-y

Singh, B. B., Shukla, G. K., and Kundu, D. (2005). Spatio-temporal models in small area estimation. *Survey Methodol.* 31, 183–195.

Strunk, J. L., Gould, P. J., Packalen, P., Gatziolis, D., Greblowska, D., Maki, C., et al. (2020). Evaluation of pushbroom DAP relative to frame camera DAP and lidar for forest modeling. *Remote Sens. Environ.* 237:111535. doi: 10.1016/j.rse.2019.111535

USDA Forest Service National Headquarters (2008). "Chapter 10 – Operational procedures," in *Forest Survey Handbook. FSH 4809.11_10* (Washington, DC: U.S. Department of Agriculture, Forest Service).

USDA Forest Service FIA (2021). *Forest Inventory and Analysis Database*. St. Paul, MN: U.S. Department of Agriculture, Forest Service, Northern Research Station.

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

Wang, H.-J., Prisley, S. P., Radtke, P. J., and Coulston, J. W. (2011). Errors In terrain-based model predictions caused by altered forest inventory plot locations in the Southern Appalachian Mountains, USA. *Math. Comput. For. Nat. Resour. Sci.* 3, 114–123.

Conflict of Interest: The 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Forests and Global Change

ORIGINAL RESEARCH
published: 25 May 2022
doi: 10.3389/ffgc.2022.779446

Small Area Estimates for National Applications: A Database to Dashboard Strategy Using *FIESTA*

Tracey S. Frescino[1], Kelly S. McConville[2], Grayson W. White[3], J. Chris Toney[1] and Gretchen G. Moisen[1]*

[1] Forest Inventory and Analysis, Rocky Mountain Research Station, USDA Forest Service, Ogden, UT, United States, [2] Department of Statistics, Harvard University, Cambridge, MA, United States, [3] RedCastle Resources, Inc., Salt Lake City, UT, United States

OPEN ACCESS

Edited by:
Brett J. Butler, Northern Research Station (USDA), United States

Reviewed by:
Steve Prisley, National Council for Air and Stream Improvement, Inc. (NCASI), United States
Andrew Finley, Michigan State University, United States

***Correspondence:**
Tracey S. Frescino
tracey.frescino@usda.gov

Specialty section:
This article was submitted to Forest Management, a section of the journal Frontiers in Forests and Global Change

Received: *18 September 2021*
Accepted: *04 May 2022*
Published: *25 May 2022*

Citation:
Frescino TS, McConville KS, White GW, Toney JC and Moisen GG (2022) Small Area Estimates for National Applications: A Database to Dashboard Strategy Using FIESTA. Front. For. Glob. Change 5:779446. doi: 10.3389/ffgc.2022.779446

This paper demonstrates a process for translating a database of forest measurements to interactive dashboards through which users can access statistically defensible estimates and analyses anywhere in the conterminous US. It taps the extensive Forest Inventory and Analysis (FIA) plot network along with national remotely sensed data layers to produce estimates using widely accepted model-assisted and small area estimation methodologies. It leverages a decade's worth of statistical and computational research on FIA's flexible estimation engine, *FIESTA*, and provides a vehicle through which scientists and analysts can share their own tools and analytical processes. This project illustrates one pathway to moving statistical research into operational inventory processes, and makes many model-assisted and small area estimators accessible to the FIA community. To demonstrate the process, continental United States (CONUS)-wide model-assisted and small area estimates are produced for ecosubsections, counties, and level 5 watersheds (HUC 10) and made publicly available through R Shiny dashboards. Target parameters include biomass, basal area, board foot volume, proportion of forest land, cubic foot volume, and live trees per acre. Estimators demonstrated here include: the simplest direct estimator (Horvitz–Thompson), model-assisted estimators (post-stratified, generalized regression estimator, and modified generalized regression estimators), and small area estimators (empirical best linear unbiased predictors and hierarchical Bayes both at the area- and unit-level). Auxiliary data considered in the model-assisted and small area estimators included maps of tree canopy, tree classification, and climatic variables. Estimates for small domain sets were generated nationally within a few hours. Exploring results across estimators and target variables revealed the progressive gains in precision using (in order of least gain to highest gain) Horvitz–Thompson, post-stratification, modified generalized regression estimators, generalized regression estimators, area-level small area models, and unit-level small area models. Substantive gains are realized by expanding model-assisted estimators beyond post-stratification, allowing FIA to continue to take advantage of

design-based inference in many cases. Caution is warranted in the use of unit-level small area models due to model mis-specification. The dataset of estimates available through the dashboards provides the opportunity for others to compare estimators and explore precision expectations over specific domains and geographic regions. The dashboards also provide a forum for future development and analyses.

Keywords: small area estimation, EBLUP, post-stratification, hierarchical Bayes, Fay and Herriot model, National Forest Inventory

INTRODUCTION

The USDA Forest Service, Forest Inventory and Analysis (FIA) program is responsible for reporting status and trends of the nation's forests and is mandated by Congress, through the 1928 McSweeney-McNary Forest Research Act and the 1974 Forest and Rangeland Renewable Resources Planning Act, to inventory and maintain a national database and provide estimates at State and National levels. The inventory was designed to provide strategic level information (Gillespie, 1999), with states being the standard reporting units, and post-stratification being the predominant estimator used in production processes (Bechtold and Patterson, 2015). Yet, there is a growing need for more precise and statistically defensible estimates to support forest land management over sub-State areas (U.S. Department of Agriculture, 2014; Prisley et al., 2021; Wiener et al., 2021).

To provide some examples, while standard FIA reporting provides analyses over entire states or regional collections of counties within a state (Witt et al., 2018; U.S. Department of Agriculture, Forest Service, 2021), estimates of forest resources are frequently needed by individual counties for county-level assessments (Morin et al., 2015; Filippelli et al., 2020) and alignment of sustainable management practices to national efforts (U.S. Department of Agriculture, Forest Service, 2020). Further, the USDA Forest Service has emphasized an ecological approach to managing forests and directing policy by its development of a hierarchical framework of Ecological Units (ECOMAP; Cleland et al., 2007). The classification was aimed at providing a scientific basis for analyzing ecosystems at different scales, depending on the management need. The ECOMAP delineations are frequently used for analyzing vegetation patterns (West et al., 1998; Hanberry et al., 2018; Miller et al., 2018) and ecological subsections provide a national collection of areas for which estimates of forest attributes would be useful. As another example, the Forest Service has recognized the need for assessing and monitoring the hydrologic systems across the US. Quantifying forest attributes within watersheds, particularly in conjunction with disturbance events, is needed for assessing variables such as stream flow and snowpack (Goeking and Tarboton, 2020). It is important to have the ability to construct estimates of forest attributes across smaller political, ecological, and hydrologic areas of interest.

One question that frequently arises is: how can we take advantage of FIA's extensive, strategic-level national database to generate estimates for areas that do not have enough sample plots using current estimation strategies to get meaningful estimates? Auxiliary data generated from remotely sensed platforms is abundant, inexpensive, and is often correlated with forest attributes of interest. One way to use the auxiliary data is to build a model for the forest attribute of interest using the FIA plot data as the response, and the auxiliary data intersected at those ground plots as the predictor variables. From this model, a wall-to-wall map of predictions of the forest attribute of interest is generated. The assumed statistical framework determines how the predicted values are aggregated to form an estimate and how the estimator accounts for the sampling design. Post-stratification is one of the simplest forms of model-assisted estimation and is the estimator currently employed in FIA's production processes. But numerous other model-assisted estimators offer further opportunity to make better use of auxiliary data (e.g., McConville et al., 2020). In model-assisted estimation, the model is simply used as a vehicle for estimating parameters in the regression estimator formula. We are not making the assumption that the population was really generated by that model. Therefore, model-assisted estimators are considered robust to model mis-specification (meaning they are asymptotically unbiased for the population attribute and the variance formulas are valid) regardless of whether or not the working model is an accurate reflection of the relationship between the variable of interest and auxiliary variables. Small area estimators (e.g., Rao and Molina, 2015), on the other hand, are needed in instances where there are too few sample plots in order to produce a reliable estimate using only data within those small domains of interest. In this case, small area estimators "borrow strength" (both sample plots and auxiliary data) from other similar areas to increase the effective sample size from which information can be produced. This borrowing process is orchestrated through a model from which measures of precision can be derived. Small area estimators rely on model-based inference which means the observations are assumed to be random realizations of some superpopulation. That is, unlike model-assisted estimators (which rely on design-based inference), we are making the assumption that the model did generate the population. One should be careful when comparing the standard error estimates of design-based and model-based methods because each paradigm conceptualizes randomness differently. In design-based inference the primary source of randomness comes from the sampling of units from the population while model-based inference considers the data to be realizations from a superpopulation model. These different conceptualizations impact how the standard error of the estimator is calculated. In addition, substantial gains in precision can be realized from model-based estimators, but

they can easily yield biased estimates if the model is mis-specified.

Recent reviews of the use of model-assisted and small area estimators in forest inventory applications are provided by Guldin (2021) and Dettmann et al. (2022). Extending beyond those reviews, the last year has seen a spike in investigations into improving precision in FIA estimates over small domains. For example, in the Interior Western US, estimates for multiple forest attributes were explored using a modified generalized regression estimator over counties (Wojcik et al., 2022). And area-level Hierarchical Bayesian and Empirical Best Linear Unbiased Predictor strategies were compared to post-stratification over ecological subsections (White et al., 2021). In the Pacific Northwest, Bell et al. (2022) compare Horvitz Thompson, generalized regression, and k-nearest neighbor synthetic estimates of aboveground live carbon. Temesgen et al. (2021) use Fay–Herriot models of above ground biomass and volume specific to stand-level inventories where variable radius plot locations may be unknown. In the Southern US, Cao et al. (2022) improve precision in volume estimates for counties using spatial area-level small area estimators. In the northern US, Harris et al. (2021) compare design- and model-based estimates in support of the National Woodland Owner Survey. Across the Western US, Gaines and Affleck (2021) estimate postfire tree density through temporal borrowing strategies. And across the conterminous US, Stanke et al. (2022) use *rFIA* to facilitate spatial Fay–Herriot models of forest carbon stocks.

Constructing estimates over non-traditional boundaries requires a shift to using these statistical estimators that can better leverage improved auxiliary remotely sensed data. *FIESTA* (Forest Inventory ESTimation for Analysis) (Frescino et al., 2020) is an R package that was originally developed to support the production of estimates consistent with current tools available from the FIA National Program, such as DATIM (Design and Analysis Toolkit for Inventory and Monitoring) and EVALIDator[1]. *FIESTA* provides an alternative data retrieval and reporting tool that is functional within the R environment, allowing customized applications and compatibility with other R-based analyses. It hosts a growing suite of model-assisted and small area estimators. While the package itself is available publicly for R users, most forest land managers need tools that do not require programming expertise. A first step in making estimates available is through distribution *via* a dashboard.

In this paper we first demonstrate a national, production-level process whereby a large collection of model-assisted and small area estimators can be rapidly applied in *FIESTA* for a variety of forest attributes and domains across the conterminous US. Second, we compare the levels of precision that can be achieved using these different estimators for different sized domains, providing benchmarks from which future improvement can be made. And third, we provide estimates and their standard errors through publicly available dashboards so others can perform analyses in different regions of the country.

[1]https://www.fia.fs.fed.us/tools-data/index.php

MATERIALS AND METHODS

FIESTA

FIESTA is an R package made up of a set of functions for compiling response data and auxiliary information for use in different estimation strategies, including simple random sampling, model-assisted, and model-based small area estimation (SAE). The functions are categorized based on different purposes: *FIESTA*'s core functions include code for querying and summarizing FIA data and different types of spatial data; *FIESTA* modules present different estimation strategies; and *FIESTA* analysis functions are wrapper functions to streamline different estimation routines.

We created an analysis function to generate and compare estimates using several different estimators for any defined domain(s) as depicted in **Figure 1**. The analysis function combines *FIESTA* core functions to: (1) extract FIA inventory data and (2) compile and summarize auxiliary information from multiple spatial data layers by domain. These are shaded in blue to indicate data compilation processes. From here, another function formats the output from these core functions, for input to the *FIESTA* Green-Book (GB), Model-Assisted (MA), and Small-Area (SA) estimation modules, including adjustments for non-response and auxiliary data standardization. The estimation modules, shaded in green, draw from a number of published R packages and generate estimates and standard errors by response for each domain.

