Canadian Journal of
Forest Research

Research Article

Model-assisted estimation of domain totals, areas, and densities in two-stage sample survey designs

Hans-Erik Andersen[a], **Göran Ståhl**[b], **Bruce D. Cook**[c], **Douglas C. Morton**[c], **and Andrew O. Finley**[d]

[a]USDA Forest Service, Pacific Northwest Research Station, Seattle, WA, USA; [b]Department of Forest Resource Management, Swedish University of Agricultural Sciences, Umeå, Sweden; [c]Goddard Space Flight Center, National Aeronautics and Space Administration, Greenbelt, MD, USA; [d]Department of Forestry, Michigan State University, East Lansing, MI, USA

Corresponding author: **Hans-Erik Andersen** (email: hans.andersen@usda.gov)

Abstract

Model-assisted, two-stage forest survey sampling designs provide a means to combine airborne remote sensing data, collected in a sampling mode, with field plot data to increase the precision of national forest inventory estimates, while maintaining important properties of design-based inventories, such as unbiased estimation and quantification of uncertainty. In this study, we present a comprehensive set of model-assisted estimators for domain-level attributes in a two-stage sampling design, including new estimators for densities, and compare the performance of these estimators with standard poststratified estimators. Simulation was used to assess the statistical properties (bias, variability) of these estimators, with both simple random and systematic sampling configurations, and indicated that (1) all estimators were generally unbiased and (2) the use of lidar in a sampling mode increased the precision of the estimators at all assessed field sampling intensities, with particularly marked increases in precision at lower field sampling intensities. Variance estimators are generally unbiased for model-assisted estimators without poststratification, while variance estimators for model-assisted estimators with poststratification were increasingly biased as field sampling intensity decreased. In general, these results indicate that airborne remote sensing data, collected as an intermediate level of sampling, can be used to increase the efficiency of national forest inventories in remote regions.

Key words: inventory, biomass, carbon, lidar, sampling

1. Introduction

Reliable and credible information on forest status and change is needed to support effective natural resource management and policy. National forest inventory (NFI) programs are usually the primary source for this type of information, since, ideally, they are designed to provide accurate (unbiased) and precise estimates at regional and national scales. For example, NFI data are used to support global forest monitoring (FAO 2016), national scale greenhouse gas monitoring (EPA 2023), and emerging markets for trading of forest carbon offset credits (Marland et al. 2017). The NFI is used to characterize various attributes (e.g., biomass, area) of the population for a multitude of domains (e.g., forest type), where estimates take the form of domain totals (e.g., total biomass of trees in forestland), areas (e.g., total forestland area), or per-unit-area densities (average tree biomass per hectare of forestland). A large network of field inventory plots is the basis for most NFI programs, and plots within homogeneous regions, or strata, are often grouped (and weighted by stratum area) to increase the precision of inventory estimates. Increasing interest in more accurate estimation across spatial scales ranging from smaller regions (e.g., county, parcel, or management unit) (Dettmann et al. 2022) to vast, highly remote regions (tropical rainforests, high-latitude boreal forests), as well as other societal drivers such the need to verify forest carbon offset credits in carbon markets, has motivated the international forest monitoring community to more effectively leverage remote sensing in inventory sampling designs, including the use of high-resolution airborne remote sensing as an intermediate level of sampling in multi-stage sampling designs. In recent years, more advanced approaches to estimation have been proposed that can more fully leverage more sophisticated models and the rich information from both satellite and airborne remote sensing in the context of NFI programs (McRoberts et al. 2014; Ståhl et al. 2016; Kangas et al. 2018; Lister et al. 2020; Westfall et al. 2022). For example, both model-assisted (Breidt and Opsomer 2017; McConville et al. 2017; Magnussen et al. 2018) and model-based techniques (McRoberts 2010) have been developed to leverage wall-to-wall remote sensing data in forest inventory.

The use of high-resolution, airborne lidar strip sampling to support forest inventories has been the focus of several recent studies in Europe and North America, where a variety of estimation approaches have been developed and evaluated, including model-assisted (Gregoire et al. 2011; Strunk et al. 2014; Ringvall et al. 2016), model-based (Saarela et al. 2016), and hybrid estimation (Ståhl et al. 2011, 2016). Gobakken et al. (2012) compared several model-dependent and model-assisted estimators in a two-stage design using lidar sampling in the Hedmark region of Norway. Ringvall et al. (2016) developed a model-assisted estimator for two-stage lidar sample surveys that addressed several significant limitations of lidar sampling in the forest inventory context, including the use of a ratio estimator to account for varying strip lengths, and poststratification to further improve precision. This estimator has since been applied to repeat collections of lidar sampling data to support monitoring of forest change in Norway (Strîmbu et al. 2017). In the context of the US NFI in interior Alaska, model-assisted (Andersen et al. 2009*a*), hierarchical Bayesian coregionalization model-based (Babcock et al. 2018) and hybrid (Ene et al. 2018) estimation approaches have been proposed as a means to combine information from NFI plots, a lidar strip sample, and wall-to-wall layers. Among these alternative approaches to integrate multiple levels of sampling, model-assisted estimation is likely best-suited for the NFI context, since it retains the statistical properties of design-based inference (approximate unbiasedness of estimators) (Särndal et al. 2003), and therefore can provide estimates for specific, remote regions (e.g., interior Alaska) that are compatible with NFI estimates for other regions of the country where a standard design and estimation approach (such as poststratification) are used.

While these more complex, multi-level sampling designs can often increase the efficiency of the inventory (higher precision at specified cost, or level of sampling effort), they often require the development of new estimators for inventory attributes, as well as new variance estimators to quantify uncertainty (i.e., variance) of the estimators. In particular, while approaches to classification (Andersen 2009; Shoot et al. 2021) and estimation (McConville et al. 2017) of domains in the NFI context have been recently investigated, no previous studies have developed ratio estimators for densities in a two-stage survey design framework. Assessing the statistical properties of these new estimators requires a simulation framework, where the true properties of the population are fixed and known and the sampling distributions of the estimators can be developed over many simulated samples. Assessment of the estimators via simulation is necessary to establish the reliability and defensibility of the estimators, especially in the context of an NFI.

The objectives of this study are to present a set of model-assisted estimators for all types of NFI estimates required to report current status (domain totals, areas, and densities) that can incorporate three levels of information available in a poststratified, two-stage sampling design: (1) a sparse sample of field inventory plots, (2) a dense strip sample of airborne lidar, and (3) a wall-to-wall stratification layer. This comprehensive set of estimators includes those presented in previous studies (Ringvall et al. 2016), as well as new *ratio-of-ratios* (*RoR*) estimators for estimating densities.

2. Methods

2.1. Case study: implementation of the national forest inventory in interior Alaska, USA

We focus on interior Alaska as an example of a region where an alternative sampling design, and corresponding estimators, are needed to support the NFI program. The US Forest Inventory and Analysis (FIA) program—the NFI for the United States—is mandated by the US Congress to assess current status and trends for all forest lands of the United States (USDA Forest Service 2016). The standard sampling framework for FIA in the conterminous United States, coastal Alaska, and Pacific islands is a systematic, unaligned grid of field plots established with a sampling intensity of 1 plot per 2428 ha. Interior Alaska is a vast region of high-latitude boreal forests (approx. 46 million hectares of forestland) with virtually no transportation infrastructure (virtually all plots are accessed via helicopter), and for this reason, FIA has implemented a modified sampling design in interior Alaska using a reduced sampling intensity for field plots (1 plot per 12 140 ha—1/5th the standard intensity—collected on a systematic grid), supplemented with high-resolution airborne imagery acquired by National Aeronautics and Space Administration Goddard's Lidar, Hyperspectral, and Thermal (G-LiHT) Airborne Imager (Cook et al. 2013) in a strip sampling mode (Cahoon and Baer 2022) covering every FIA plot. The FIA field inventory of interior Alaska was initiated within the Tanana inventory unit (13.5 million ha) and carried out over a 5-year period (2014–2018), and the G-LiHT collection was carried out in 2014 (see Fig. 1).

2.2. Standard FIA sampling design and estimation

The sample-based estimators of population totals, areas, and densities (and corresponding variance estimators) used in standard FIA reporting assume that the plots are collected as a simple random sample (*SRS*), and the precision of these estimators is further improved through poststratification, as documented in Bechtold and Patterson (2005). The poststrata used in FIA are generally developed based on available land cover maps, as well as ownership and other GIS layers.