Domains of Interest

To illustrate this database to dashboard process, three national datasets were used as targets for constructing forest population estimates: (1) Cleland Ecomap Subsections (Cleland et al., 2007), (2) County boundaries (U.S. Census Bureau, 2019) and (3) Watershed Boundary Dataset (WBD) – hydrological unit code (HUC) 10 (U.S. Geological Survey [USGS], 2013). The Cleland Ecomap dataset consists of a set of polygon feature classes across the conterminous United States, delineated from a nested, hierarchical classification based on ecological associations, including climate, physiography, hydrology, soil, and vegetative characteristics. The ecosubsection polygon feature classes are the smallest unit of Ecomap classification, ranging from 55 thousand acres (222 square kilometers) to over 8 million acres (32,375 square kilometers) in size. The US Census Bureau delineation of counties is based on political boundaries, without any consideration of ecological characteristics. Here, the Federal Information Processing Standards (FIPS) codes were used as domain identifiers. The sizes of the counties range from 292 thousand acres (1181 square kilometers) to approximately 6 million acres (24,281 square kilometers). The hydrological units (HU) are from a standardized, nested hierarchical system made up of delineations based on topographic, hydrologic, and other relevant landscape characteristics, defining surface water drainage across the United States. The HUC levels range from the largest, first-level (HUC-2) region, averaging approximately 123 million acres (496 square kilometers) to the smallest, sixth-level (HUD-12) sub-watershed, averaging approximately 26 thousand acres (107 square kilometers). We

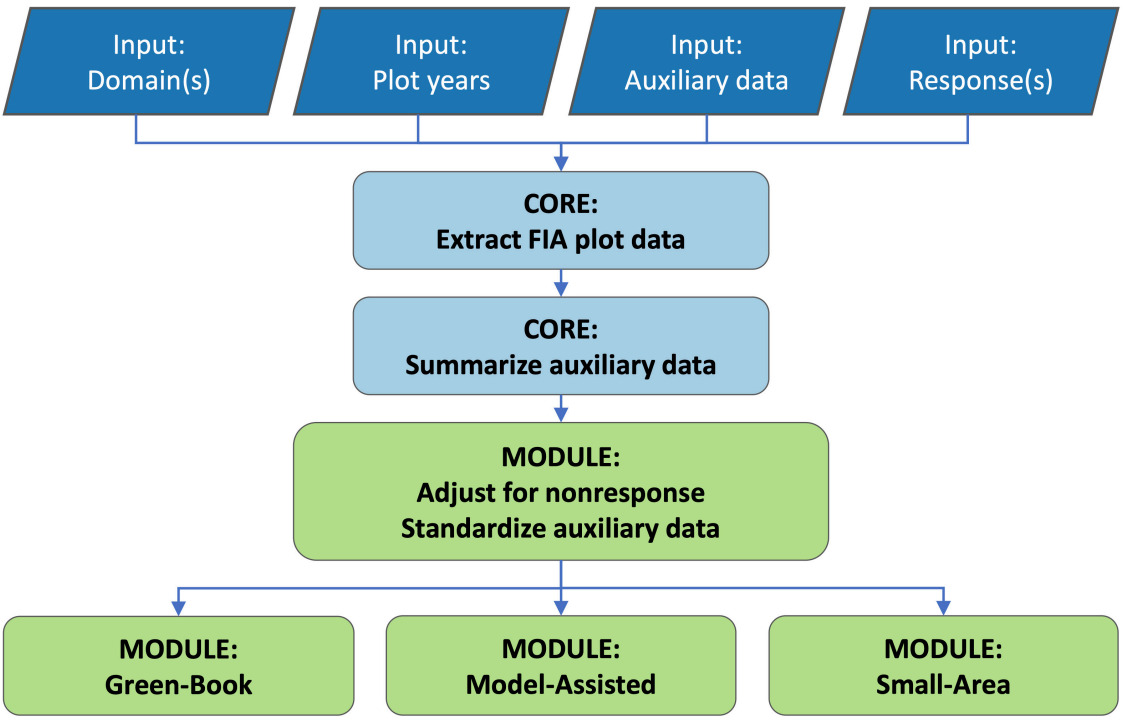


FIGURE 1 | Flowchart of *FIESTA* functions used for dashboard estimates.

used the fifth-level (HUC-10) watershed as our domains of interest, with areas averaging approximately 144 thousand acres (585 square kilometers).

We used the Cleland Ecomap Province boundaries to define areas for which our small area estimators would borrow strength from, based on the assumption that within-province domains are more homogenous for fitting models, and will therefore offer a collection of similar plots to increase our effective sample size with, and help constrain the variance of estimates. For ease of processing, we generated post-stratified and model-assisted estimates by province as well for each domain. There are a total of 39 provinces across the conterminous US, ranging from approximately 3 million acres (12 thousand square kilometers) to 195 million acres (789,000 square kilometers) in size. Polygon domains that crossed more than one province were assigned to a province based on a plurality overlap. **Figure 2** illustrates the designation of ecosubsections, counties, and watersheds within province boundaries.

Response Data

The FIA updates and maintains a comprehensive database of forest inventory data across the U.S. based on a sample of plots, each representing approximately one acre of land. The database stores: tree-level measurements, including diameter and height; forest condition observations, including stand size and forest type; and a slew of calculated attributes, including basal area, volume, and biomass. The response data used in this analysis were extracted from the FIA database based on the most current measurement of each sampled plot at the time of download (2021 July 29). Only single intensity plots were used for this analysis to assure equal sampling probabilities across the populations. It should be noted that only the unit-level, model-based estimators require data from an equal probability sample design while all of the other estimators can account for unequal probability samples.

We used six different forest attributes as the focus for this analysis: forest area; live basal area (sqft) of trees 1.0 inch diameter and greater; number of live trees 1.0 inch diameter and greater; net board-foot (International 1/4-inch Rule) volume of live trees; cubic-foot volume of live trees; and biomass of live trees 1.0 inch diameter and greater, in tons (Burrill et al., 2021). All response data were expanded to the acre and adjusted for non-response at the plot-level; then summarized by the domain of interest. Thus, a plot that was partially sampled was assumed to be representative of the entire plot. A *FIESTA* function was used to extract and compile the data for each set of domains within each province. Plots were retrieved by intersecting states from a pre-built SQLite database and assigned to each province based on the Global Positioning System (GPS) plot location center.

Auxiliary Data

We used a limited set of auxiliary information for simplicity and consistency in the analyses. The data included two satellite-based classified images to represent current vegetative cover: (1) the 2016 USGS National Land Cover Dataset (NLCD), analytical tree canopy cover raster (Yang et al., 2018), including values from 1 to 100 representing the percent of tree canopy cover on the ground (*tcc*), and (2) the LANDFIRE 2014 Existing Vegetation Type (EVT) product (Rollins, 2009) re-classed to two classes,

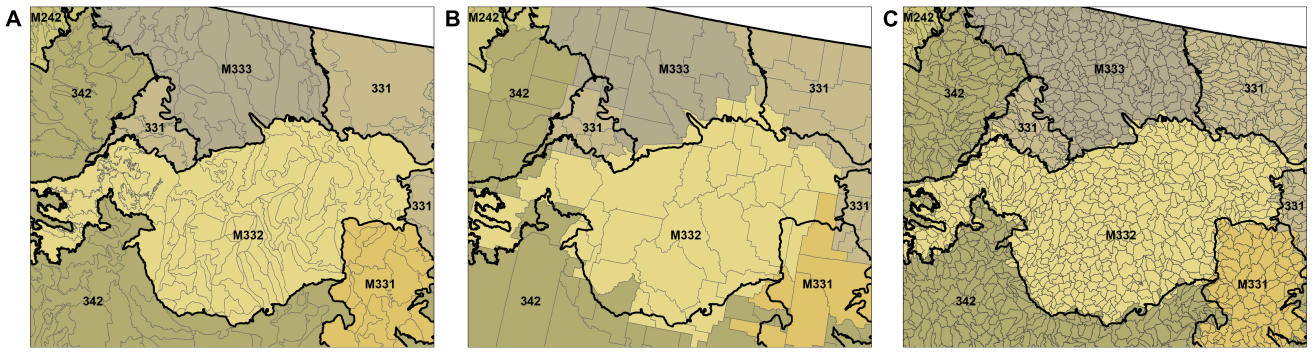


FIGURE 2 | Illustration of provincial modeling domain and the **(A)** ecosubsections, **(B)** counties, and **(C)** watersheds they contain. Domains are assigned to modeling domain based on plurality of occupancy.

representing the dominant lifeform (1: tree; 2: non-tree) (*tnt2*). The NLCD layer was resampled to 90 m using the average of the original 30 m pixels to correspond to the acre-size FIA plot more closely (Nelson et al., 2009). Similarly, the LANDFIRE classified map was resampled to 90 m using the majority value within a focal window of 3 × 3 pixels.

The next three spatial layers are from the PRISM (Parameter-elevation Regressions on Independent Slopes Model) dataset (PRISM Climate Group, 2004), and represent influential climate patterns. The data layers include 30-year normals (Daly, 2002) describing average annual precipitation (*ppt*), average annual temperature (*tmean*), and average minimum temperature (*tmin01*) for the month of January over the period 1981–2010.

The last layer was chosen to understand the local altitude characteristics of the site, the LANDFIRE 2010, elevation dataset, derived from the National Elevation Dataset (NED), representing land height, in meters, above mean sea level (*elev*). This layer was resampled from 30 m resolution to 90 m using cubic-convolution interpolation.

A *FIESTA* function was used to assign values from each auxiliary spatial layer at each FIA plot location as well as calculate zonal mean statistics by domain within each province. The function uses the Geospatial Data Abstraction Library (GDAL) for low-level access to raster and vector geospatial data formats (GDAL/OGR contributors, 2019) and C++ to increase performance for large datasets. Predictors were standardized by subtracting the mean and dividing by the standard deviation for all observations within the modeling extent (i.e., province for small area estimates and domains for post-stratified and model-assisted estimates).

Estimators

Using the same input datasets, we generated estimates for three national datasets, using one estimator programmed in *FIESTA* and seven other estimators available from packages in the Comprehensive R Archival Network (CRAN[2]), integrated through *FIESTA*. This example illustrates *FIESTA*'s versatility to call upon a variety of estimation packages and also allows a user to compare output from multiple estimation strategies within a dashboard environment.

Mimicking FIA's current estimation strategy, we produced post-stratified estimates based on the *tnt2* variable through *FIESTA*'s Green-Book module which implements estimators documented in Bechtold and Patterson (2015). We also generated estimates based on a generalized regression estimator (GREG; Sarndal, 1984; McConville et al., 2020) that was implemented through *FIESTA*'s Model-Assisted module that makes use of the *mase* R package (McConville et al., 2018).

Through *FIESTA*'s Small-Area module, we integrated multiple estimators from the *JoSAE* R package (Breidenbach, 2018), including: area-level and unit-level empirical best linear unbiased prediction (EBLUP) estimators based on the Battese–Harter–Fuller unit-level model (Battese et al., 1988) and the Fay–Herriot area-level model (Fay and Herriot, 1979); a modified generalized regression (Rao and Molina, 2015); and a Horvitz–Thompson estimator (HT; Horvitz and Thompson, 1952). Area-level EBLUPs were also fit using the *sae* R package (Molina and Marhuenda, 2015). Note that unit-level estimators rely on models that relate specific plot-level responses to specific plot-level predictors, while area-level estimators rely on models that relate averaged area-level responses to averaged area-level predictors. To obtain the EBLUP estimates, the model parameters were estimated using restricted maximum likelihood within both the *JoSAE* and *sae* packages. We also generated hierarchical Bayesian (HB) estimates using the *hbsae* R package (Boonstra, 2012). We again used the Battese–Harter–Fuller model for the unit-level HB and the Fay–Herriot model for the area-level HB now with flat priors on all of the model parameters except the ratio of the between and within area variance where a half-Cauchy prior was used (White et al., 2021). The estimators described above are consolidated in **Table 1**, along with associated acronyms used for those estimators, as well as the publically available packages and functions called by *FIESTA* to construct those estimates. The source code for the back-end estimation done in *FIESTA* is publicly available *via* the *FIESTAutils* R package (Frescino et al., 2022), particularly in the SAest.pbar and MAest.pbar functions. In this implementation of *FIESTA*, we did not use any spatial covariance structure in our models, instead

[2]https://cran.r-project.org/

TABLE 1 | Estimators and associated short names/acronyms, R packages, and specific R functions within those packages.

Estimator	Short name(s) used in paper and dashboards	Packages(s) used to fit estimator	Function(s) used to fit estimator
Horvitz–Thompson	Horvitz–Thompson, HT	mase	horvitzThompson()
Post-stratified	Post-stratified, PS	mase	postStrat()
Modified Generalized Regression	Modified GREG	JoSAE	eblup.mse.f.wrap()
Generalized Regression	GREG	mase	greg()
Unit-level empirical best linear unbiased prediction based on the Battese–Harter–Fuller model	Unit-level EBLUP, unit EBLUP	JoSAE	eblup.mse.f.wrap()
Area-level empirical best linear unbiased prediction based on the Fay–Herriot model	Area-level EBLUP, area EBLUP	sae, JoSAE	mseFH(), sae.al.f()
Unit-level hierarchical Bayesian prediction with half-Cauchy prior based on the Battese–Harter–Fuller model	Unit-level HB, unit HB	hbsae	fSAE.Unit()
Area-level hierarchical Bayesian prediction with half-Cauchy prior based on the Fay–Herriot model	Area-level HB, area HB	hbsae	fSAE.Area()

borrowing strength from ecologically similar areas serving as surrogates both for spatial proximity as well as similarity in other dimensions.

Relevant predictors were selected for each small area model (both unit- and area- level, as well as the modified GREG) using the elastic net component of the gregElasticNet function in the *mase* R package (McConville et al., 2018). The elastic net is a regularized regression method, which controls for multicollinearity and performs variable selection (Zou and Hastie, 2005). The regularization is a linear combination of a lasso (L1) penalty and a ridge (L2) penalty. The mixing of these two penalties is controlled by alpha, where $\alpha = 1$ is purely lasso and $\alpha = 0$ is ridge. The variables were selected using $\alpha = 0.5$. If no variables were selected, then the function was rerun with $\alpha = 0.2$. If again, no variables were selected, NA was returned for all domains in the province. Variable selection was also implemented within *mase* for the GREGs, also using the elastic net procedure.

In addition, for area-level small area models, domains were identified up front where models would fail (e.g., where number of observations per domain were less than or equal to 1, or where variance of the response within that domain was 0) and returned with NA values.

Dashboards

We created three dashboards for this article, each associated with each different national dataset used in this article: an ecosubsection dashboard, a fifth-level, HUC10 watershed dashboard, and a county dashboard. The dashboards were built using the R packages *flexdashboard* (Iannone et al., 2020) and *shiny* (Chang et al., 2021). The dashboards utilize interactive spatial data mapping R packages such as *leaflet* (Cheng et al., 2021) in order to display results across the nation. The use of *leaflet* allows for users to zoom into regions of interest and click on interactive polygons to obtain estimate information at the domain level through the visual aid of an interactive map. We also use R packages such as *ggplot2* (Wickham, 2016) and *plotly* (Sievert, 2020) to visualize estimates graphically and the R package *DT* (Xie et al., 2021) to create interactive data tables.

RESULTS

Continental United States Processing

Estimates and standard errors from eight different estimators were generated across the conterminous US for six forest responses using the *FIESTA* R package. We ran a compiled set of *FIESTA* functions for each national dataset that performed: database extractions, auxiliary data summaries, and estimation preprocessing calculations, for integration with five different estimation R packages.

Estimates for all domains within all three national datasets were completed overnight using a Windows 10, 32.0 GB RAM, 64-bit, single core, i5-6300U CPU, 2.40 GHz processor. There was an average of 963 million, 90 m pixels across our three national datasets. **Table 2** shows total times for one run by each national dataset, broken down by data compilation and estimation processes. Data compilation was a combination of plot data extraction and auxiliary spatial summaries, including pixel counts and zonal statistics for each domain across all provinces. Estimation processing included generation of small area estimates (and modified GREG) from *JoSAE*, *sae*, and *hbsae* packages, along with post-stratification from *FIESTA*, and GREG estimates from the *mase* package. Processing times also included a model selection routine from *mase* for all small area estimates (and modified GREG) and GREG estimates. On average, the GREG estimates consumed over 50% of the total estimation time. This was because a model was fit for each response for each domain (i.e., ecosubsection, county, watershed) within a province, different than small area estimates (and modified GREG), where only one model was fit for each response for

TABLE 2 | Processing times for generating eight different estimates for five response variables across the three national datasets.

	Total domains	Total minutes for data compilation	Total minutes for estimates
Counties	3,100	75	54
Ecosubsections	1,232	86	58
Watersheds	15,456	177	152

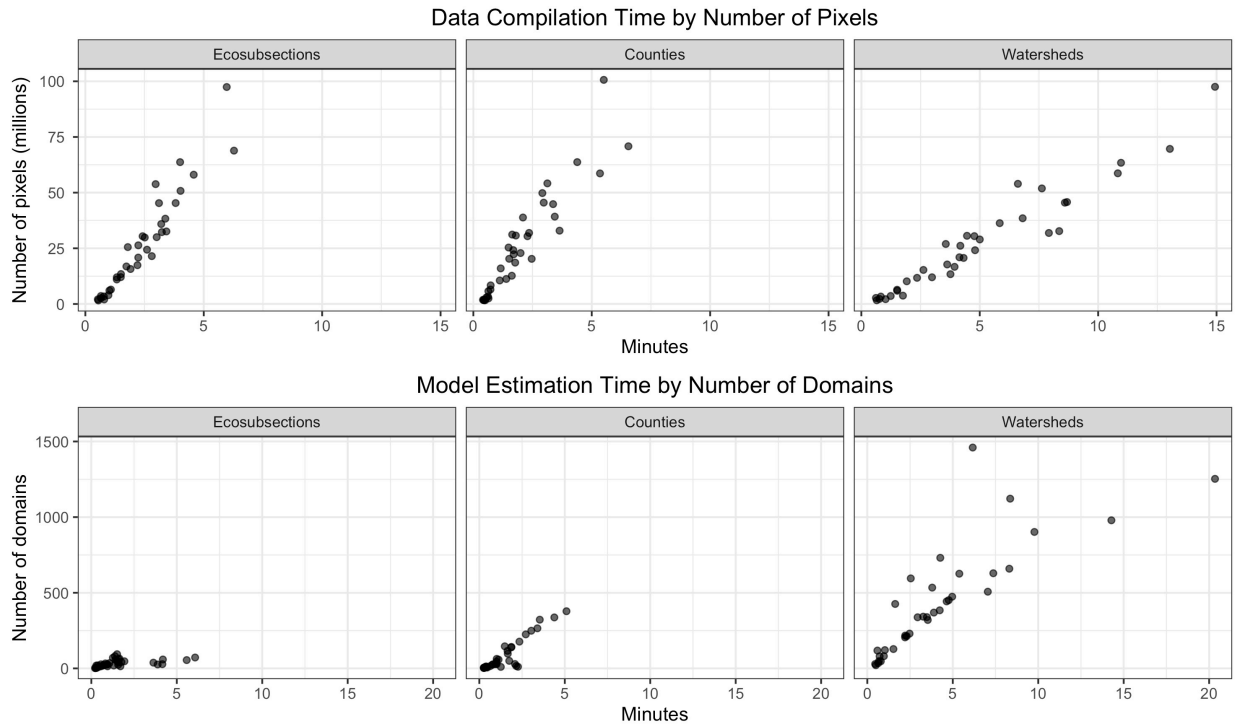


FIGURE 3 | Nationwide processing times for data summarization and model estimation. Times are expressed in minutes by number of domains and number of pixels for each of the national datasets: ecosubsections, counties, and watersheds.

each province. Processing times are further explored in **Figure 3** by national dataset as a function of number of pixels for data compilation (with the number of pixels increasing as predictors are added), and as a function of number of domains for the estimation processes. In both cases, processing time follows a linear trend, although the slope of the trend varies by national dataset. The number of domains shows a slightly stronger influence on time.

The estimation challenges posed by these three national datasets are explained by looking at a summary of the number of domains and plots within each province available to our suite of estimators. In general, ecosubsections are relatively large domains for which FIA would customarily rely on direct estimators. Contrarily, HUC 10 watersheds pose applications better suited for SAE. The sizes of counties in the US vary dramatically and are typically much smaller in the eastern US than they are in the western US. The distribution of number of domains in each province (**Figure 4A**) shows that on average, area-level models had over 30 domains to work with for the county and ecosubsection national datasets, whereas the smaller watershed delineation resulted in an average of over 200 domains per area-level model. However, both the ecosubsection and county national datasets posed challenges for the area-level models in instances where number of domains fell into the single digits. Area-level models occasionally failed in production runs, most often for domains for which there was a combination of too few domains and too weak a relationship with auxiliary data at the area-level. For the unit-level models (both model-assisted and small area) the median number of plots available (**Figure 4B**) was over 100, well within the recommended sample sizes for direct estimators. However, for watersheds, the average number of plots across provinces was only around 3. **Figure 4C** reflects how many provinces had extremely small numbers of plots at the domain level. For ecosubsections, very few did, with the minimum never falling below about 10 plots. However both the county and watershed national datasets had a number of provinces where only 1 plot was available in some of the domains, precluding the use of area-level models in those cases. Finally, the maximum number of plots by domain within province in **Figure 4D** illustrates how rarely there are a sufficient number of plots within watersheds for direct estimation.

Variable selection was part of the nationwide processing to minimize model failure rates and improve model specification. Although all the auxiliary data made available to the estimation modules were known to have some relationship to FIA response variables, that relationship is naturally different across provinces and estimators. To provide a sense of variable importance nationally, **Figure 5** illustrates the percentage of times each predictor variable was selected by the elastic net for unit-level and area-level EBLUPs of basal area for the watershed national data set. The *tcc* and *tnt2* predictors are most often included in both unit- and area-level models, with *ppt*, *elev*, *tmean*, and *tmin01* selected less often. With the weaker relationships at the unit-level than the area-level, the elastic net most commonly selected 2 predictors at the unit-level and 4 predictors at the

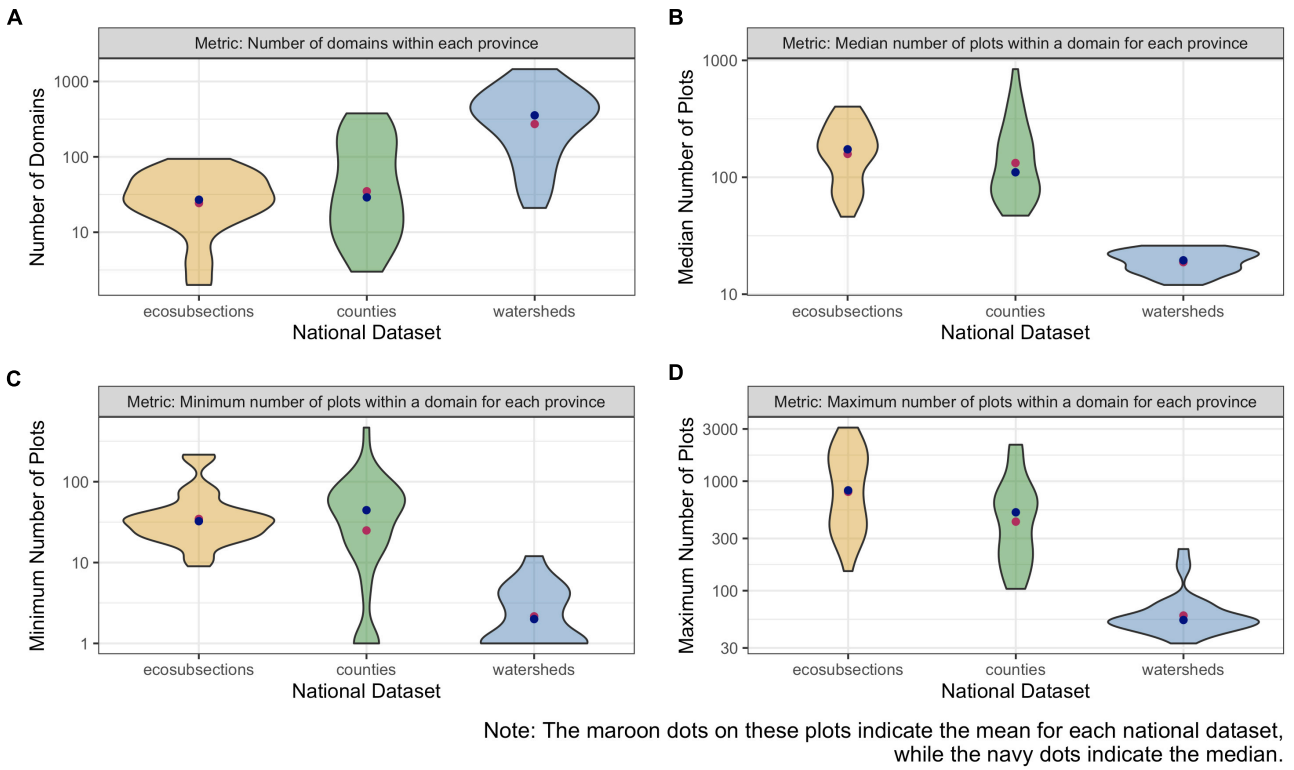


FIGURE 4 | Violin plots reflecting the distribution, by national dataset, of **(A)** number of domains within the provinces, as well as **(B)** median, **(C)** minimum, and **(D)** maximum number of plots per domain within the provinces. The maroon dots on these plots indicate the mean for each national dataset, while the navy dots indicate the median. The color of the violin plot indicates the national dataset it represents: yellow for ecosubsections, green for counties, and blue for watersheds. Note that the *y*-axis is spaced on a log10 scale.

area-level. Strongly correlated predictors, such as *tcc* and *tnt2,* exhibited a grouping effect where they were either all included or excluded from the model, a known phenomenon for the elastic-net procedure (Zou and Hastie, 2005).

Results from this nationwide processing are represented by **Figures 6A–C**, which depict the small area estimates of basal area for ecosubsections, counties, and watersheds, respectively. The estimator used is the *sae* area-level EBLUP. Missing values were filled with *JoSAE*'s area-level EBLUP and then with the Horvitz-Thompson estimator if needed. This process filled all holes except for 12 ecosubsections. For those 12 ecosubsections, there were no sampled plots with response variable greater than zero, so these were given a value of zero.

Precision and Bias

With *FIESTA*'s ability to compute a wide range of estimators, we can now easily make comparisons of the performance of different estimation approaches. **Figure 7** displays the median relative efficiency of each of the eight estimators of basal area over all domains across continental United States (CONUS). Numbers in each cell reflect the median ratio of the variance of the estimates derived under the estimator named in the column over the variance of the estimator named in the row. Reading an estimator's median variance ratio down the column allows one to see its median variance ratio where it is in the numerator, while reading an estimator's median variance ratio across the row allows one to see the ratio where it is in the denominator. Red cells indicate higher valued ratios, meaning that the estimator in the denominator is less variable, while blue cells indicate lower valued ratios, meaning that the estimator in the numerator is less variable.

From **Figure 7**, we see that the direct estimators tend to have higher median variance estimates than the indirect estimators. Among the direct estimators, the modified GREG and the GREG, which incorporate more of the auxiliary data, tends to be less variable than the HT, which utilizes no auxiliary data, and the PS, which uses one categorical, auxiliary data layer. The variance estimates tend to be slightly lower when modeling at the domain (GREG) instead of the province (modified GREG), which may be explained by the GREG's tendency to underestimate the variance when using an internal model (Kangas et al., 2016). In general, the best direct estimator was the GREG and its average relative efficiency over a Horvitz Thompson estimator for the three national datasets ranged from 0.35 to 0.45. In fact, the GREG is fairly competitive with the indirect estimators and in some cases results in a smaller median variance estimate, especially for ecosubsection domains which tend to have larger sample sizes. Among the indirect estimators, the HB and EBLUP approaches show strong agreement, which isn't surprising given the moderately sized samples, large number of domains, and

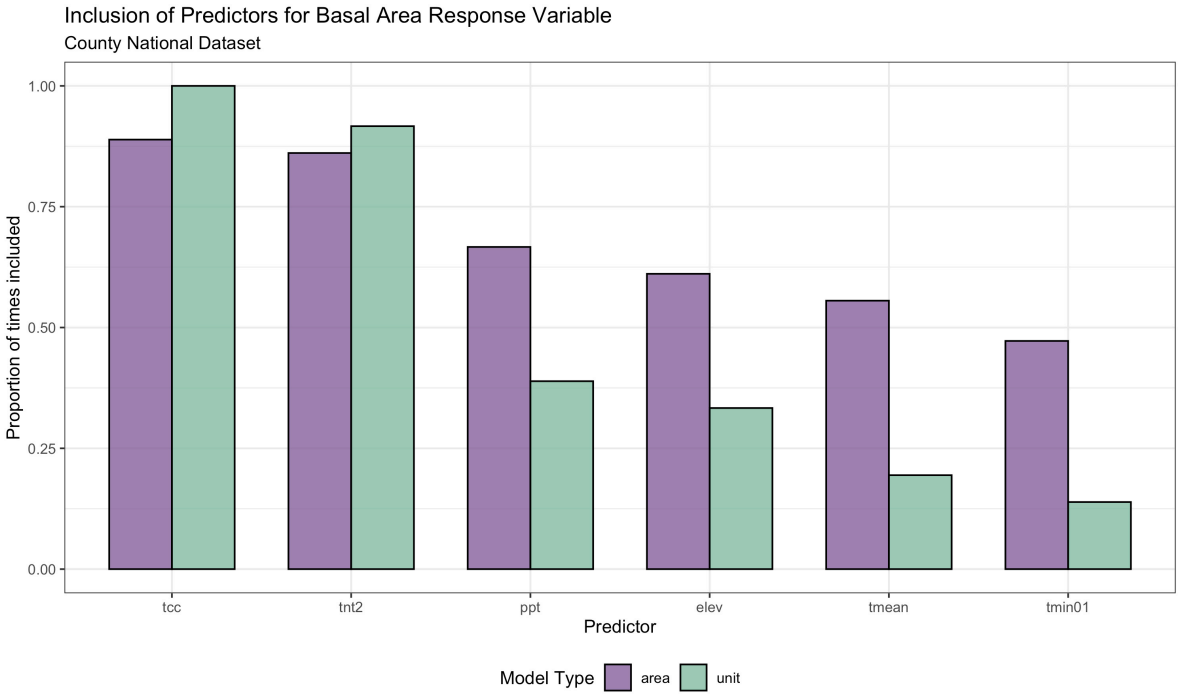


FIGURE 5 | Proportion of times each of the predictor variables was selected for area-level (purple) and unit-level (green) model-based estimators (both EBLUP and HB) applied to the county national dataset through the elastic net variable selection process used in this paper.

weakly informative prior on the ratio of the within and between variation. Relative efficiencies range from 0.96 to 1.0.