2.2.1. Poststratified estimation of a domain total

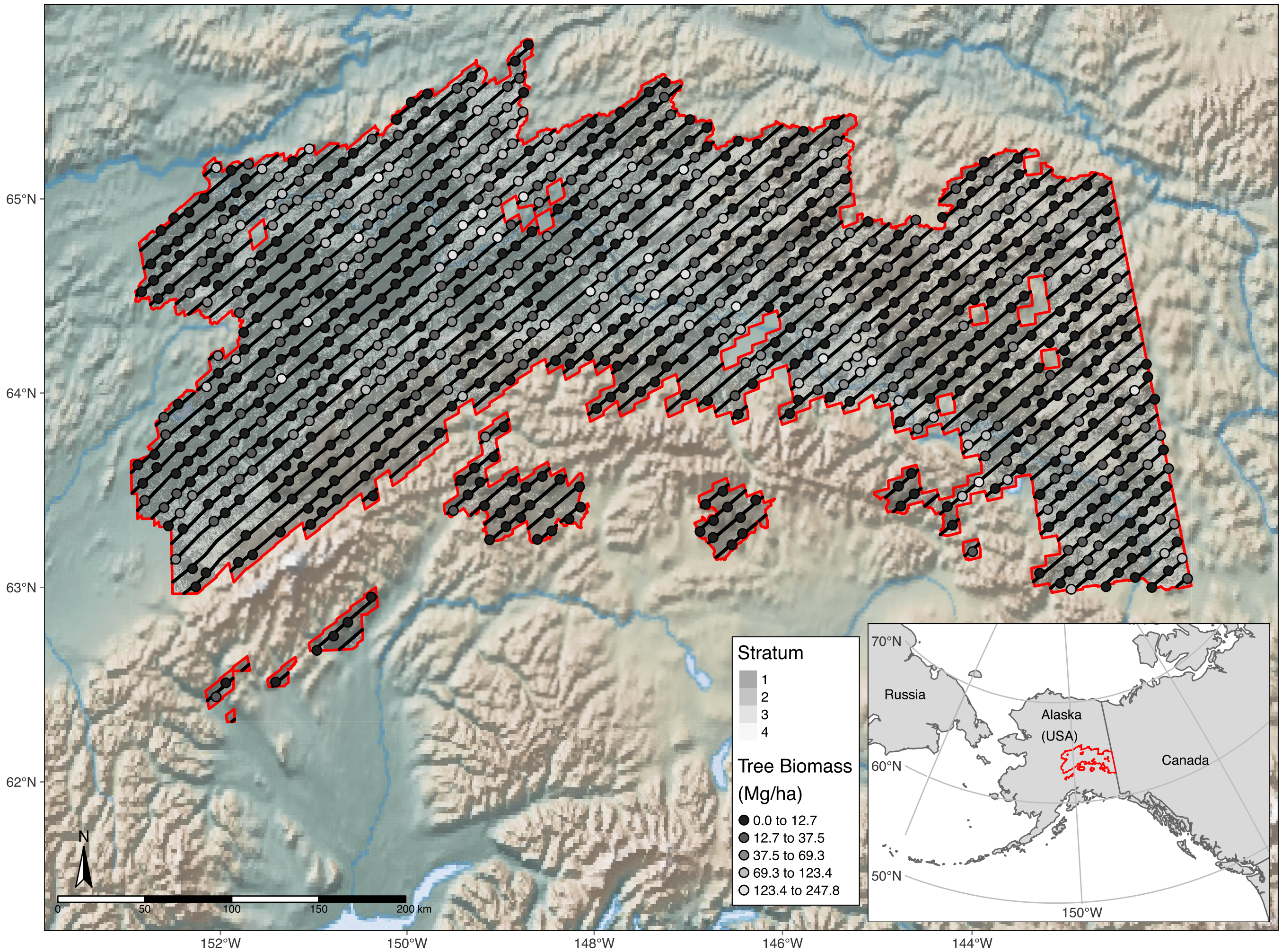
Estimates of population totals, such as tree biomass, or carbon, are usually primary attributes of interest in an NFI. In the standard estimation framework of FIA, these values are provided by the following estimators (including variance):

$$(1)\qquad \bar{t}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$$

$$(2)\qquad \hat{t}_{SRS,PS} = A_T \sum_{h=1}^{H} W_h \bar{t}_h$$

where y_{hi} is the plot-level value (per ha) for an inventory attribute of interest (e.g., tree biomass, carbon, volume, etc.)

Fig. 1. Tanana inventory unit, interior Alaska, United States. Black lines indicate location of airborne Goddard's Lidar, Hyperspectral, and Thermal (G-LiHT) flight lines. Dots indicate (approximate) location of forest inventory and analysis (FIA) field plots (colored by biomass). Underlying raster indicates stratification, red outline indicates the domain sampled by G-LiHT airborne sensor. Figure was created using R version 4.4.1 and assembled from the following data sources: FIA field plots (https://apps.fs.usda.gov/fia/datamart/datamart.html) and G-LiHT data (http://gliht.gsfc.nasa.gov). Base map from Natural Earth (https://www.naturalearthdata.com).



on plot i in stratum h, n_h is the number of plots in stratum h, A_T is the total area of the population (ha), and W_h is the weight (area proportion) for stratum h, assumed to be known without error.

The variance estimator is:

$$(3)\qquad \widehat{V}\left(\hat{t}_{SRS,PS}\right) = A_T^2\left[\sum_{h=1}^{H} W_h n_h \widehat{V}\left(\bar{t}_h\right) + \sum_{h=1}^{H}\left(1-W_h\right)\frac{n_h}{n}\widehat{V}\left(\bar{t}_h\right)\right]$$

$$(4)\qquad \widehat{V}\left(\bar{t}_h\right) = \frac{\sum_{i=1}^{n_h} y_{hi}^2 - n_h \bar{t}_h^2}{n_h\left(n_h-1\right)}$$

2.2.2. Poststratified estimation of domain area

Estimation of domain area (e.g., total forestland area in hectares, etc.) is another primary output of an NFI, and are provided by the following estimators (including variance):

$$(5)\qquad \bar{a}_h = \frac{\sum_{i=1}^{n_h} a_{hi}}{n_h}$$

$$(6)\qquad \widehat{A}_{SRS,PS} = A_T\sum_{h=1}^{H} W_h \bar{a}_h$$

where a_{hi} is the proportion of plot i, in stratum h, within a given domain (e.g., forestland, forest type, etc.).

The variance estimator is:

$$(7)\qquad \widehat{V}\left(\widehat{A}_{SRS,PS}\right) = A_T^2\left[\sum_{h=1}^{H} W_h n_h \widehat{V}\left(\bar{a}_h\right) + \sum_{h=1}^{H}\left(1-W_h\right)\frac{n_h}{n}\widehat{V}\left(\bar{a}_h\right)\right]$$

$$(8)\qquad \widehat{V}\left(\bar{a}_h\right) = \frac{\sum_{i=1}^{n_h} a_{hi}^2 - n_h \bar{a}_h^2}{n_h\left(n_h-1\right)}$$

2.2.3. Poststratified estimation of a domain density

The final category of estimates required in an NFI include densities for various inventory attributes on a per-unit-area basis. In the case where a density is estimated over the entire area of the population, the estimator is obtained by simply dividing the estimator of the total $\widehat{t}_{SRS,PS}$ by the known area of the population A_T, and the corresponding variance estimator is obtained by dividing the variance of the total by A_T^2. However, often a density is estimated for a domain (e.g., tree carbon per forested hectare, where the domain is forestland). Because both the numerator and denominator of these estimators are themselves estimators, they are considered *ratio* estimators. The poststratified estimator of domain density used in FIA is given by:

$$\widehat{D}_{SRS,PS} = \frac{\sum_{h=1}^{H} W_h \bar{t}_h}{\sum_{h=1}^{H} W_h \bar{a}_h} \tag{9}$$

with variance estimator:

$$\widehat{V}\left(\widehat{D}_{SRS,PS}\right) = \frac{1}{\widehat{A}_{SRS,PS}^2}\left[\widehat{V}\left(\widehat{t}_{SRS,PS}\right) + \widehat{D}_{SRS,PS}^2 \widehat{V}\left(\widehat{A}_{SRS,PS}\right) - 2\widehat{D}_{SRS,PS}\ Cov\left(\widehat{t}_{SRS,PS}, \widehat{A}_{SRS,PS}\right)\right] \tag{10}$$

$$Cov\left(\widehat{t}_{SRS,PS}, \widehat{A}_{SRS,PS}\right) = \frac{A_T^2}{n}\left[\sum_{h=1}^{H} W_h n_h Cov\left(\bar{t}_h, \bar{a}_h\right) + \sum_{h=1}^{H}\left(1 - W_h\right)\frac{n_h}{n} Cov\left(\bar{t}_h, \bar{a}_h\right)\right] \tag{11}$$

$$Cov\left(\bar{t}_h, \bar{a}_h\right) = \frac{\sum_{i=1}^{n_h} y_{hi} a_{hi} - n_h \bar{t}_h \bar{a}_h}{n_h\left(n_h - 1\right)} \tag{12}$$

2.3. Ratio estimators for model-assisted estimation in two-stage sampling designs

2.3.1. Sampling design and notation

The addition of another level of sampling, e.g., provided by high-resolution, airborne data, collected as a strip sample, requires a different set of domain-level estimators. The notation and formulas used to describe these estimators mainly follow the notation and formulas used by Ringvall et al. (2016). The study area is partitioned into M nonoverlapping strips, denoted as *PSU*s. These strips have unequal lengths and thus unequal areas, and the set U_1 is the population of *PSU*s. A first-stage sample, S_1, of *PSU*s is selected. Each *PSU* is partitioned into N_i grid-cells (*SSU*s) with a size roughly corresponding to the size of the field plots. In the second stage, we randomly select a sample S_i of *SSU*s from each *PSU*. Subsampling within selected *PSU*s are invariant of which *PSU*s are selected in the first stage. The total number of grid-cells in the study area is $N = \sum_{i \in U_1} N_i$.

In inventory programs where the plot is actually a cluster of several subplots (as is the case with many NFIs, including FIA), a slight adjustment to the sampling framework can be used to maintain consistency in the spatial support (i.e., plot footprint) for sampling units at the different sampling stages. To this end, in this study N is the number of points on the 200 m hexagonal grid across the study area, N_i is the points on this grid along the ith *PSU*. The per-ha average is estimated for each attribute in N and then scaled up to the level of the population by multiplying by the population area, A_T. However, for the purposes of clarity and consistency, we maintain the more general notation used in Ringvall et al. (2016) in the following sections (it should be noted that the above adjustment does not change the results). Additional details related to the measurements collected at FIA plots and remote sensing plots in this study are provided in Section 3.1.