The relative gain in precision we obtain using any one of these estimators is further clarified as a function of sample size in **Figures 8A–C** for ecosubsections, watersheds, and counties respectively. Here, smoothed curves of standard errors for basal area are plotted against sample size for each estimator. We see a consistent pattern across national datasets. As expected the direct estimators yield the highest variances with direct being the worst, followed by improvements with post-stratification, modified GREG, and GREG. Also as expected, the model-based estimators show considerable improvement over smaller sample sizes, with the unit-level EBLUP and HB, as well as the area-level EBLUPs and HB yielding similar results with a slight improvement from the unit-level estimators. Similar patterns were seen for the other response variables.

Overall, between the indirect, area-level and indirect, unit-level estimators, neither is consistently more efficient than the other. However, **Figure 9** illustrates the potential for model misspecification in unit-level models. **Figure 9A** we see the challenging relationship between basal area and *tcc* at the unit level for province M221 with counties as the national dataset. **Figure 9B** depicts the clear linear relationship between the same variables, in the same province and with the same national dataset, but at the area level. The consequences of using the unit- versus the area-level estimators are also depicted where the Horvitz Thompson estimates are plotted against the unit-level EBLUP in **Figure 9C**, and against the area-level EBLUP in **Figure 9D**. The potential for over predicting biomass in the unit level instance is apparent at the zero tick for the x-axis. Nationally, this pattern persists for all the unit-level models (EBLUP, HB, and modified GREG) as shown in **Figure 10**. Negative estimates are also occasionally provided by the unit-level models.

Dashboards

While we are able to draw several useful conclusions using the static tables and figures provided in the results above, the *FIESTA* dashboards provide an interactive venue for users to explore these estimators in greater depth. For ecosubsections, counties, and watersheds, these can be found at https://ncasi-shiny-tools.shinyapps.io/Ecosubsections/, https://ncasi-shiny-tools.shinyapps.io/Counties/, and https://ncasi-shiny-tools.shinyapps.io/Watersheds/, respectively. In these dashboards, the tables and figures adapt dynamically to the user's choices for state, attributes, and estimators.

Figure 11 gives an example of the watershed dashboard. From the "Maps and Data" tab (A) the user can select the state, the forest attribute, and which of the design- or model-based estimates they are interested in. Clicking on any individual watershed reveals the watershed name, as well as the estimate and standard error for the attribute selected. Then from the "Estimator Comparisons" tab, the user can compare the performance of any of the estimators for their attribute and state of choice through graphs depicting the distribution of estimates and standard errors, as well as a table of relative efficiencies comparable to **Figure 7C**, but for the users specified state and attribute.

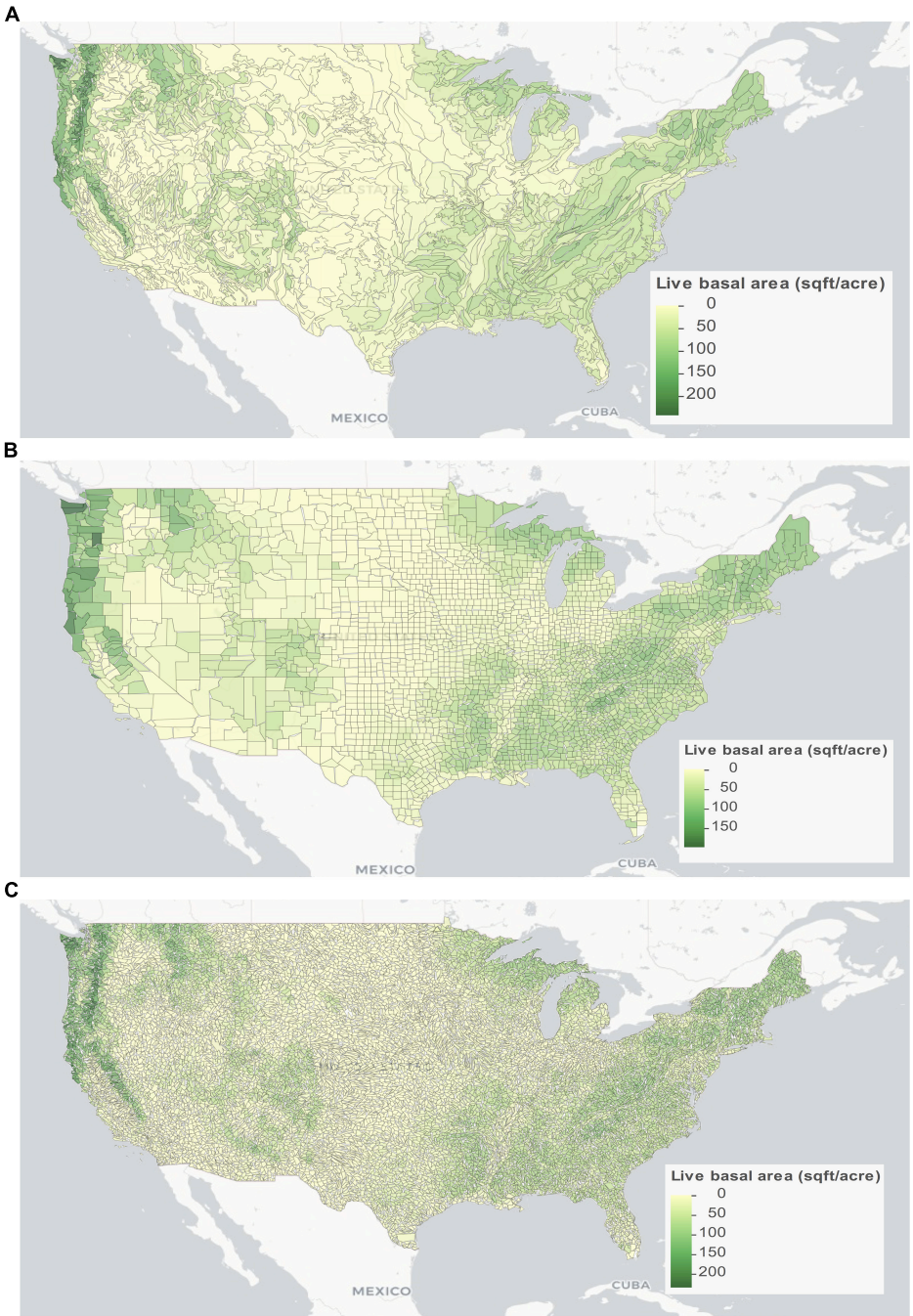


FIGURE 6 | Small area estimates of basal area for **(A)** ecosubsections, **(B)** counties, and **(C)** watersheds based predominantly on area-level EBLUPs.

DISCUSSION

In this paper, we demonstrated a process for using FIA's extensive, strategic-level, national database for generating estimates within non-traditional extents across the US, and present results from a wide range of alternative estimators in a user-friendly dashboard environment. The study of SAE and other alternative estimation strategies for forest inventories is rapidly evolving. *FIESTA* offers the flexibility to accommodate these estimation strategies, along with integrating unique responses, multiple auxiliary data sources, and different model fitting specifications, to continue this evolution through user-friendly delivery systems, and it reveals a pathway for using FIA data in more creative ways for answering forest research questions. The demonstration presented in this paper of a nationwide processing system, precision analyses, and dashboards, answered several questions

A

Ecosubsections

Estimator (Denominator)	Estimator (Numerator) Horvitz-Thompson	Post-Stratified	Modified GREG	GREG	Unit EBLUP	Unit HB	Area EBLUP	Area HB
Horvitz-Thompson	1.000	0.754	0.510	0.450	0.415	0.414	0.677	0.646
Post-Stratified	1.326	1.000	0.694	0.631	0.577	0.576	0.888	0.846
Modified GREG	1.961	1.440	1.000	0.926	0.857	0.852	1.172	1.136
GREG	2.222	1.585	1.080	1.000	0.939	0.930	1.377	1.324
Unit EBLUP	2.407	1.734	1.166	1.065	1.000	1.000	1.401	1.332
Unit HB	2.416	1.736	1.174	1.075	1.000	1.000	1.410	1.339
Area EBLUP	1.477	1.126	0.853	0.726	0.714	0.709	1.000	0.972
Area HB	1.547	1.181	0.880	0.755	0.751	0.747	1.029	1.000

B

Counties

Estimator (Denominator)	Estimator (Numerator) Horvitz-Thompson	Post-Stratified	Modified GREG	GREG	Unit EBLUP	Unit HB	Area EBLUP	Area HB
Horvitz-Thompson	1.000	0.723	0.467	0.417	0.294	0.293	0.305	0.280
Post-Stratified	1.382	1.000	0.678	0.611	0.428	0.426	0.425	0.392
Modified GREG	2.143	1.474	1.000	0.917	0.657	0.654	0.620	0.561
GREG	2.400	1.636	1.091	1.000	0.725	0.723	0.726	0.670
Unit EBLUP	3.402	2.338	1.521	1.379	1.000	0.999	0.936	0.865
Unit HB	3.415	2.350	1.528	1.383	1.001	1.000	0.937	0.869
Area EBLUP	3.281	2.351	1.614	1.377	1.069	1.067	1.000	0.962
Area HB	3.567	2.552	1.783	1.493	1.156	1.150	1.039	1.000

C

Watersheds

Estimator (Denominator)	Estimator (Numerator) Horvitz-Thompson	Post-Stratified	Modified GREG	GREG	Unit EBLUP	Unit HB	Area EBLUP	Area HB
Horvitz-Thompson	1.000	0.798	0.491	0.347	0.210	0.209	0.282	0.275
Post-Stratified	1.254	1.000	0.683	0.498	0.305	0.302	0.394	0.383
Modified GREG	2.037	1.464	1.000	0.766	0.461	0.457	0.558	0.544
GREG	2.882	2.006	1.306	1.000	0.677	0.670	0.846	0.827
Unit EBLUP	4.752	3.281	2.171	1.477	1.000	0.997	1.304	1.282
Unit HB	4.790	3.313	2.188	1.493	1.003	1.000	1.313	1.290
Area EBLUP	3.541	2.540	1.792	1.181	0.767	0.762	1.000	0.986
Area HB	3.640	2.608	1.837	1.209	0.780	0.775	1.014	1.000

FIGURE 7 | The relative efficiency of each of the eight estimators for basal area averaged over **(A)** ecosubsections, **(B)** counties, and **(C)** watersheds for the entire study region. Numbers in each cell reflect the variance of the estimates derived under the estimator named in the column divided by the variance of the estimator named in the row. Shades of blue reflect values less than 1, with the deepest blue set at the minimum value for that specific table. Conversely, shades of red reflect values greater than 1 with the deepest red set at the maximum value for that specific table.

surrounding SAE for forest inventories, but also posed additional topics warranting further research.

Inventory Attributes

Here we constructed estimates of six key FIA attributes to demonstrate the process. But FIA has information on hundreds of attributes and *FIESTA* can access any of these from FIA's extensive database to construct estimates of interest. FIA's current estimation process does not just focus on one variable at a time to conduct specific inference, rather it reports on a multitude of estimates that must be internally consistent, accommodating generic inference. *FIESTA* can

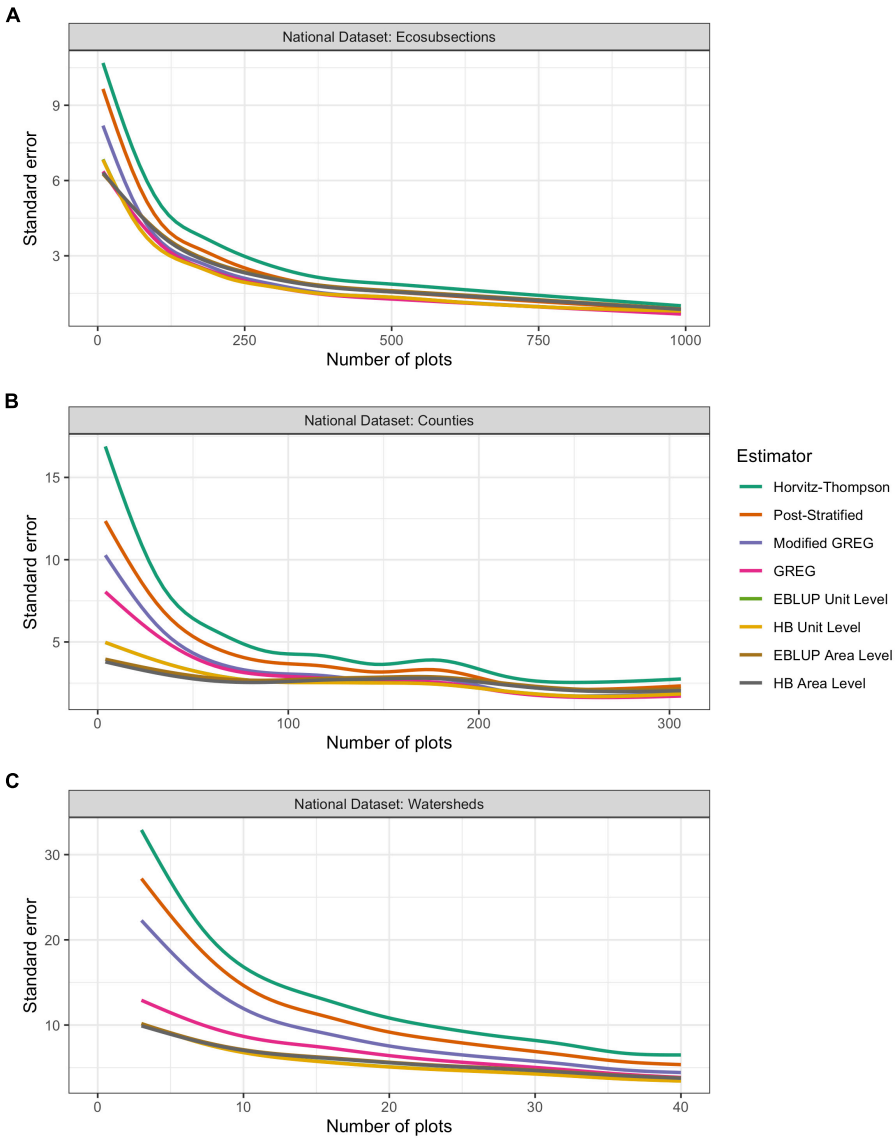


FIGURE 8 | Smoothed standard errors by number of plots for all estimates of basal area obtained over **(A)** ecosubsections, **(B)** counties, and **(C)** watersheds. The lines used to represent the data were created with a generalized additive model using smoothing splines with a cubic spline basis. The *y*-axis represents the standard error of each estimator and the *x*-axis represents the number of plots within the domain of interest. For each plot, we trimmed the number of plots to the 0.95 quantile in order to avoid high leverage points in the smoothing algorithm. This 0.95 quantile point was found to be 991 for ecosubsections, 277 for counties, and 40 for watersheds. These plots were created only for regions where no estimators produced NA values.

mimic this process using post-stratification through its Green-Book or Model-Assisted modules, and thus be compatible with current estimates. However, the opportunity exists to improve precision in these direct estimates by simply moving to other model-assisted methods such as a GREG where additional auxiliary data in either continuous or discrete format can be used, and still retain the ability for generic inference that is important to any sample survey organization. Going beyond that, though, there are instances where specific inference is called for and small area estimates are constructed through model-based methods targeting a single attribute. *FIESTA* has the ability to accommodate these types of problems through its SA module. More work is needed to provide guidelines on transitioning from model-assisted to model-based estimators in cases of specific inference. In addition, more work is needed in small domains to model FIA variables jointly to preserve their ecological consistency.

Auxiliary Data

The development and distribution of alternative auxiliary data layers relevant to forest inventory is an active area of research. The set of predictors used here provide a sensible place to start for estimates of forest status. But finer resolution and

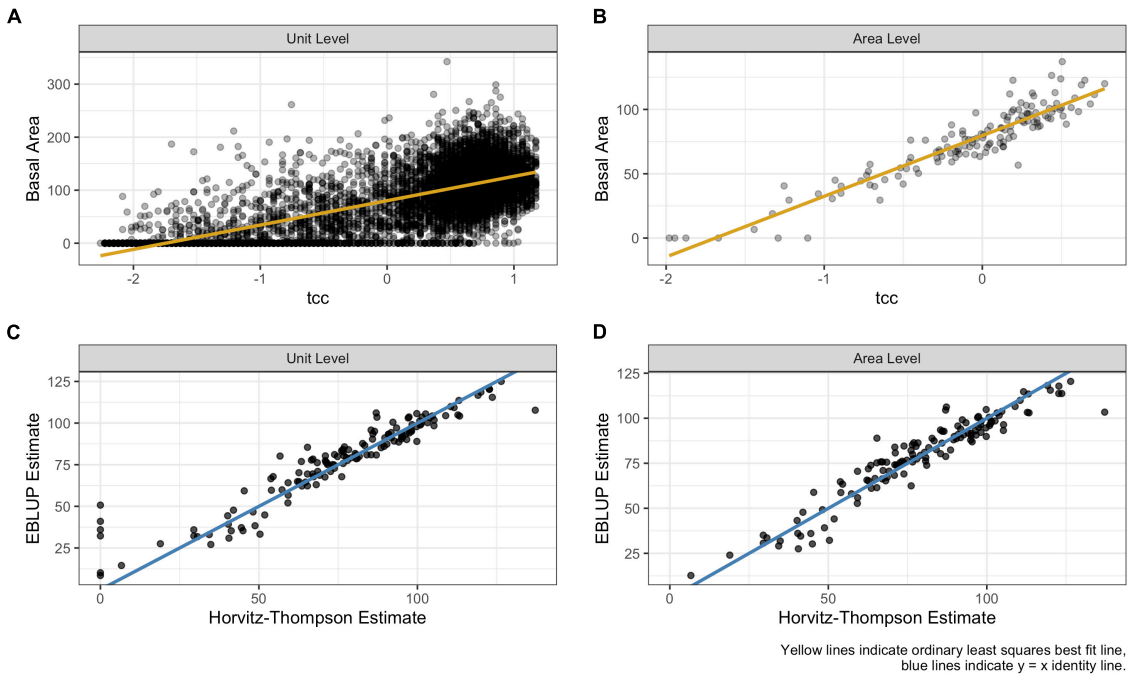


FIGURE 9 | The relationship between a response variable (basal area) and a predictor variable (total canopy cover, "tcc"). **(A)** Depicts the relationship between basal area and tcc at the unit level for province M221 with counties as the national dataset. **(B)** Depicts the relationship between the same variables, in the same province and with the same national dataset, but at the area level. **(C)** Unit and **(D)** area-level EBLUP estimates compared to the Horvitz–Thompson estimates in province M221 for the county domains. On plots **(A,B)**, the yellow line indicates the ordinary least squares best fit line and on plots **(C,D)** the blue line is the identity line.

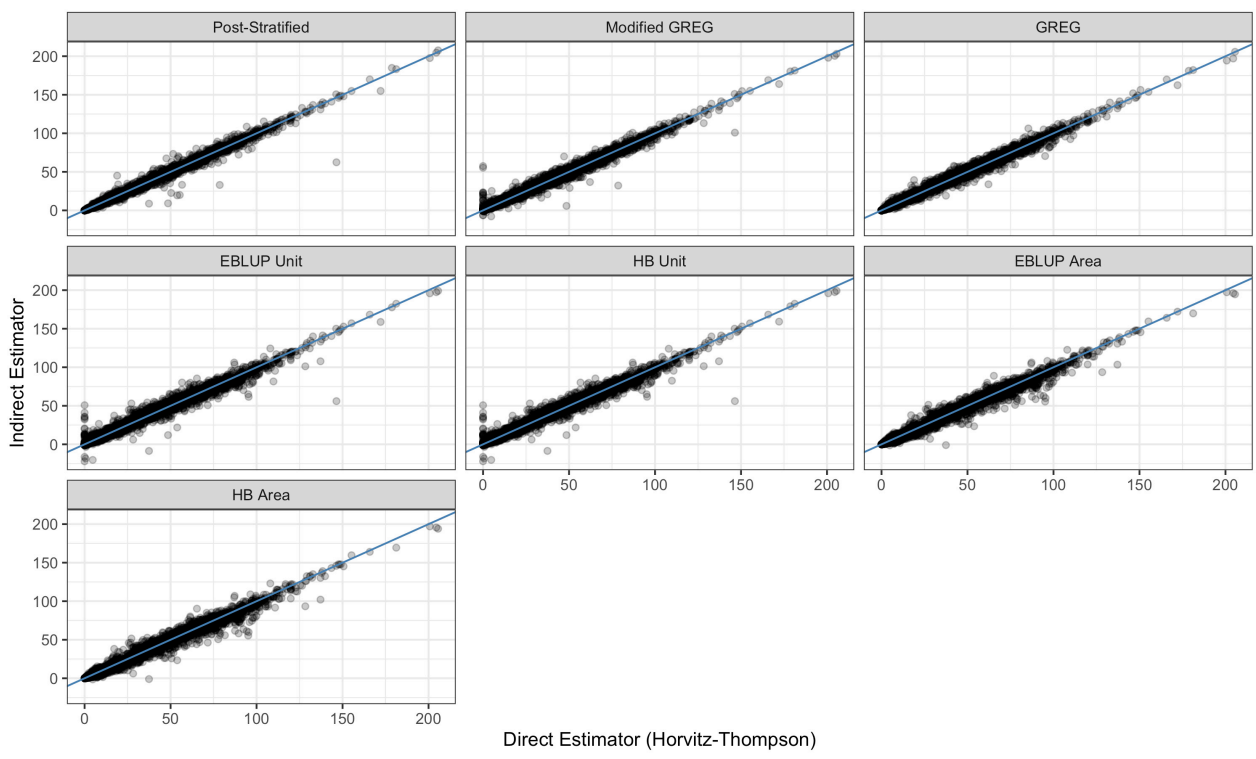


FIGURE 10 | Estimates of basal area (sqft) from each estimator compared to the Horvitz–Thompson for all counties in across the conterminous US.

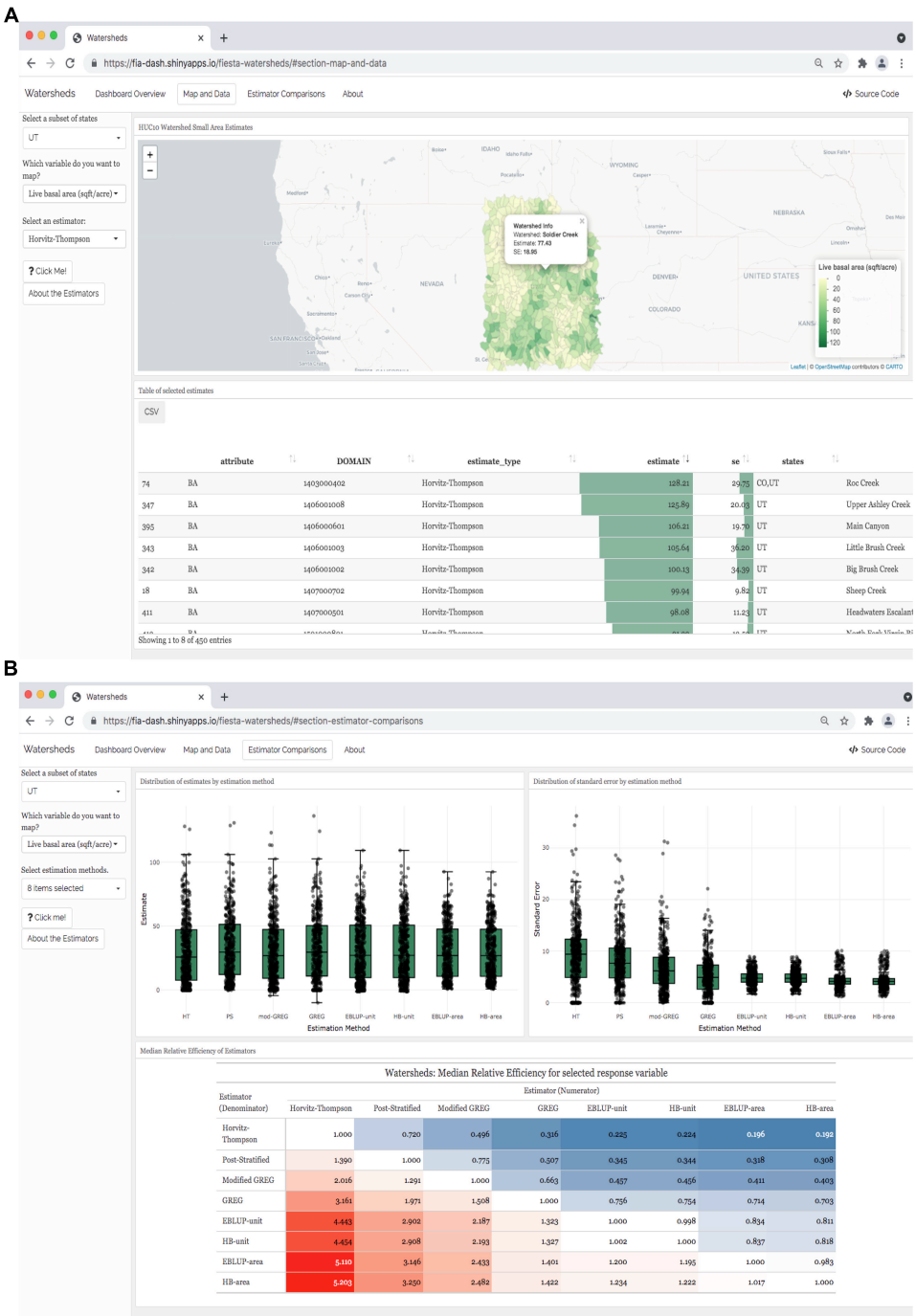


FIGURE 11 | Example views from the watershed dashboard available at https://ncasi-shiny-tools.shinyapps.io/Watersheds/. From the Maps and Data Tab **(A)** the user can select a state, which variable they are interested in and specify which of the design- or model-based estimates they would like to see. Then from the Estimator Comparisons Tab **(B)**, the user can compare the performance of any of the estimators for their variable and state of choice through graphs depicting the distribution of estimates and standard errors, as well as a table of relative efficiencies comparable to **Table 2**, but for the users specified state and attribute. Dashboards in the same format have also been constructed for ecosubsections and counties, which can be accessed at https://ncasi-shiny-tools.shinyapps.io/Ecosubsections/, and https://ncasi-shiny-tools.shinyapps.io/Counties/ respectively.

higher quality data is coming online rapidly and could offer substantial improvements in precision of inventory estimates. (See Lister et al., 2020 for a review of evolving remotely sensed products.) In addition, looking beyond status to variables that reflect change, such as growth removals and mortality, requires a very different set of auxiliary data in order to establish good

models between the response and predictors (e.g., Coulston et al., 2021). *FIESTA* is designed to use any appropriately scaled auxiliary data and allows the user to easily assess the contribution of a given set of predictors on the precision of an estimator. Although critical to defining the sample design, a nationwide layer depicting sampling intensity by year has yet to be developed and is needed to enable the use of all inventory plots, not just those collected at standard spatial and temporal scales.

Estimators

The eight estimators applied to the three national datasets illustrate the ability of *FIESTA* to draw from numerous alternative estimation packages in R. In this case, the packages *mase*, *JoSAE*, *sae*, and *hbsae* were called upon. There are many options associated with these packages that can be tapped, while *FIESTA* was also designed to plug-and-play alternative packages and arguments as needed. As new estimation packages, or user-built analysis functions become available, they can be added to the comparison to continue to get the best results for the questions asked.