Several other details regarding the sampling design should be noted. First, in practice lidar strips are often allocated systematically, traversing the entire area of study, but we assume them to be randomly allocated from the point of view of the variance formulas. Likewise, field plots are often distributed systematically (i.e., at a regular spacing) within lidar strips, we assume them to be randomly allocated within strips, at a specified sampling intensity, from the point of view of the variance formulas. We assume that all the *SSU*s (FIA plots) fall entirely within a given *PSU*, that is, there are no border issues and no plots straddling a *PSU* boundary. As is the case with many standard NFI designs, the entire study area is divided into strata, and this map information is incorporated in the estimators through poststratification. It should also be noted that the lidar strip sampling is carried out independent of the stratification. In addition, the strata normally do not coincide with the domains of interest, since areas belonging to the domain potentially appear scattered across several strata and they cannot be perfectly identified on maps, but only in the field on the subplots. Prediction models for the variable of interest (e.g., biomass per ha), as well as the domain (e.g., forestland), are developed from the collected field data.

2.3.2. A ratio estimator of the domain total

With simple random sampling without replacement in both stages, a ratio (R) estimator of the variable of interest within a given domain (cf. Ringvall et al. 2016; eq. (5)) is

$$\widehat{t}_R = N\frac{\sum_{i=1}^{m} \widehat{t}_{ri}}{\sum_{i=1}^{m} N_i} \tag{13}$$

where m is the number of selected *PSU*s and $\widehat{t}_{ri}$ is the model-assisted estimator of total biomass in a given domain (where the r subscript denotes a regression estimator) in the ith *PSU* given as

$$\widehat{t}_{ri} = \sum_{k=1}^{N_i} \widehat{q}_k \widehat{y}_k + \frac{N_i}{n_i}\sum_{k=1}^{n_i} e_k \tag{14}$$

where n_i is the number of selected *SSU*s within *PSU* i, $\widehat{y}_k$ is a predicted value of the domain-level attribute of interest based on a regression (or similar) model within a grid cell/plot, $\widehat{q}_k$ is an indicator variable (e.g., based on logistic regression) predicting whether or not the grid cell/plot belongs to the domain (or a prediction of the proportion in a domain, e.g., Van den Boogaart and Tolosana-Delgado 2013), and e_k is a deviation between the measured and predicted plot-level value for the variable of interest in the domain. In cases where the domain membership is incorporated into the regression model predicting the variable of interest (e.g., biomass is always zero on "nonforest" plots), the $\widehat{q}_k$ term can be omitted. It should be noted that predictions of the target variable and domain membership are made at the scale of the entire plot. If predictions are made at subplot scale in the case of cluster plots, it is assumed that these predictions, and associated error terms, are averaged across subplots to obtain a plot-level value.

The variance estimator of $\widehat{t}_R$ is

$$(15)\quad \widehat{V}\left(\widehat{t}_R\right) = \frac{N^2}{\widehat{N}^2}\left[M^2\left(\frac{1}{m}-\frac{1}{M}\right)s_r^2 + \frac{M}{m}\sum_{S_1} N_i^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right)s_e^2\right]$$

where $s_r^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(\widehat{t}_{ri} - \widehat{R}N_i\right)^2$ with $\widehat{R} = \frac{\sum_{i=1}^{m}\widehat{t}_{ri}}{\sum_{i=1}^{m}N_i}$, $s_e^2 = \frac{1}{n_i-1}\sum_{k=1}^{n_i}(e_k - \bar{e}_i)^2$ with $\bar{e}_i = \frac{1}{n_i}\sum_{k=1}^{n_i}e_k$, and $\widehat{N} = (M/m)\sum_{i=1}^{m}N_i$. Additional details related to the derivation of this ratio estimator are provided in Appendix A1 of Ringvall et al. (2016).

2.3.3. A ratio estimator of the domain total in poststratified estimation

When m *PSU*s are selected with simple random sampling without replacement in the first stage and n_i units are selected with simple random sampling with replacement in the second stage (as before), an estimator of a stratum total is

$$(16)\quad \widehat{t}_{Rh} = N_h\frac{\sum_{i=1}^{m}\widehat{t}_{rhi}}{\sum_{i=1}^{m}N_{hi}}$$

with

$$(17)\quad \widehat{t}_{rhi} = \sum_{k=1}^{N_{hi}}\widehat{q}_{hk}\widehat{y}_{hk} + \frac{N_{hi}}{n_{hi}}\sum_{k=1}^{n_{hi}}e_{hk}$$

where the notation follows from the previously used notation, but refers to a specific stratum h. Note that e_{hk} involves either (1) prediction of domain membership for subplots and averaging across subplots, or (2) prediction of domain membership for the entire plot. In addition, all subplots are assumed to belong to the same stratum.

The poststratified estimator of the domain total is:

$$(18)\quad \widehat{t}_{R,PS} = \sum_{h=1}^{H}\widehat{t}_{Rh}$$

The variance estimator becomes (from Ringvall et al. 2016)

$$(19)\quad \widehat{V}\left(\widehat{t}_{Rh}\right) = \left(\frac{N_h}{\widehat{N}_h}\right)^2\left(M^2\left(\frac{1}{m}-\frac{1}{M}\right)s_{rh}^2 + \frac{M}{m}\sum_{S_1}\widehat{V}\left(\widehat{t}_{rhi}\right)\right)$$

where $s_{rh}^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(\widehat{t}_{rhi} - \widehat{R}_hN_{hi}\right)^2$ with $\widehat{R}_h = \frac{\sum_{S_I}\widehat{t}_{rhi}}{\sum_{S_I}N_{hi}}$ and $\widehat{N}_h = (M/m)\sum_{i=1}^{m}N_{hi}$. The within *PSU* variance $V\left(\widehat{t}_{rhi}\right)$ is estimated as

$$(20)\quad \widehat{V}\left(\widehat{t}_{rhi}\right) = N_{hi}^2\left(\frac{1}{n_{hi}}-\frac{1}{N_{hi}}\right)s_{e,h}^2$$

with $s_{e,h}^2 = \frac{1}{n_{hi}-1}\sum_{k=1}^{n_{hi}}(e_{hk}-\bar{e}_{hi})^2$ and $\bar{e}_{hi} = \frac{1}{n_{hi}}\sum_{k=1}^{n_{hi}}e_{hk}$.

Finally, the poststratified estimator of the domain total is $\widehat{t}_{R,PS} = \sum_{h=1}^{H}N_h\frac{\sum_{i=1}^{m}\widehat{t}_{rhi}}{\sum_{i=1}^{m}N_{hi}}$. The variance estimator is

$$(21)\quad \widehat{V}\left(\widehat{t}_{R,PS}\right) = \sum_{h=1}^{H}\widehat{V}\left(\widehat{t}_{Rh}\right) + \sum_{h=1}^{H}\sum_{g\neq h}\frac{N_h}{\widehat{N}_h}\frac{N_g}{\widehat{N}_g}M^2\left(\frac{1}{m}-\frac{1}{M}\right) \times \frac{\sum_{i=1}^{m}\left(\left(\widehat{t}_{rhi}-\widehat{R}_hN_{hi}\right)\left(\widehat{t}_{rgi}-\widehat{R}_gN_{gi}\right)\right)}{m-1}$$

with $\widehat{V}\left(\widehat{t}_{Rh}\right)$ given by eq. (19). In this case, the computations can be limited to those strata where the domain has a potential to occur. Note that because in our case sampling is not conducted independently between strata, the variance estimator (eq. (21)) must account for covariances among strata (i.e., the last term on the right hand side). Additional details related to the derivation of the poststratified ratio estimator are provided in appendix A2 of Ringvall et al. (2016).

2.3.4. A ratio estimator of the domain area

So far, the population total within the domain has been addressed. Following the previous notation and estimation principles, a domain area estimator (without poststratification) would be

$$(22)\quad \widehat{A}_R = N\frac{\sum_{i=1}^{m}\widehat{a}_{ri}}{\sum_{i=1}^{m}N_i}$$

where $\widehat{a}_{ri}$ is the model-assisted estimator of total domain area in the ith *PSU* given as

$$(23)\quad \widehat{a}_{ri} = \sum_{k=1}^{N_i}\widehat{q}_k\ a\_cell + \frac{N_i}{n_i}\sum_{k=1}^{n_i}e_k$$

where *a_cell* is the area of a grid-cell. The other notation is the same as previously, and notably e_k is a deviation between a measured and predicted plot/grid-cell level value for the proportion of the plot within the domain. Like for the target variable of interest, e_k should be the averaged difference across subplots, recalculated to correspond to the size of a grid-cell.

The variance estimator of $\widehat{A}_R$ is

$$(24)\quad \widehat{V}\left(\widehat{A}_R\right) = \frac{N^2}{\widehat{N}^2}\left[M^2\left(\frac{1}{m}-\frac{1}{M}\right)s_r^2 + \frac{M}{m}\sum_{S_1}N_i^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right)s_e^2\right]$$

where $s_r^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(\widehat{a}_{ri} - \widehat{R}N_i\right)^2$ with $\widehat{R} = \frac{\sum_{i=1}^{m}\widehat{a}_{ri}}{\sum_{i=1}^{m}N_i}$ and $s_e^2 = \frac{1}{n_i-1}\sum_{k=1}^{n_i}(e_k-\bar{e}_i)^2$ with $\bar{e}_i = \frac{1}{n_i}\sum_{k=1}^{n_i}e_k$.

2.3.5. A ratio estimator of the domain area in poststratified estimation

So far, the population total within the domain has been addressed. Following the previous notation and estimation principles, a domain area estimator (with poststratification) would be

$$(25)\quad \widehat{A}_{Rh} = N_h\frac{\sum_{i=1}^{m}\widehat{a}_{rhi}}{\sum_{i=1}^{m}N_{ih}}$$

where $\widehat{a}_{rhi}$ is the model-assisted estimator of total domain area, in stratum h, in the ith *PSU* given as

$$(26)\quad \widehat{a}_{rhi} = \sum_{k=1}^{N_{hi}}\widehat{q}_{hk}\ a\_cell + \frac{N_{hi}}{n_{hi}}\sum_{k=1}^{n_{hi}}e_{hk}$$

where *a_cell* is the area of a grid-cell.