This nationwide processing of domains of different sizes over CONUS revealed some important information about the efficiency of different estimators given the set of auxiliary data provided. Evaluated at a national scale the increasingly superior precision performance of post-stratification, modified GREG, and GREG is not surprising. However, the smaller variances produced by GREG over the modified GREG warrant further investigation. Also, the precision gains from model-based estimators over the design-based estimators for very small sample sizes is also not surprising. And although unit-level models produced estimates compatible with direct estimates in areas where direct estimates were reliable, the effect of model mis-specification on bias in estimates should be more fully explored. Important to note are the gains that can be realized from model-assisted methods such as GREG for smaller sample sizes while still maintaining asymptotic unbiasedness even if the models are mis-specified, and retaining the ability to conduct generic inference. This automated processing also makes tests for simply improving FIA's current post-stratification process with new auxiliary data much easier.

Model Fitting

While it is ideal to construct individual models for every variable over every geographic region, a production system needs to be automated and robust to model mis-specification. For example, while it is good to have a number of auxiliary data layers available for estimation models, variable selection techniques, like the elastic net employed here, can help ensure only meaningful data are contributing to the estimation process. This paper demonstrates the strides that have been taken to automate model-fitting strategies such as variable selection and handle inevitable issues that arise from non-convergence, insufficient data, and small numbers of domains in a production environment. However, more work should be done to evaluate variable contributions, especially in the presence of collinearity and complex, non-linear relationships.

In these nationwide runs, borrowing strength occurred at the ecological province level. However, *FIESTA* is set up to allow a user to specify a different borrowing strategy. For example, White et al. (2021) suggest, for some response variables, ecological sections may provide a better borrowing strategy for small area models, as they are smaller, more homogenous regions. Alternatively, management strategies across different forest land ownerships might suggest a reason to distinguish borrowing across public vs. private land ownerships. In addition, users may have access to higher quality or higher resolution auxiliary data in their specific geographic region and wish to constrain the borrowing area to the extent of the better data. In addition, as future work we hope to add more flexible small area models to *FIESTA* in order to account for spatial structure, as these models have been shown to increase precision in a forestry context (Ver Planck et al., 2018).

Computing and Delivery

The dashboards presented here provide a mechanism for scientists, statisticians, and other users to explore potential for precision gains and for setting expectations in geographically specific regions of the country. All the graphics presented in the results at the national scale can be subset for specific provinces or states within the dashboards. The dashboards also demonstrate an opportunity through which users of forest inventory data can explore small area perimeters and specific forest inventory variables for which estimates are needed until such time as interactive online tools are available for them to fulfill their information needs. Following the same process shown here, estimates will soon be derived for past and present wildfire perimeters across the nation to obtain a sample-based picture of resources lost to fire. Although similar strategies will be used, new challenges arise from the diversity of sizes and extents across non-contiguous boundaries.

All estimates, tables, graphics, maps, and dashboards were processed within the R environment. This project highlights the power, versatility, and magnitude of R. Although some aspects may be more efficient in other software, keeping it in one platform minimizes complexity in programming and analysis. Work is underway to increase *FIESTA*'s processing speed through conversion of spatial functions to Python. Beyond *FIESTA* access provided to novice users through dashboards like those illustrated here, plans for other distribution veins proceed as follows: for expert users, the *FIESTA* package is currently distributed on GitHub[3], and will soon be available on CRAN (see footnote 2); for novice-to-intermediate users, a stand-alone desktop application of *FIESTA* is currently rolling out; and for Esri users, *FIESTA* is being integrated into ArcGIS Pro. In addition to the backend estimation code already available in the *FIESTAutils* R package on CRAN, all other code used in this paper will be available through the open-source delivery of *FIESTA* along with additional resources in vignettes and the associated *FIESTAnalysis* package which provides wrapper functions to streamline analyses

[3]https://github.com/USDAForestService/FIESTA

using *FIESTA* functions and includes estimate diagnostics found in this paper.

CONCLUSION

Leveraging a decade's worth of statistical and computational research on FIA's flexible estimation engine, *FIESTA*, we demonstrated a process for translating information in FIA's extensive national database to interactive dashboards through which users can easily access statistically defensible estimates anywhere in the conterminous US. We combined FIA plot data with national remotely sensed data layers to produce estimates over collections of small domains using published and widely accepted model-assisted and SAE methodologies. Based on national analyses, the order of estimator performance for smaller sample sizes (ranging from best to worst precision) was unit-level small area models, area-level small area models, generalized regression estimators, modified generalized regression estimators, post-stratification, and Horvitz–Thompson. But the gains in precision for unit-level over the area-level small area models do not offset the potential for bias due to model mis-specification in unit-level models. Further, for moderate sample sizes, substantive gains in precision can be realized by simply moving beyond post-stratification to alternative model-assisted estimators like generalized regression, to capitalize on information from auxiliary data and retain the advantages of direct design-based estimators. The extensive dataset of estimates available through the dashboards provides the opportunity for others to compare estimators and explore precision expectations over specific domains and geographic areas of the country. The dashboards also provide a forum for future development and analyses. This project also illustrates one pathway to moving statistical research into operational inventory processes, providing a vehicle through which FIA scientists and analysts can share their own tools and analytical processes with others.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: the datasets generated and analyzed for this study can be found in the publicly available dashboards created for this study. The dashboards for ecosubsections, counties, and watersheds can be found at: https://ncasi-shiny-tools.shinyapps.io/Ecosubsections/, https://ncasi-shiny-tools.shinyapps.io/Counties/, and https://ncasi-shiny-tools.shinyapps.io/Watersheds/, respectively.

AUTHOR CONTRIBUTIONS

TF and GM: conceptualization, methodology, validation, and writing—original draft preparation. TF, GW, and JT: software. TF: formal analysis and data curation. TF, GM, and KM: investigation. GM and TF: resources. TF, GM, KM, and GW: writing—review and editing. GW: visualization. KM: supervision. GM: project administration and funding acquisition. All authors read and agreed to the published version of the manuscript.

FUNDING

KM's contributions were supported by the USDA Forest Inventory and Analysis Program (*via* agreement 19-JV-11221638-112).

ACKNOWLEDGMENTS

We want to thank Madelon Basil, Isabelle Caldwell, Alex Flowers, Sam Olson, and Olek Wojcik for creating the template used in the *FIESTA* dashboards.

REFERENCES

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *J. Am. Stat. Assoc.* 83, 28–36. doi: 10.1080/01621459.1988.10478561

Bechtold, W. A., and Patterson, P. L. (2015). *The Enhanced Forest Inventory and Analysis Program National Sampling Design and Estimation Procedures*. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station, doi: 10.2737/SRS-GTR-80

Bell, D. M., Wilson, B. T., Werstak, C. E., Oswalt, C. M., and Perry, C. H. (2022). Examining k-nearest neighbor small area estimation across scales using national forest inventory data. *Front. For. Glob. Change* 5:763422. doi: 10.3389/ffgc.2022.763422

Boonstra, H. J. (2012). *hbsae: Hierarchical Bayesian Small Area Estimation. R package version 1.0*. Available Online at: https://CRAN.R-project.org/package=hbsae (accessed July 1, 2021).

Breidenbach, J. (2018). *JoSAE: Unit-Level and Area-Level Small Area Estimation. R package version 0.3.0*. Available Online at: https://CRAN.R-project.org/package=JoSAE (accessed July 1, 2021).

Burrill, E. A., DiTommaso, A. M., Turner, J. A., Pugh, S. A., Christensen, G., Perry, C. J., et al. (2021). *FIA Database Description and User Guide for Phase 2 (version: 9.0.1). [WWW Document]*. St Paul MN: U.S. Department of Agriculture, Forest Service, North Central Research Station.

Cao, Q., Dettmann, G. T., Radtke, P. J., Coulston, J. W., Derwin, J. M., Thomas, V. A., et al. (2022). Increased precision in county-level volume estimates in the U.S. National Forest Inventory with area-level SAE. *Front. For. Glob. Change* 5:769917. doi: 10.3389/ffgc.2022.769917

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., et al. (2021). *shiny: Web Application Framework for R. R package version 1.6.0*. Available Online at: https://CRAN.R-project.org/package=shiny (accessed July 1, 2021).

Cheng, J., Karambelkar, B., and Xie, Y. (2021). *leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. R package version 2.0.4.1*. Available Online at: https://CRAN.R-project.org/package=leaflet (accessed July 1, 2021).

Cleland, D. T., Freeouf, J. A., Keys, J. E. Jr., Nowacki, G. J., Carpenter, C., and McNab, W. H. (2007). *Ecological Subregions: Sections and Subsections of the Conterminous United States [1:3,500,000] [CD-ROM]. Sloan, A.M., cartog. Gen. Tech. Report WO-76*. Washington, DC: U.S. Department of Agriculture, Forest Service.

Coulston, J. W., Green, P. C., Radtke, P. J., Prisley, S. P., Brooks, E. B., Thomas, V. A., et al. (2021). Enhancing the precision of broad-scale forestland removals estimates with small area estimation techniques. *For. Int. J. For. Res.* 50, 1–15. doi: 10.1093/forestry/cpaa045

Daly, C. (2002). *Climate division normals derived from topographically-sensitive climate grids. 13th AMS Conf. on Applied Climatology*. Portland, OR: American Meteorological Society.

Dettmann, G. T., Radtke, P. J., Coulston, J. W., Green, P. C., Wilson, B. T., and Moisen, G. G. (2022). Review and synthesis of estimation strategies to meet small area needs in forest inventory. *Front. For. Glob. Change* 5:813569. doi: 10.3389/ffgc.2022.813569

Fay, R. E., 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, 269–277. doi: 10.2307/2286322

Filippelli, S. K., Falkowski, M. J., Hudak, A. T., Fekety, P. A., Vogeler, J. C., Khalyani, A. H., et al. (2020). Monitoring pinyon-juniper cover and aboveground biomass across the Great Basin. *Environ. Res. Lett.* 15:025004. doi: 10.1088/1748-9326/ab6785

Frescino, T. S., Moisen, G. G., Patterson, P. L., Toney, J. C., and Freeman, E. A. (2020). "Demonstrating a progressive FIA through FIESTA: A bridge between science and production," in *Celebrating progress, possibilities, and partnerships: Proceedings of the 2019 Forest Inventory and Analysis (FIA) Science Stakeholder Meeting; November 19–21, 2019; Knoxville, TN. e-Gen. Tech. Rep. SRS–256*, ed. T. J. Brandeis (Asheville, NC: U.S. Department of Agriculture Forest Service, Southern Research Station), 199–200.

Frescino, T. S., Toney, C., and White, G. W. (2022). *FIESTAutils: Utility Functions for Forest Inventory Estimation and Analysis. R package version 1.0.0*. Available Online at: https://CRAN.R-project.org/package=FIESTAutils (accessed April 5, 2022).

Gaines, G. C., and Affleck, D. L. R. (2021). Small area estimation of postfire tree density using continuous forest inventory data. *Front. For. Glob. Change* 4:761509. doi: 10.3389/ffgc.2021.761509

GDAL/OGR contributors (2019). *GDAL/OGR Geospatial Data Abstraction software Library*. Beaverton, Oregon: Open Source Geospatial Foundation.

Gillespie, A. J. R. (1999). Rationale for a National Annual Forest Inventory Program. *J. For.* 97, 16–20. doi: 10.1093/jof/97.12.16

Goeking, S. A., and Tarboton, D. G. (2020). Forests and water yield: a synthesis of disturbance effects on streamflow and snowpack in western coniferous forests. *J. For.* 118, 172–192. doi: 10.1093/jofore/fvz069

Guldin, R. W. (2021). A systematic review of small domain estimation research in forestry during the twenty-first century from outside the United States. *Front. For. Glob. Change* 4:695929. doi: 10.3389/ffgc.2021.695929

Hanberry, B. B., Brzuszek, R. F., Foster, H. T. II, and Schauwecker, T. J. (2018). Recalling open old growth forests in the Southeastern mixed forest province of the United States. *Ecoscience* 26, 11–22. doi: 10.1080/11956860.2018.1499282

Harris, V., Caputo, J., Finley, A., Butler, B. J., Bowlick, F., and Catanzaro, P. (2021). Small-area estimation for the USDA Forest Service, National Woodland Owner Survey: creating a fine-scale land cover and ownership layer to support county-level population estimates. *Front. For. Glob. Change* 4:745840. doi: 10.3389/ffgc.2021.745840

Horvitz, D. G., and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* 47, 663–685. doi: 10.7717/peerj.1634

Iannone, R., Allaire, J., and Borges, B. (2020). *flexdashboard: R Markdown Format for Flexible Dashboards. R package version 0.5.2*. Available Online at: https://CRAN.R-project.org/package=flexdashboard (accessed July 1, 2021).

Kangas, A., Myllymaki, M., Gobakken, T., and Næsset, E. (2016). Model-assisted forest inventory with parametric, semiparametric, and nonparametric models. *Can. J. For. Res.* 46, 855–868. doi: 10.1139/cjfr-2015-0504

Lister, A. J., Andersen, H., Frescino, T., Gatziolis, D., Healey, S., Heath, L. S., et al. (2020). Use of Remote Sensing Data to Improve the Efficiency of National Forest Inventories: a Case Study from the United States National Forest Inventory. *Forests* 11:1364. doi: 10.3390/f11121364

McConville, K., Tang, B., Zhu, G., Cheung, S., and Li, S. (2018). *mase: Model-Assisted Survey Estimation. R package version 0.1. 2*. Available Online at: https://cran.r-project.org/package=mase (accessed July 1, 2021).

McConville, K. S., Moisen, G. G., and Frescino, T. S. (2020). A Tutorial on Model-Assisted Estimation with Application to Forest Inventory. *Forests* 11:244. doi: 10.3390/f11020244

Miller, K., McGill, B., Mitchell, B. R., and Comiskey, J. (2018). Eastern national parks protect greater tree species diversity than unprotected matrix forests. *For. Ecol. Manage.* 414, 74–84.

Molina, I., and Marhuenda, Y. (2015). sae: an R Package for Small Area Estimation. *R J.* 7, 81–98. doi: 10.32614/rj-2015-007

Morin, R. S., Pugh, S. A., Liebhold, A. M., and Crocker, S. J. (2015). "A regional assessment of emerald ash borer impacts in the Eastern United States: ash mortality and abundance trends in time and space," in *Pushing boundaries: new directions in inventory techniques and applications: Forest Inventory and Analysis (FIA) symposium 2015. 2015 December 8–10; Portland, Oregon. Gen. Tech. Rep. PNW-GTR-931*, eds S. M. Stanton and G. A. Christensen (Portland, OR: U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station), 233–236.