The poststratified estimator of the total population area is:

$$(27)\quad \widehat{A}_{R,PS} = \sum_{h=1}^{H} \widehat{A}_{Rh}$$

with variance estimator

$$(28)\quad \widehat{V}\left(\widehat{A}_{Rh}\right) = \left(\frac{N_h}{\widehat{N}_h}\right)^2 \left(M^2\left(\frac{1}{m} - \frac{1}{M}\right)s_{rh}^2 + \frac{M}{m}\sum_{S_1}\widehat{V}\left(\widehat{a}_{rhi}\right)\right)$$

where $s_{rh}^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(\widehat{a}_{rhi} - \widehat{R}_{a,h}N_{hi}\right)^2$ and $\widehat{R}_{a,h} = \frac{\sum_{S_1}\widehat{a}_{rhi}}{\sum_{S_1}N_{hi}}$. $V\left(\widehat{t}_{rhi}\right)$ is estimated as

$$(29)\quad \widehat{V}\left(\widehat{a}_{rhi}\right) = N_{hi}^2\left(\frac{1}{n_{hi}} - \frac{1}{\widehat{N}_{hi}}\right)s_{e,h}^2$$

with $s_{e,h}^2 = \frac{1}{n_{hi}-1}\sum_{k=1}^{n_{hi}}(e_{hk} - \bar{e}_{hi})^2$ and $\bar{e}_{hi} = \frac{1}{n_{hi}}\sum_{k=1}^{n_{hi}}e_{hk}$.

A variance estimator for the total domain area is given by:

$$(30)\quad \widehat{V}\left(\widehat{A}_{R,PS}\right) = \sum_{h=1}^{H}\widehat{V}\left(\widehat{A}_{Rh}\right) + \sum_{h=1}^{H}\sum_{g\neq h}\frac{N_h}{\widehat{N}_h}\frac{N_g}{\widehat{N}_g}M^2\left(\frac{1}{m} - \frac{1}{M}\right) \times \frac{\sum_{i=1}^{m}\left(\left(\widehat{a}_{rhi} - \widehat{R}_{a,h}N_{hi}\right)\left(\widehat{a}_{rgi} - \widehat{R}_{a,g}N_{gi}\right)\right)}{m-1}$$

with $\widehat{V}\left(\widehat{A}_{Rh}\right)$ given by eq. (28).

2.3.6. A *ratio-of-ratios* estimator of the domain density

In the context of forest inventory, we are often also interested in domain densities, in addition to domain totals and areas. Because the form of the estimator of domain density is a ratio estimator, where the numerator and denominator *are themselves ratio estimators*, we term the form of this new estimator as *RoR* estimator.

$$(31)\quad \widehat{D}_{RoR} = \frac{\widehat{t}_R}{\widehat{A}_R} = \frac{N\frac{\sum_{i=1}^{m}\widehat{t}_{ir}}{\sum_{i=1}^{m}N_i}}{N\frac{\sum_{i=1}^{m}\widehat{a}_{ir}}{\sum_{i=1}^{m}N_i}} = \frac{\sum_{i=1}^{m}\widehat{t}_{ir}}{\sum_{i=1}^{m}\widehat{a}_{ir}}$$

where $\widehat{t}_R$ is the ratio domain biomass estimator and $\widehat{A}_R$ is the ratio domain area estimator. Introducing the new variables $u_i = \widehat{t}_i - \widehat{D}_{RoR}\widehat{a}_i$ and $v_{ik} = e_{ik}^y - \widehat{D}_{RoR}e_{ik}^a$ the variance estimator for $\widehat{D}_{RoR}$ is

$$(32)\quad \widehat{V}\left(\widehat{D}_{RoR}\right) = \frac{1}{\widehat{A^*}^2}\left[M^2\left(\frac{1}{m} - \frac{1}{M}\right)s_u^2 + \frac{M}{m}\sum_{S_1}N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)s_{vi}^2\right]$$

where $s_u^2 = \frac{1}{m-1}\sum_{i=1}^{m}(u_i - \bar{u})^2$, $s_{vi}^2 = \frac{1}{n_i-1}\sum_{k=1}^{n_i}(v_{ik} - \bar{v}_i)^2$, and the term $\widehat{A}^* = M\left(\frac{\sum_{i=1}^{m}\widehat{a}_{ir}}{m}\right)$, i.e., the average estimated area from the sampled strips, times the total number of strips in the population. Details regarding the derivation of $\widehat{V}\left(\widehat{D}_{RoR}\right)$ are provided in Appendix A.

2.3.7. A *ratio-of-ratios* estimator of the domain density with poststratification

The variance of the *RoR* estimator takes on a more complex form in a poststratified (or prestratified) setting, since in the case of poststratification we must account for dependencies (i.e., covariance) between the estimators for different strata, since a single survey line may cross over several strata.

The *RoR* estimator with poststratification is:

$$(33)\quad \widehat{D}_{RoR,PS} = \frac{\widehat{t}_{R,PS}}{\widehat{A}_{R,PS}} = \frac{\sum_h\widehat{t}_{Rh}}{\sum_h\widehat{A}_{Rh}} = \frac{\sum_h N_h\frac{\sum_{i=1}^{m}\widehat{t}_{rhi}}{\sum_{i=1}^{m}N_{hi}}}{\sum_h N_h\frac{\sum_{i=1}^{m}\widehat{a}_{rhi}}{\sum_{i=1}^{m}N_{hi}}}$$

Note that summation in (eq. (33)) is across all sample strips, even if a stratum is not present in that strip. In such cases, the contribution from the strip is zero. Alternatively, (see Ringvall et al. 2016) an estimator could be based on summation only across those strips that contain the stratum, in which case the basic estimator (eq. (33)) would remain the same but the variance and variance estimators would be slightly different. In the following, the variances are developed for the case of summation across all strips.

Details related to the derivation of the variance estimator of $\widehat{D}_{RoR,PS}$ are provided in Appendix B, which results in the following:

$$(34)\quad \widehat{V}\left(\widehat{D}_{RoR,PS}\right) \approx \frac{1}{\widehat{A}_{R,PS}^2}\widehat{V}\left(\widehat{t}_{R,PS} - \widehat{D}_{RoR,PS}\widehat{A}_{R,PS}\right) = \frac{1}{\widehat{A}_{R,PS}^2}\left[\widehat{V}\left(\widehat{t}_{R,PS}\right) + \widehat{D}_{RoR,PS}^2\widehat{V}\left(\widehat{A}_{R,PS}\right) - 2\widehat{D}_{RoR,PS}\,\widehat{Cov}\left(\widehat{t}_{R,PS}, \widehat{A}_{R,PS}\right)\right]$$

with

$$(35)\quad \widehat{V}\left(\widehat{t}_{R,PS}\right) = \sum_h\frac{N_h^2}{\widehat{N}_h^2}\left[M^2\left(\frac{1}{m} - \frac{1}{M}\right)s_{rh}^2 + \frac{M}{m}\sum_{S1}N_{hi}^2\left(\frac{1}{n_{hi}} - \frac{1}{N_{hi}}\right)s_{e,h}^2\right] + \sum_h\sum_{g\neq h}\frac{N_hN_g}{\widehat{N}_h\widehat{N}_g}M^2\left(\frac{1}{m} - \frac{1}{M}\right)\widehat{Cov}\left(r_h, r_g\right)$$

and

$$(36)\quad \widehat{V}\left(\widehat{A}_{R,PS}\right) = \sum_h\frac{N_h^2}{\widehat{N}_h^2}\left[M^2\left(\frac{1}{m} - \frac{1}{M}\right)s_{ra,h}^2 + \frac{M}{m}\sum_{S1}N_{hi}^2\left(\frac{1}{n_{hi}} - \frac{1}{N_{hi}}\right)s_{ea,h}^2\right] + \sum_h\sum_{g\neq h}\frac{N_hN_g}{\widehat{N}_h\widehat{N}_g}M^2\left(\frac{1}{m} - \frac{1}{M}\right)\widehat{Cov}\left(r_{a,h}, r_{a,g}\right)$$

and

$$(37)\quad \widehat{Cov}\left(\widehat{t}_{R,PS}, \widehat{A}_{R,PS}\right) = \sum_h\sum_g\frac{N_hN_g}{\widehat{N}_h\widehat{N}_g}M^2\left(\frac{1}{m} - \frac{1}{M}\right)\widehat{Cov}\left(r_h, r_{a,g}\right)$$

where $r_h = \widehat{t}_{rhi} - \widehat{R}_hN_{hi}$ and $r_{a,h} = \widehat{a}_{rhi} - \widehat{R}_{a,h}N_{hi}$, and the covariance terms estimated in the usual way across the m survey strips.

3. Evaluation

3.1. Empirical material and artificial population

3.1.1. FIA field plot measurements

The standard FIA plot is made up of four 1/60th ha (7.3 m radius) fixed-area, circular subplots spaced 36.6 m apart (for

Can. J. For. Res. 54: 1425–1442 (2024) | dx.doi.org/10.1139/cjfr-2024-0039

details on the FIA plot design and measurements, see Cahoon and Baer 2022). There were 880 FIA field plots within the Tanana G-LiHT coverage (Fig. 1). The coordinates of each subplot were obtained (with less than 2 m error) with GLONASS-enabled Trimble GeoXH mapping-grade Global Navigation Satellite System (GNSS) receivers (Andersen et al. 2009*b*, 2022; McGaughey et al. 2017). Aboveground dry biomass (Mg) for each measured tree was calculated using standard FIA procedures, and aggregated at the plot level (Mg/ha). In addition, the proportion of the FIA plot classified as accessible forestland was recorded as a fractional value between 0 and 1 (inclusive).

Table 1. Goddard's Lidar, Hyperspectral, and Thermal lidar specifications.

Instrument	Riegl VQ-480
Laser wavelength	1550 nm
Flying height	335 m AGL
Beam divergence	0.3 mrad
Footprint size	10 cm
Half-scanning angle	15 degrees
Average pulse density	3 pulses/m^2
Swath width	400 m

Note: AGL, Above Ground Level.

3.1.2. G-LiHT airborne laser scanner sampling

To augment the relatively sparse sample of field plots, single, linear swaths of high-resolution airborne remote sensing measurements, placed approximately 9.2 km apart, oriented in a northwest-southwest direction and covering every FIA plot, were acquired using G-LiHT mounted on a fixed-wing (Piper Cherokee) platform (Cook et al. 2013) (see Fig. 1). In this study, only the airborne lidar data from the G-LiHT instrument was utilized. The specifications for the lidar data collection are provided in Table 1.

Lidar point cloud data were processed into 1 m resolution Digital Terrain and Canopy Height Models (DTMs and CHMs, respectively). In order to allow for a direct comparison between high-resolution lidar-based canopy heights and the field-measured FIA inventory parameter (e.g., tree biomass), and to avoid any bias or inefficiency due to a difference in spatial support regions between the FIA plot footprint and the airborne lidar measurements, in this study the lidar metrics were extracted within a discrete number of "remote sensing plots" corresponding exactly to the FIA plot footprint (i.e., four 1/60th ha subplots, etc.), and spaced at 200 m intervals along the center of the G-LiHT swaths. The approach resulted in 59 090 total remote sensing plots within the Tanana G-LIHT coverage (Fig. 1), the average of the lidar CHM gridcells within a FIA plot footprint (1/15th ha) was calculated for (1) each actual field-measured FIA plot and (2) remote sensing plots distributed over the entire lidar coverage area.

3.1.3. Stratification layer

A stratification layer was generated from the National Land Cover Database (Homer et al. 2015) by reclassifying "Evergreen Forest" (Class 42) as Class 2, "Mixed Forest" (Class 43) as Class 3, "Deciduous Forest" (Class 41) as Class 4, and all other classes as Class 1. These strata were chosen so that stratum was an ordinal variable and biomass has a positive correlation with stratum, a desirable feature in developing the simulated population (described in next section).

3.1.4. Artificial population

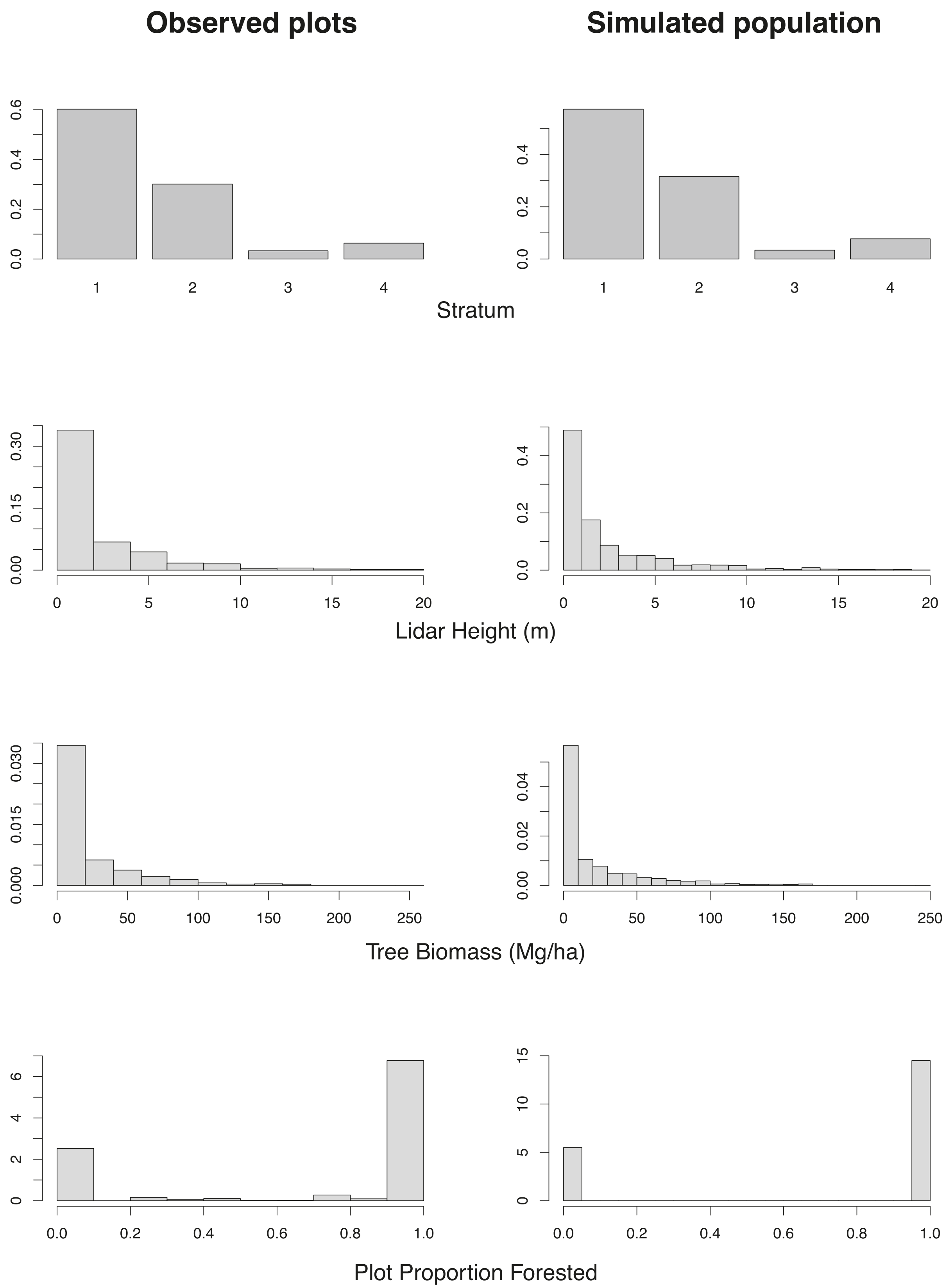
Simulation can be a useful approach to gain insight into the statistical properties of various survey estimators, especially in the case of somewhat complex, multi-level sampling designs (Ene et al. 2012; Saarela et al. 2017). When generating a simulated environmental population, it is desirable to include realistic correlations between the response variable (e.g., biomass, forest type, etc.) and the predictor variables used in the estimator (e.g., lidar height, strata, etc.), realistic marginal distributions for the attributes, and realistic spatial variation across multiple scales.

In this study, we use a combination of modeling approaches to develop a simulated population over the Tanana study area, with the objectives of generating values for independent and dependent variables that have realistic *marginal distributions*, *correlation structure*, and *spatial distributions*. Using the methods described in (Ene et al. 2012), a Gaussian copula model was used to generate a large (10 000 000) sample of simulated plots with marginal distributions and correlation structure similar to the observed plot data in the Tanana study area. To simulate the value for forestland proportion, we developed a logistic regression model using the observed data with forestland class (forested vs. nonforested) as the dependent variable and lidar height as the independent variable. For each of the 10 000 000 simulated values generated from the above copula model, we draw a random value from a binomial distribution where the parameter is the predicted probability from logistic regression. To introduce realistic spatial variability in the simulated population, we generate a dense grid of points (grid spacing of 200 m) across the entire Tanana study area (2703 291 total points), and at each of these points attach the stratum from the FIA stratification layer (based on NLCD (National Land Cover Database) land cover class; see Fig. 1). Finally, simulated plot values (stratum, lidar height, tree biomass, forestland proportion) at each point in the grid are imputed as random draws from the stratum-level simulated population.

In this way, the marginal distributions (Fig. 2) and correlation structure (Table 2) in the original plot data are generally maintained in the simulated population, while association with the mapped land cover (stratification) layer ensures that the simulated population is realistically distributed in the spatial domain (Fig. 3). In addition, each point in the dense grid is assigned a strip number to enable selection of random strip samples. This simulated population can then be used to select samples and assess the statistical properties (bias, variability) of alternative estimators.

At each iteration, an *SRS* of *m PSUs* (lidar strips) is drawn from the population, and from this a subsample of *n SSUs*

Fig. 2. Marginal distributions from observed plots (left column) and simulated population (right column).