Nelson, M. D., McRoberts, R. E., Holden, G. R., and Bauer, M. E. (2009). Effects of satellite image spatial aggregation and resolution on estimates of forest land area. *Int. J. Remote Sens.* 30, 1913–1940. doi: 10.3390/s80 63767

Prisley, S., Bradley, J., Clutter, M., Friedman, S., Kempka, D., Rakestraw, J., et al. (2021). Needs for small area estimation: perspectives from the US private forest sector. *Front. For. Glob. Change* 4:746439. doi: 10.3389/ffgc.2021.746439

PRISM Climate Group (2004). *PRISM Climate Data*. Oregon: Oregon State University.

Rao, J. N., and Molina, I. (2015). *Small Area Estimation*. Hoboken: John Wiley & Sons.

Rollins, M. G. (2009). LANDFIRE: a Nationally Consistent Vegetation, Wildland Fire, and Fuel Assessment. *Int. J Wildland Fire* 18, 35–49.

Sarndal, C. (1984). Design-consistent versus model-dependent estimation for small domains. *J. Am. Stat. Assoc.* 79, 624–631. doi: 10.2307/2288409

Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. London: Chapman and Hall.

Stanke, H., Finley, A. O., and Domke, G. M. (2022). Simplifying small area estimation with rFIA: a demonstration of tools and techniques. *Front. For. Glob. Change* 5:745874. doi: 10.3389/ffgc.2022.745874

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. Glob. Change* 4:745916. doi: 10.3389/ffgc.2021.745916

U.S. Census Bureau (2019). *TIGER/Line Shapefiles*. Suitland: U.S. Census Bureau.

U.S. Department of Agriculture (2014). *Farm bill*. Washington, D.C: U.S. Department of Agriculture.

U.S. Department of Agriculture, Forest Service (2020). *County Governments and the USDA Forest Service: A guidebook for working together*. Washington, DC: National Association of Counties and USDA Forest Service, 66.

U.S. Department of Agriculture, Forest Service (2021). *Forests of Georgia, 2019. Resource Update FS-310*. Asheville, NC: U.S. Department of Agriculture, Forest Service, 2.

U.S. Geological Survey [USGS], U.S. Department of Agriculture, and Natural Resources Conservation Service (2013). *Federal standards and procedures for the National Watershed Boundary Dataset (WBD); 2013; TM; 11-A3; Section A: Federal Standards in Book 11 Collection and Delineation of Spatial Data*. Reston, VA: U.S. Geological Survey.

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

West, N. E., Tausch, R. J., and Tueller, P. T. (1998). *A Management-Oriented Classification of Pinyon-Juniper Woodlands of the Great Basin. Gen. Tech. Rep. RMRS-GTR-12*. Ogden. UT: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, 42.

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. Glob. Change* 4:752911. doi: 10.3389/ffgc.2021.752911

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Wiener, S. W., 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. Glob. Change* 4:763487. doi: 10.3389/ffgc.2021.763487

Witt, C., DeRose, R. J., Goeking, S. A., and Shaw, J. D. (2018). *Idaho's forest resources, 2006-2015. Resour. Bull. RMRS-RB-29*. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, 84.

Wojcik, O. C., Olson, S. D., Nguyen, P. V., McConville, K. S., Moisen, G. G., and Frescino, T. S. (2022). GREGORY: a Modified Generalized Regression Estimator Approach to Estimating Forest Attributes in the Interior Western US. *Front. For. Glob. Change* 4:763414. doi: 10.3389/ffgc.2021.763414

Xie, Y., Cheng, J., and Tan, X. (2021). *DT: A Wrapper of the JavaScript Library 'DataTables'. R package version 0.18*. Available Online at: https://CRAN.R-project.org/package=DT (accessed July 1, 2021).

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

Zou, H., and Hastie, T. (2005). Regularization and variable selection *via* the elastic net. *J. R. Stat. Soc. Ser. B* 67, 301–320. doi: 10.1093/brain/awv075

Conflict of Interest: GW is 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Forests and Global Change

DATA REPORT
published: 11 July 2022
doi: 10.3389/ffgc.2022.769268

RegRake: A Web-Based Application for Custom Small Area Estimation and Mapping of Forest Survey Data With Regularized Raking

Todd A. Schroeder[1*†], *Nicholas N. Nagle*[2†] *and Joseph M. McCollum*[1]

[1] United States Department of Agriculture (USDA) Forest Service, Southern Research Station, Knoxville, TN, United States,
[2] Department of Geography & Sustainability, University of Tennessee, Knoxville, Knoxville, TN, United States

Keywords: small area estimation (SAE), regularized raking estimator, Forest Inventory and Analysis (FIA), R Shiny application, expansion factor map

OPEN ACCESS

Edited by:
Steve Prisley,
National Council for Air and Stream Improvement, Inc. (NCASI), United States

Reviewed by:
Edson Vidal,
University of São Paulo, Brazil
Philip Radtke,
Virginia Tech, United States

****Correspondence:***
Todd A. Schroeder
todd.schroeder@usda.gov

†These authors share first authorship

Specialty section:
This article was submitted to Forest Management, a section of the journal Frontiers in Forests and Global Change

Received: *01 September 2021*
Accepted: *13 June 2022*
Published: *11 July 2022*

Citation:
Schroeder TA, Nagle NN and McCollum JM (2022) RegRake: A Web-Based Application for Custom Small Area Estimation and Mapping of Forest Survey Data With Regularized Raking. Front. For. Glob. Change 5:769268. doi: 10.3389/ffgc.2022.769268

INTRODUCTION

The U.S. Forest Service, Forest Inventory and Analysis (FIA) program manages and implements a design-based network of permanent sample plots that are used to derive a suite of estimates describing the current extent, status and condition of the nation's forest resources (Bechtold and Patterson, 2005). Like many national forest inventory (NFI) systems, the FIA sample is designed for strategic-level estimation of forest characteristics to meet broad scale monitoring and reporting requirements. FIA produces official estimates for entire states or multiple county areas (referred to as survey units), which are large enough to guarantee sufficient sample sizes for direct estimation (USDA, 2008). Users of FIA data often desire estimates for smaller areas (e.g., single counties, watersheds, burn perimeters, etc.) however, in most cases these smaller areas have insufficient sampling densities to support direct estimation alone. In addition, the expansion factors reported by FIA (Bechtold and Patterson, 2005) have been designed to estimate the area of forest within a survey unit, and thus are not appropriate for use on small areas and may not be suitable for use with all forest attributes (e.g., basal area, volume, biomass, etc.). To overcome these limitations, we use a previously published regularized raking algorithm (Nagle et al., 2019) to develop spatial expansion factors which can be combined with FIA plot data to derive statistically valid estimates in areas that are too small to support direct estimation with FIA data alone.

One advantage of regularized raking is that instead of producing a single set of expansion factors representing a plot's contribution to the survey unit (such as those published by FIA), a wall-to-wall map of expansion factors is produced, which describes the probability that any location can be represented by a particular FIA plot. This map of design weights facilitates mapping of survey estimates for small areas and allows post-stratified estimates to be derived for any forest attribute in the FIA database. Furthermore, because the expansion factors are map-based they can be delivered and used on the backend of web-based applications, allowing end users the ability to interactively derive small area estimates for specific areas of interest without having to directly interact with the FIA database. To demonstrate the potential utility of this approach we present RegRake, a new, web-based application that uses pre-developed expansion factor maps to support on-the-fly estimation of FIA forest attributes for user defined small areas in the state of South Carolina, USA. In this article we briefly describe the regularized raking algorithm and ancillary input data used to produce the small area expansion factor maps. We also use RegRake to develop a series of raster and vector-based forest attribute estimates to help evaluate the amount of wood supply surrounding a proposed biomass energy plant.

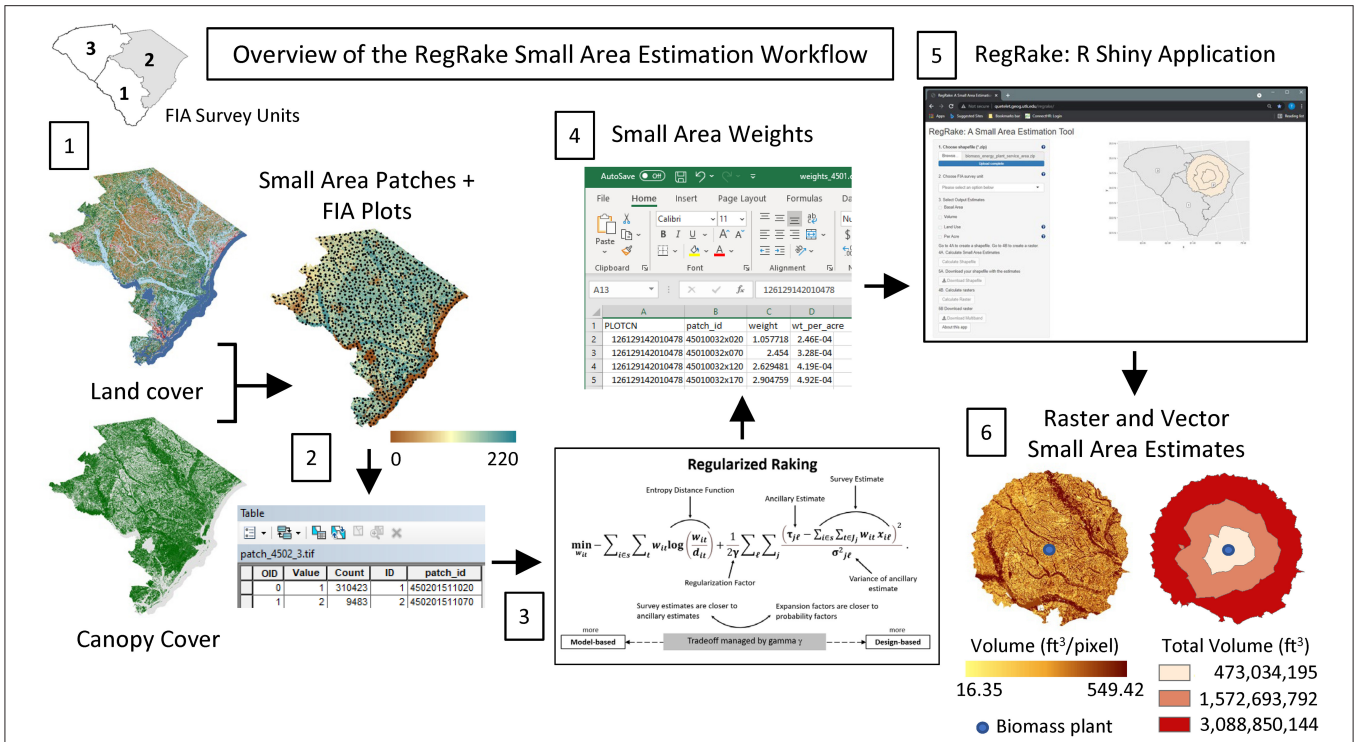


FIGURE 1 | Overview of the six main steps in the RegRake small area estimation workflow, which include: 1. Combining maps of land cover and canopy cover to produce small area patches; 2. Assigning FIA plot identifiers to the small area patches; 3. Running the regularized raking algorithm; 4. Developing a table of small area expansion factors; 5. Using the RegRake application to combine small area patches and small area expansion factors for user supplied GIS shapefile; and 6. Outputting small area estimates in raster and vector format.

SURVEY ESTIMATION WITH REGULARIZED RAKING

First, we provide a brief overview of the six main steps (shown in **Figure 1**) involved with developing custom small area estimates with the RegRake R (R Core Team, 2020) Shiny application (Chang et al., 2020). The reader is assumed to have a basic understanding of forest survey statistics (Köhl, 2004), including the use of sample plots and estimators using survey weights or expansion factors (e.g., Horvitz-Thompson, or HT; Cochran, 2007; Thompson, 2012) to derive population totals (e.g., means and variances). As a strategic-level inventory FIA uses spatially balanced, randomly selected sample plots to derive estimates for a variety of forest attributes, however at a density of one plot per 2,400 ha the sampling intensity requires the use of multi-county areas (referred to as survey units) to derive valid statistical estimates for reporting (see **Figure 1** for the three FIA survey units in SC). Expansion factors (or survey weights) published by FIA are designed for large-area estimation of forest area at the scale of the survey unit, therefore, to expand the use of FIA data to smaller areas (or regions that can't support direct estimation with plot data alone) and other attributes, we use a modification of the dasymetric mapping technique (Nagle et al., 2014) known as "regularized raking" (Nagle et al., 2019). Regularized raking develops a new set of expansion factors, w_{it} for each survey unit by matching each FIA plot (i) to a map of homogenous small area patches (t). Unlike traditional expansion factors that produce a single weight representing each plot's contribution to the survey unit (or in the case of FIA the number of acres each plot represents in the entire sample population), our approach results in a map describing the probability that each pixel or patch can be represented by a particular FIA plot. This expansion factor map, which is the main output of the regularized raking approach, can be used to produce both small area estimates and wall-to-wall maps for every attribute in the FIA database.

To develop new expansion factor maps for SC, Nagle et al. (2019) combined land cover and tree canopy cover maps (binned into 11 and 20 classes, respectively) from the 2011 National Land Cover Database (Homer et al., 2015) to form a series of up to 220 homogenous patches for each county in SC (**Figure 1**, step 1). Next, they developed predictive models for basal area and volume within each patch. Finally, they used these predictive values as ancillary information and developed expansion factors for FIA plots sampled between 2007 and 2011 (Cycle 7) using the regularized raking estimator. Both Deville et al. (1993) and Nagle et al. (2019) discuss the use of unregularized raking and calibration to adjust survey weights to match population totals derived from ancillary data. This process, also known as post-stratification, is currently used by FIA to calibrate survey design weights to land- and canopy cover maps, however this is done at the survey unit level. This approach cannot be directly applied to the large volume of ancillary data considered here for several

reasons. First, ancillary data derived from predictive models violate the assumption of raking and calibration that ancillary data are perfectly known. Second, when there are large numbers of ancillary data sets, raking and calibration algorithms often produce erratic and unreliable expansion factors, or the raking algorithm can even fail to converge. The regularized raking algorithm of Nagle et al. (2019) provides a feasible solution as it estimates the expansion factors w_{it} by solving the minimization problem (shown in **Figure 1**, step 3):

$$\min_{w_{it}} -\sum_{i\in s}\sum_{t} w_{it}\log\left(\frac{w_{it}}{d_{it}}\right)+\frac{1}{2\gamma}\sum_{\ell}\sum_{j}\frac{\left(\tau_{j\ell}-\sum_{i\in s}\sum_{t\in J_j} w_{it}x_{i\ell}\right)^2}{\sigma^2{}_{j\ell}} \tag{1}$$

where, $x_{i\ell}$ are the survey data for plot i and attribute ℓ, and the ancillary data for region j and attribute ℓ has the estimated value $\tau_{j\mathrm{l}}$ and variance $\sigma^2_{j\mathrm{l}}$, and $\mathrm{d}_{i\mathrm{t}}$ are prior weights determined by the sample design probabilities.