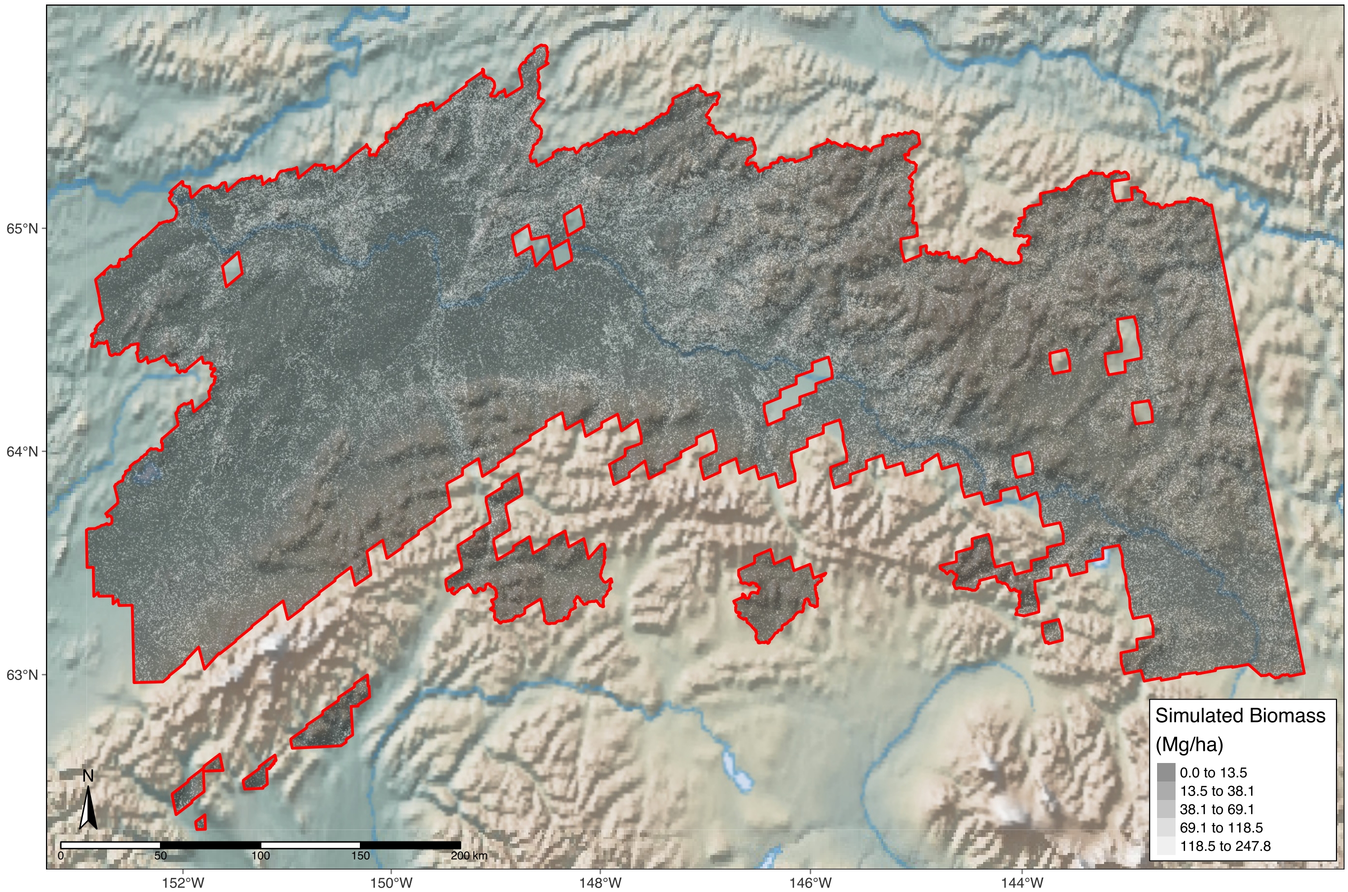


(plots) is drawn as an *SRS*, at a specified plot sampling intensity (defined as the ratio $\frac{\text{field plots}}{\text{remote sensing plots}}$), similar to the approach used in Ringvall et al. (2016). This approach allows for assessment of statistical properties for different sampling intensities for field plots (in this study we assume that the sampling intensity of the lidar is fixed) (Fig. 4). In order to evaluate the estimators under systematic sampling, simulated samples were drawn where the airborne remote sensing strips

Table 2. Correlation matrices for (a) observed field data and (b) simulated population.

(a)					(b)				
	Stratum	Lidar Ht.	*y*	*a*		Stratum	Lidar Ht.	*y*	*a*
Stratum	1.00	0.61	0.56	0.35	Stratum	1.00	0.52	0.58	0.24
Lidar Ht.	0.61	1.00	0.89	0.37	Lidar Ht.	0.52	1.00	0.85	0.36
y	0.56	0.89	1.00	0.36	*y*	0.58	0.85	1.00	0.31
a	0.35	0.37	0.36	1.00	*a*	0.24	0.36	0.31	1.00

Fig. 3. Spatial distribution of simulated population within Tanana study area. Base map from Natural Earth (https://www.naturalearthdata.com).



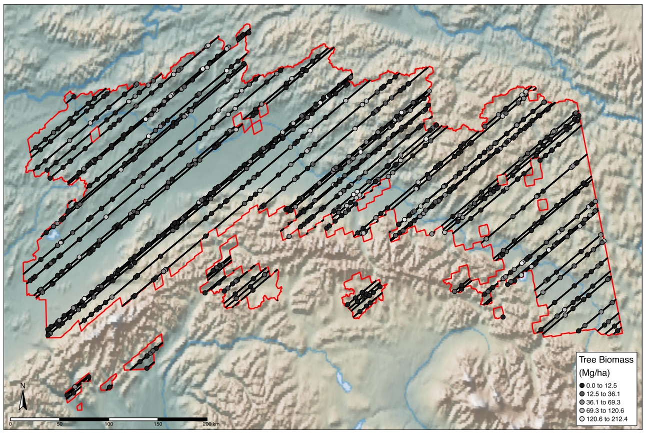
were collected at regular intervals and the field plots were distributed systematically within the airborne strips (with the same number of strips and field sampling intensities as the *SRS* sampling) (Fig. 5).

3.2. Simulation-based metrics

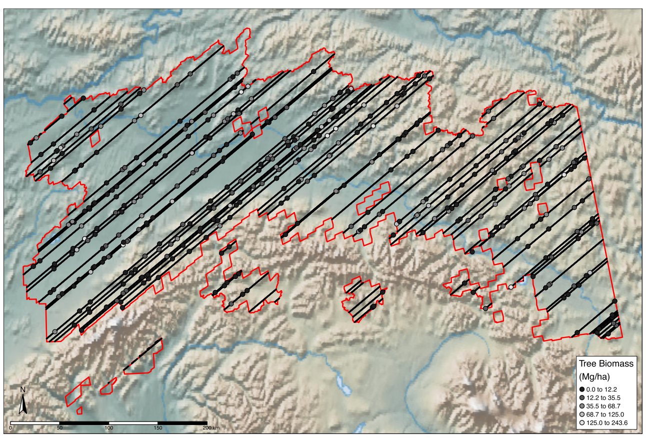
The properties of each estimator are assessed by drawing a large number (K) of simulated samples from the population and using these draws to develop a sampling distribution for each point estimator and corresponding variance estimator, which can be used to evaluate the bias, variance, as well as other measures such as the empirical coverage probability. In this study, we used $K = 10\,000$, which was considered to be an adequately large series in Särndal et al. (2003). For example, for an estimator of an arbitrary population parameter Y (i.e., a total, area, or density), at each iteration, the point estimator $\widehat{Y}_k$ and the standard deviation estimator $\sqrt{\widehat{V}\left(\widehat{Y}_k\right)}$ are calculated based on the equations presented in the previous sections, and the mean, standard deviation, and observed bias (in percent) are calculated for both the point estimator and standard deviation estimator based on the K simulated samples (see Ringvall et al. (2016) for details on these calculations). In addition, the *empirical coverage probability of the* 95% *confidence interval* for the point estimator was calculated as the proportion of simulations where the calculated 95% confidence interval:

$\left(\widehat{Y} - z_{0.025}\sqrt{\widehat{V}\left(\widehat{Y}\right)}, \widehat{Y} + z_{0.975}\sqrt{\widehat{V}\left(\widehat{Y}\right)}\right)$ contains the true value Y. The empirical coverage probability provides an indication of how reliable the CI (confidence interval) is as measure of uncertainty. An empirical coverage probability near 0.95 is an indicator that the 95% confidence intervals

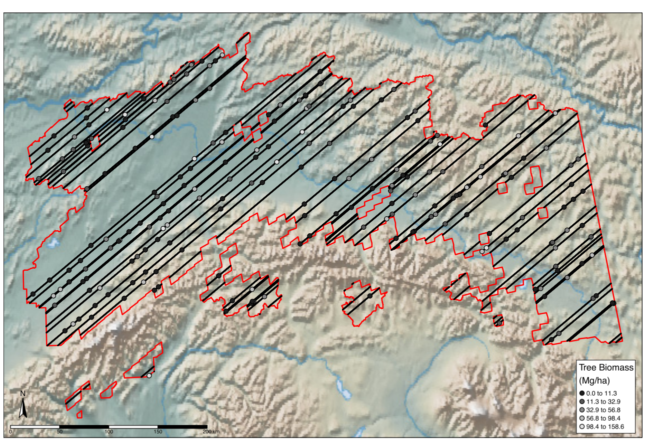
Fig. 4. Simulated samples at different sampling intensities with simple random sampling. Black lines indicate random sampled of airborne lidar flight lines. Dots indicate simulated field plots, color-coded by biomass. Base map from Natural Earth (https://www.naturalearthdata.com).



(a) Field plot sampling intensity: 0.015

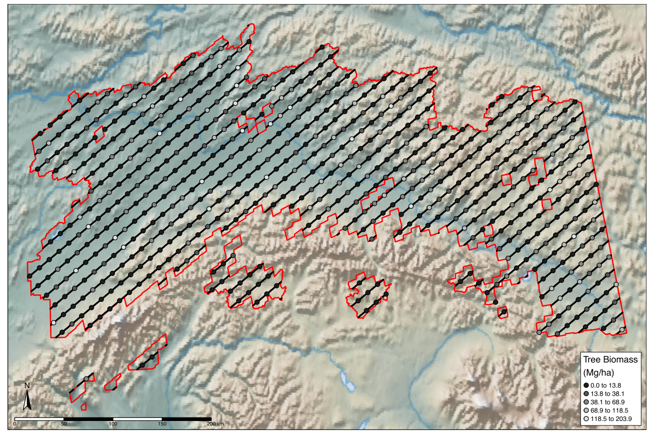


(b) Field plot sampling intensity: 0.0075

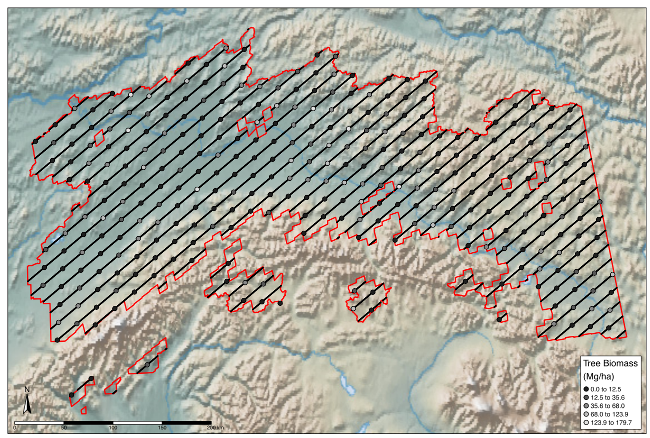


(c) Field plot sampling intensity: 0.00375

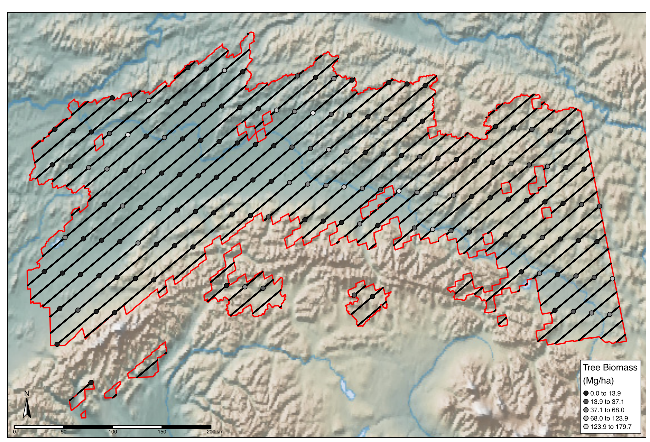
Fig. 5. Simulated samples at different sampling intensities with systematic sampling. Black lines indicate random sampled of airborne lidar flight lines. Dots indicate simulated field plots, color-coded by biomass. Base map from Natural Earth (https://www.naturalearthdata.com).



(a) Field plot sampling intensity: 0.015



(b) Field plot sampling intensity: 0.0075



(c) Field plot sampling intensity: 0.00375

 Can. J. For. Res. 54: 1425–1442 (2024) | dx.doi.org/10.1139/cjfr-2024-0039

Table 3. Simulation results, simple random sampling, and biomass (*kt*).

t	Estimator	$\hat{t}$ Mean	$\hat{t}$ St.Dev.	$\hat{t}$ Bias (%)	$\sqrt{\hat{V}(\hat{t})}$ Mean	$\sqrt{\hat{V}(\hat{t})}$ St.Dev.	$\sqrt{\hat{V}(\hat{t})}$ Bias (%)	95% CI Cov. Prob.
235 976				**Plot sampling intensity: 0.015**				
	SRS, PS	235 956	10 322	−0.0087	10 381	605	0.57	0.948
	R	235 979	8895	0.0012	8813	1035	−0.92	0.942
	R, PS	236 058	6990	0.035	6356	794	−9.1	0.928
				Plot sampling intensity: 0.0075				
	SRS, PS	235 961	14 746	−0.0065	14 760	1108	0.098	0.948
	R	235 918	11 198	−0.025	11 028	1359	−1.5	0.941
	R, PS	236 025	9902	0.02	8248	1180	−17	0.906
				Plot sampling intensity: 0.00375				
	SRS, PS	235 814	21 186	−0.069	20 821	2123	−1.7	0.941
	R	236 036	14 931	0.025	14 328	1942	−4	0.931
	R, PS	236 229	13 836	0.11	9711	1632	−30	0.820

Note: *SRS*, simple random sample; *PS*, Poststratified.

Table 4. Simulation results, simple random sampling, area of forestland (km^2).

A	Estimator	$\hat{A}$ Mean	$\hat{A}$ St.Dev.	$\hat{A}$ Bias (%)	$\sqrt{\hat{V}(\hat{A})}$ Mean	$\sqrt{\hat{V}(\hat{A})}$ St.Dev.	$\sqrt{\hat{V}(\hat{A})}$ Bias (%)	95% CI Cov. Prob.
78 404				**Plot sampling intensity: 0.015**				
	SRS, PS	78 387	1586	−0.022	1580	61	−0.41	0.947
	R	78 406	1467	0.0028	1449	155	−1.2	0.943
	R, PS	78 403	1494	−0.00096	1457	158	−2.5	0.928
				Plot sampling intensity: 0.0075				
	SRS, PS	78 400	2264	−0.0049	2240	97	−1.1	0.947
	R	78 395	2054	−0.012	2018	216	−1.8	0.942
	R, PS	78 389	2171	−0.018	2044	226	−5.8	0.906
				Plot sampling intensity: 0.00375				
	SRS, PS	78 384	3199	−0.025	3151	155	−1.5	0.944
	R	78 391	2853	−0.016	2812	294	−1.4	0.941
	R, PS	78 385	3033	−0.024	2708	304	−11	0.915

Note: *SRS*, simple random sample.

calculated using this estimator are reliable. Empirical coverage probabilities less than 0.95 indicate that the calculated 95% CIs are giving a falsely precise estimate of uncertainty, while coverage probabilities greater than 0.95 indicate that the 95% CIs obtained from this estimator are overly conservative.

3.3. Results and discussion

The ratio estimators for total biomass ($\hat{t}_R$ and $\hat{t}_{R,PS}$), forestland area ($\hat{A}_R$ and $\hat{A}_{R,PS}$), and biomass density in forestland ($\hat{D}_{RoR}$ and $\hat{D}_{RoR,PS}$) were generally unbiased at all sampling intensities and sampling configurations (simple random sampling vs. systematic), as was the standard poststratified estimator ($\hat{t}_{SRS,PS}$, $\hat{A}_{SRS,PS}$, and $\hat{D}_{SRS,PS}$) (Tables 3–8). While the observed bias was low (< 0.38%) for all estimators, it was higher for ratio estimators than the standard poststratified estimator. It should be noted that model-assisted estimators, such as the regression estimator that is a component of the ratio estimators used here, are *approximately* design-unbiased, which means that they may be biased for small sample sizes (Särndal et al. 2003). Furthermore, in the case of poststratified, two-stage sampling designs, one must consider the sample sizes by *stratum* and *PSU*, which can be quite small, potentially introducing bias in the estimators.

The ratio estimators of total biomass with simple random sampling (Table 3) were significantly more precise than the standard poststratified estimator $\hat{t}_{SRS,PS}$, with a reduction of approx. 32% in standard error (relative to standard poststratified estimator) for $\hat{t}_{R,PS}$ and 14% for $\hat{t}_R$ at the 0.015 sampling intensity, which also indicates the benefit of using poststratification. Ratio estimators of forestland area (Table 4) were more precise than the standard poststratified estimator, but the gains were modest and, interestingly, poststratification did not improve the precision of the ratio estimator (8% reduction in standard error for $\hat{A}_R$ in relation to $\hat{A}_{SRS,PS}$, but only 6% reduction for $\hat{A}_{R,PS}$ at 0.015 sampling intensity). There

Table 5. Simulation results, simple random sampling, average biomass in forestland (Mg/ha).

D	Estimator	$\widehat{D}$ Mean	St.Dev.	Bias (%)	$\sqrt{\widehat{V}(\widehat{D})}$ Mean	St.Dev.	Bias (%)	95% CI Cov. Prob.
				Plot sampling intensity: 0.015				
30.10	*SRS*, *PS*	30.11	1.33	0.035	1.33	0.08	0.1	0.950
	RoR	30.10	1.19	0.023	1.18	0.14	−0.36	0.946
	RoR, *PS*	30.12	1.06	0.072	0.98	0.12	−7	0.928
				Plot sampling intensity: 0.0075				
	SRS, *PS*	30.11	1.90	0.044	1.89	0.15	−0.6	0.946
	RoR	30.11	1.58	0.046	1.56	0.19	−1.6	0.942
	RoR, *PS*	30.13	1.52	0.12	1.32	0.18	−13	0.906
				Plot sampling intensity: 0.00375				
	SRS, *PS*	30.11	2.71	0.039	2.67	0.28	−1.5	0.940
	RoR	30.15	2.18	0.17	2.10	0.28	−3.4	0.935
	RoR, *PS*	30.18	2.13	0.28	1.63	0.25	−24	0.863

Note: *SRS*, simple random sample; *RoR*, *ratio-of-ratios*.

Table 6. Simulation results, systematic sampling, biomass (*kt*).

t	Estimator	$\widehat{t}$ Mean	St.Dev.	Bias (%)	$\sqrt{\widehat{V}(\widehat{t})}$ Mean	St.Dev.	Bias (%)	95% CI Cov. Prob.
				Plot sampling intensity: 0.015				
235 976	*SRS*, *PS*	235 991	12 156	0.0063	12 184	658	0.24	0.950
	R	236 026	8039	0.021	10 369	1253	29	0.986
	R, *PS*	236 006	8025	0.012	7372	1025	−8.1	0.925
				Plot sampling intensity: 0.0077				
	SRS, *PS*	236 034	17 216	0.025	17 298	1344	0.48	0.947
	R	236 109	11 286	0.056	12 913	1718	14	0.967
	R, *PS*	236 143	11 417	0.07	9675	1533	−15	0.903
				Plot sampling intensity: 0.00385				
	SRS, *PS*	235 848	25 007	−0.055	24 607	2811	−1.6	0.939
	R	236 201	16 277	0.095	16 909	2602	3.9	0.948
	R, *PS*	236 306	16 268	0.14	11 540	2191	−29	0.822

Note: *SRS*, simple random sample.

Table 7. Simulation results, systematic sampling, area of forestland (km^2).

A	Estimator	$\widehat{A}$ Mean	St.Dev.	Bias (%)	$\sqrt{\widehat{V}(\widehat{A})}$ Mean	St.Dev.	Bias (%)	95% CI Cov. Prob.
				Plot sampling intensity: 0.015				
78 404	*SRS*, *PS*	78 405	1869	0.002	1851	37	−0.97	0.945
	R	78 407	1688	0.0043	1695	203	0.41	0.945
	R, *PS*	78 403	1735	−0.00068	1685	205	−2.9	0.925
				Plot sampling intensity: 0.0077				
	SRS, *PS*	78 407	2632	0.0037	2621	76	−0.43	0.952
	R	78 407	2384	0.0045	2363	282	−0.89	0.939
	R, *PS*	78 398	2497	−0.0072	2379	291	−4.7	0.903
				Plot sampling intensity: 0.00385				
	SRS, *PS*	78 379	3733	−0.032	3716	160	−0.47	0.945
	R	78 406	3400	0.0026	3331	399	−2	0.936
	R, *PS*	78 382	3581	−0.028	3195	410	−11	0.911

Note: *SRS*, simple random sample.

Table 8. Simulation results, systematic sampling, average biomass in forestland (Mg/ha).

		$\widehat{D}$			$\sqrt{\widehat{V}\left(\widehat{D}\right)}$			95% CI Cov. Prob.
D	Estimator	Mean	St.Dev.	Bias (%)	Mean	St.Dev.	Bias (%)	
				Plot sampling intensity: 0.015				
	SRS, *PS*	30.11	1.55	0.032	1.56	0.088	0.82	0.949
	RoR	30.12	1.21	0.063	1.38	0.18	14	0.970
	RoR, *PS*	30.12	1.23	0.062	1.14	0.15	−7.3	0.925
				Plot sampling intensity: 0.0077				
30.10	*SRS*, *PS*	30.12	2.19	0.076	2.21	0.18	0.84	0.949
	RoR	30.14	1.70	0.14	1.82	0.25	7	0.953
	RoR, *PS*	30.15	1.74	0.18	1.54	0.23	−12	0.903
				Plot sampling intensity: 0.00385				
	SRS, *PS*	30.12	3.18	0.084	3.15	0.37	−0.96	0.941
	RoR	30.18	2.46	0.28	2.48	0.38	0.55	0.943
	RoR, *PS*	30.21	2.50	0.38	1.93	0.34	−23	0.862

Note: *SRS*, simple random sample; *RoR*, *ratio-of-ratios*.

were moderate gains in precision from using the ratio estimators in estimation of average tree biomass in forestland (Table 5) (20% reduction in standard error for $\widehat{D}_{RoR,PS}$ in relation to $\widehat{D}_{SRS,PS}$, 11% reduction for $\widehat{D}_{RoR}$ at the 0.015 sampling intensity).

The relative precision of the ratio estimators (vs. standard poststratified estimators) is more pronounced as the field sampling intensity decreased. For example, at the lowest sampling intensity of 0.0375, the reduction in standard error (relative to standard poststratified estimators) for ratio estimators was 30% for $\widehat{t}_R$ and 35% for $\widehat{t}_{R,PS}$, 11% for $\widehat{A}_R$ and 5% for $\widehat{A}_{R,PS}$, 20% for $\widehat{D}_{RoR}$ and 21% for $\widehat{D}_{RoR,PS}$.

The SE estimators in the standard FIA poststratification estimation framework are generally unbiased (i.e., bias < 1%) for all attributes and at all sampling intensities. The observed bias of SE estimators for the ratio estimators without poststratification was low, while *PS* ratio estimators are increasingly biased at low sampling intensities. The problem of negative bias in the variance of ratio estimators with smaller sample sizes has been reported in previous studies (Knottnerus and Scholtus 2019), and the further development of these estimators to reduce this bias should be a focus of future research.

The use of variance estimators that assume simple random sampling, when the sampling is actually systematic results in higher estimates of standard error (Tables 6–8). This result is often observed when sampling natural populations (such as forests) where attributes are spatially correlated, because systematic samples force a minimum distance between samples, whereas simple random sampling does not (Cochran 1977; Gregoire and Valentine 2007). Although the observed effect of this overestimation of SE in this study is generally small, it does result in several cases (notably at higher field sampling intensities) where 95% CI coverage probabilities for *SRS*, *PS*, *R*, and *RoR* estimators exceed 0.95. The use of *SRS* variance estimators with systematic samples is usually considered to be acceptable in NFI programs because it typically results in a conservative estimate for precision (i.e., overestimate of SE).

In the context of forest inventory in remote regions, it can be useful to express gains in precision by using an alternative design in terms of the increase in plot sample size needed to achieve this precision under the standard design. For example, to achieve the 32% reduction in standard error from using the $\widehat{t}_{R,PS}$ estimator would require 2.2 times more field plots in the standard FIA sampling design and estimation framework. Even modest reductions in standard error can lead to significant cost savings; for example, a reduction in standard error of 10% for $\widehat{A}_{R,PS}$ equates to using 1.23 times more plots in the standard design, a significant cost savings when cost per plot is high (e.g., ∼$10 000 USD/plot in interior Alaska) and annual number of plots installed exceeds 100 (Cahoon and Baer 2022). Although the G-LiHT airborne data were collected as part of a research study and it is therefore difficult to estimate the total cost of data collection, if the total cost of the G-LiHT acquisition (including transit, salaries, etc.) is conservatively estimated at $50/km, this represents a fraction of the cost of the increased field plot sample required to match the statistical precision of the two-stage design.

It should be noted that although the form of the variance estimators can appear complex, the fundamental principles behind both two-stage sampling and regression estimation still apply. For example, the precision of the model-assisted ratio estimators in a two-stage design will depend on how evenly the domain of interest occurs across *PSU*s. With a large variation between *PSU*s, the s_r^2 (or s_{rh}^2) term would also most likely be large. This, however, does not necessarily mean that the *RoR* estimator of the domain density has a large variance. Also, it is apparent that improvement in the models to predict y_i and a_i, will reduce the error terms s_e^2, leading to a smaller variance for the estimators.

In conclusion, the methods and results presented here indicate that airborne remote sensing data—collected in a sampling mode—can increase the efficiency of national forest inventories, especially in remote regions. The *RoR* estimator developed in this study enables estimation of population densities at the domain level within the context of the two-stage sampling design, and together with the ratio estimators previously presented in Ringvall et al. (2016), provide a full suite of model-assisted estimators that can serve as alternatives to the standard poststratified estimators in Bechtold and Patterson (2005) when an intermediate level of sampled airborne data is available. Model-assisted estimation, which includes both standard poststratification and regression-based approaches, supports design-unbiased estimation of inventory attributes, and associated confidence intervals, and therefore is well-suited to the NFI context, where a nationally consistent estimation framework is required.

It should be noted that the focus of this study was on the estimation framework, and not on the challenges of classifying domains or prediction of inventory attributes via airborne remote sensing. It is expected that future efforts will be devoted to the development of more sophisticated models for domain (e.g., forest type) and attribute characterization, which could significantly increase the precision of estimators presented here. For example, the use of machine learning algorithms (presumably trained on a large, external (i.e., independent) datasets to avoid overfitting), and applied to high-resolution, hyperspectral/hyperspatial imagery has the potential to markedly improve domain classification and forest attribute prediction (Dalponte et al. 2014; Breidt and Opsomer 2017). In addition, the estimators presented here focus on inventory attributes associated with forest *status*, and not *trends*. While several recent studies have reported that a similar two-stage sampling design and model-assisted estimation framework can be extended to estimate of change (Strîmbu et al. 2017), further research is required to develop and evaluate the full set of change estimators required in an NFI context.

Acknowledgements

We would like to thank the interior Alaska FIA field crews of the Alaska Department of Natural Resources, Division of Forestry for high-quality data collection under extremely challenging conditions, and GIS support from Thomas Thompson of the USDA Forest Service, Anchorage Forestry Sciences Laboratory. The authors would also like to thank Larry Corp and Ross Nelson for their significant contributions in the design, planning, and implementation of this study, and Tad Fickel for his participation in the 2014 G-LiHT campaign.

Article information

History dates

Received: 15 February 2024
Accepted: 25 July 2024
Accepted manuscript online: 7 August 2024
Version of record online: 31 October 2024

Copyright

Data availability

G-LiHT airborne remote sensing data used in this study are available through the G-LiHT webpage, http://gliht.gsfc.nasa.gov or by directly connecting to the anonymous ftp site, fusionftp.gsfc.nasa.gov. FIA field plot data used in this study are available at https://apps.fs.usda.gov/fia/datamart/datamart.html. However, the exact coordinates of plot locations are kept confidential as required by Federal law.

Author information

Author ORCIDs

Hans-Erik Andersen https://orcid.org/0000-0002-5546-9522
Göran Ståhl https://orcid.org/0000-0001-9030-8057
Bruce D. Cook https://orcid.org/0000-0002-8528-000X
Douglas C. Morton https://orcid.org/0000-0003-2226-1124
Andrew O. Finley https://orcid.org/0000-0002-2277-2912

Author contributions

Conceptualization: HA, GS, BC, DM, AF
Formal analysis: HA, GS
Funding acquisition: HA, BC, DM, AF
Investigation: HA, GS
Methodology: HA, GS, BC, DM, AF
Resources: BC, DM
Writing – original draft: HA, GS

Competing interests

The authors declare there are no competing interests.

Funding information

Funding for this project was provided by the USDA Forest Service Pacific Northwest Research Station and the National Aeronautics and Space Administration (NASA). This research was part of the project “A Joint USFS-NASA Pilot Project to Estimate Forest Carbon Stocks in Interior Alaska by Integrating Field, Airborne and Satellite Data” funded by the NASA Carbon Monitoring System (CMS) (NNX13AQ51G).

References

Andersen, H.-E. 2009. Using airborne light detection and ranging (LIDAR) to characterize forest stand condition on the Kenai Peninsula of Alaska. West. J. Appl. For. **24**(2): 95–102.

Andersen, H.-E., Barrett, T., Winterberger, K., Strunk, J., and Temesgen, H. 2009a. Estimating forest biomass on the western lowlands of the Kenai Peninsula of Alaska using