Deville and Särndal (1992) showed that the only difference between raking and the generalized regression estimator (or GREG) is the use of the entropy distance function (i.e., the logarithmic term in Equation 1) instead of the chi-square function. Although GREG is a commonly used model-assisted approach that can produce lower squared errors than the more generalized raking approach, it can produce negative design weights which can be problematic. In addition to being asymptotically design unbiased (Guggemos and Tille, 2010), the output of our generalized raking estimator (shown in **Figure 1**, step 4) is a set of strictly positive expansion factors w_{it} in units of acres of patch t represented by plot i. Since the weights are normalized by the small-area's size $\frac{w_{it}}{\sum_t w_{it}}$, they form a vector representing the probability density across all the samples found in each small patch, allowing for broader use with all attributes in the FIA database.

Calibrating on too many ancillary variables can result in erratic expansion weights or non-convergence (in the case of raking) or negative weights (in the case of GREG). Negative weights are problematic for survey estimation, especially for agencies that publish them, thus, to ensure weights stay positive and to avoid overfitting a regularization approach is employed. Inspired by ridge GREG and LASSO regression (McConville et al., 2017), we use a global regularization parameter gamma (γ in Equation 1), which allows the raking estimator to converge when the design weights "approximately" fit the ancillary data. Essentially the regularization parameter strikes a balance between producing expansion factors that closely match the unbiased design weights vs. finding factors that closely match the ancillary totals. While this tradeoff tends to sacrifice some small amount of finite-sample bias in the predictions it significantly reduces the overall mean square error. One challenge is finding a suitable value for γ. In the limit as γ gets large, the resulting weights will match the survey-unit HT weights without regard for the patch-level data. At the other extreme, as γ approaches zero, the resulting weights are a purely model-based estimate without regard to the sample design. Here, a cross-validation procedure was used to find γ. This process (described in full in Nagle et al., 2019) involved running the regularized raking algorithm on a series of simulated auxiliary totals, then using the resulting expansion factors to predict forest volume. Errors from these simulations were then compared across a range of regularization values and the γ with the lowest mean square error was selected. We recognize our use of the survey data to determine γ is considered endogenous, and while these approaches have been shown to be unbiased (Breidt and Opsomer, 2008) more work is needed to determine their legitimacy for use in survey estimation with FIA data. Lastly, to proportionally assign the errors to the quality of the various ancillary data sets we set the denominator $\sigma^2_{j\mathrm{l}}$ in Equation (1) to the variance of the ancillary total ($\tau_{j\mathrm{l}}$) as proposed in Nagle et al. (2014).

DELIVERING CUSTOM SMALL AREA ESTIMATES VIA WEB-BASED APPLICATION

Although the regularized raking equation can produce a map of small area expansion factors, these weights (in acres) can also be stored in tabular form once they have been merged with the FIA plot identifiers. This table of small area weights (shown in **Figure 1**, step 4) can then be used with the map of small area patches (**Figure 1**, step 2) to produce small area estimates. Here, instead of delivering these inputs as separate products, we opt to distribute them on the backend of the RegRake R Shiny application (shown in step 5, **Figure 1**), which can combine both inputs on the fly to produce vector and raster-based estimates for user defined areas of interest (shown in step 6, **Figure 1**). Disseminating the small area weights and patch raster map *via* the R Shiny application simplifies the process of developing small area estimates, giving users the ability to derive small area estimates by simply uploading a polygon shapefile into the web-based interface.[1] Although the small area weights developed with regularized raking can be used to develop small area estimates for any attribute in the FIA database, currently RegRake only supports estimation of a limited number of key variables including total basal area of live trees >1.0 inch (based on the FIA condition variable, BALIVE in ft^2/acre), total net volume of wood in the merchantable stem of sample trees > 5.0 inches (based on the FIA tree variable, VOLCFNET in ft^3), and total acres of Forest, Agriculture, Developed and Other land use classes (derived by simplifying the FIA condition variable LAND_USE_SRS according to the cross-walk table found in the RegRake user's guide available at the above url). In addition, RegRake also offers the option to run these estimates on a per acre basis. For basal area and volume, the per acre estimates are derived by dividing the estimates of total basal area and total volume by the estimated number of forest acres in the user's polygon shapefile, while for land use, the per acre estimates are reported as a percentage of the polygon's area found in each land use class (summing to 100% across all land uses). In addition to these tabular

[1] http://quetelet.geog.utk.edu/regrake/

TABLE 1 | Population totals and per acre estimates for FIA basal area, net volume and land use variables derived for a series of concentric rings representing driving distances to a hypothetical biomass energy plant.

Distance to biomass plant (miles)	0–20	20–40	40–80
ACRES_UNADJ	480,391	1,565,802	2,694,707
BALIVE_TOT (ft^2)	26,744,795	88,923,486	175,528,455
VOL_TOT (ft^3)	473,034,195	1,572,693,792	3,088,850,144
FOREST_TOT (acres)	283,896	945,211	1,890,851
AG_TOT (acres)	126,524	371,062	428,540
URBAN_TOT (acres)	61,831	223,308	329,006
OTHER_TOT (acres)	8,140	26,221	46,309
BALIVEperAC (ft^2/acre)	94.21	94.08	92.83
VOLperAC (ft^3/acre)	1,666.23	1,663.85	1,633.58
FOREST_percent (%)	0.59	0.60	0.70
AG_percent (%)	0.26	0.24	0.16
URBAN_percent (%)	0.13	0.14	0.12
OTHER_percent (%)	0.02	0.02	0.02

and vector-based estimates, RegRake can also produce raster-based estimates, which represent the estimated amount of each variable found in each 30 m pixel (matching the resolution of the patch raster map). Therefore, when the per pixel values are multiplied by the area represented by all the pixels in each volume, basal area and land use bin and summed, the resulting values match the estimates reported in the RegRake tabular output.

As an example, we used RegRake to derive tabular, vector and raster-based small area estimates for a set of concentric rings representing various driving distances from a proposed biomass energy plant. First, the individual files making up the polygon shapefile are zipped and uploaded into the RegRake application (shown in step 5, **Figure 1**). Note, shapefiles must be in polygon format and can include one or many polygons in the same file. Here, our shapefile contains 3 polygons representing driving distances of 0–20, 20–40, and 40–80 miles from the proposed biomass energy plant. We recommend that polygons be at least 10,000 acres to ensure enough FIA plots are used to produce valid estimates. We also note that if the uploaded shapefile overlaps multiple survey units, each one must be run individually and later combined to produce estimates for the full area. After running the population totals and per acre estimates the results appear directly in the table tab of the RegRake application. These estimates (shown in **Table 1**) can then be copied and pasted into a spreadsheet or downloaded as a new shapefile with the results appended to the attribute table (e.g., the shapefile showing total volume for the three driving distances in our example is shown in **Figure 1**, step 6). As a final step the user can also develop raster-based estimates, which can be downloaded as a single or multi-band file depending on the number of variables selected (e.g., the raster-based total volume map for our example shapefile is shown in **Figure 1**, step 6). In this hypothetical example, a plant manager could overlay the raster total volume estimates with other GIS data sets (e.g., roads, land use, protected areas, ownership, etc.) to help pinpoint areas where supply is plentiful, allowing cost estimates associated with accessing and delivering the necessary raw materials to be considered when evaluating potential sites for building a new bioenergy plant.

ADVANTAGES AND FUTURE OPPORTUNITIES OF REGULARIZED RAKING

In this article we describe an approach for using a web-based, R Shiny application called RegRake to deliver a new set of expansion factors that can be used with U.S. Forest Service FIA data to develop on the fly, custom small area estimates for several forest attributes in the state of South Carolina, USA. The regularized raking estimator used to develop these new expansion factors has several appealing qualities that help facilitate the development of small area estimates. These include automated calibration of survey design weights using multiple ancillary data sets that can be applied to all FIA variables (instead of just 1 variable, as is common with the GREG estimator). The new, strictly positive expansion factors also produce consistent spatial and tabular estimates, that scale seamlessly across domains of interest, and which maintain relatively low levels of uncertainty despite the tendency for variance to increase when small area estimates are made with FIA data alone. Although RegRake is only available in South Carolina we anticipate adding new locations as the requisite expansion factor maps are developed for other states across the southeastern U.S. We also plan to add other attributes and estimates of uncertainty, giving users even greater flexibility to develop and evaluate small area estimates for a more robust suite of forest characteristics for their area of interest.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: forest inventory data - https://apps.fs.usda.gov/fia/datamart/ RegRake R Shiny Application - http://quetelet.geog.utk.edu/regrake/.

AUTHOR CONTRIBUTIONS

NN and TS: conceptualization and supervision. NN: methodology. NN and JM: software, data curation, R programming, and writing—review and editing. TS: writing—original draft, visualization, project administration, and funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

Funding for this research was provided by the U.S. Forest Service, Southern Research Station, and Forest Inventory and Analysis (FIA) program (*via* agreement 17-CR-11330145-057).

REFERENCES

Bechtold, W. A., and Patterson, P. L. (2005). "The enhanced forest inventory and analysis program—National sampling design and estimation procedures", in *General Technical Report SRS-80* (Asheville, NC: US Department of Agriculture, Forest Service, Southern Research Station).

Breidt, F. J., and Opsomer, J. D. (2008). Endogenous post-stratification in surveys: classifying with a sample-fitted model. *Ann. Stat.* 36, 403–427. doi: 10.1214/009053607000000703

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., and McPherson, J. (2020). *shiny: Web Application Framework for R. R package Version* 1.5.0. Available online at: https://shiny.rstudio.com/reference/shiny/1.4.0/shiny-package.html

Cochran, W. G. (2007). *Sampling Techniques, 3rd Edn*. Hoboken, NJ: John Wiley & Sons.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *J. Am. Stat.* 87, 376–382. doi: 10.1080/01621459.1992.10475217

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *J. Am. Stat.* 88, 1013–1020. doi: 10.1080/01621459.1993.10476369

Guggemos, F., and Tille, Y. (2010). Penalized calibration in survey sampling: design-based estimation assisted by mixed models. *J. Stat. Plan. Inference*. 140, 3199–3212. doi: 10.1016/j.jspi.2010.04.010

Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., et al. (2015). Completion of the 2011 National Land Cover Database for the Conterminous United States representing a decade of land cover change information. Photogramm. *Eng. Rem. S.* 81, 345–354. Available online at https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NERL&dirEntryId=309950

Köhl, M. (2004). Inventory | forest inventory and monitoring. *Enc. For. Sci.* 403–409. doi: 10.1016/B0-12-145160-7/00154-X

McConville, K. S., Breidt, F. J., Lee, T. C. M., and Moisen, G. G. (2017). Model-assisted survey regression estimation with the LASSO. *J. Surv. Stat. Methodol.* 5, 131–158. doi: 10.1093/jssam/smw041

Nagle, N. N., Buttenfield, B. P., Leyk, S., and Spielman, S. (2014). Dasymetric modeling and uncertainty. *Ann. Assoc. Am. Geogr.* 104, 80–95. doi: 10.1080/00045608.2013.843439

Nagle, N. N., Schroeder, T. A., and Rose, B. (2019). A regularized raking estimator for small-area mapping from forest inventory surveys. *Forests*. 10, 111045. doi: 10.3390/f10111045

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Thompson, S. K. (2012). *Sampling, 3rd Edn*. Hoboken, NJ: John Wiley & Sons.

USDA (2008). *Chapter 10 Operational Procedures*. United States Department of Agriculture (USDA), Forest Service Handbook No. 4809.11, 24. Available online at: https://www.fs.fed.us/dirindexhome/fsh/4809.11/4809.11_10.doc (accessed June 24, 2022).

Author Disclaimer: This study was subject to agency review and approved for publication. Any use of product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. This paper was written and prepared by U.S. Government employees on official time, and is therefore, in the public domain and not subject to copyright.

Conflict of Interest: The 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

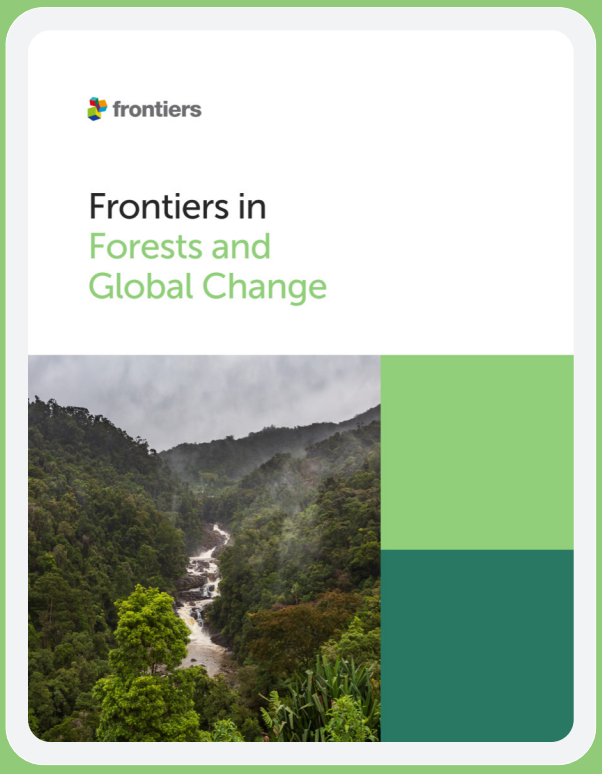
Frontiers in
Forests and Global Change

Informs and promotes sustainable management of the world's forests

An innovative journal that places forests at the forefront of attention for scientists, policy makers and the public. It advances our understanding of how forests 'work', spanning from molecules to ecosystems to the biosphere.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact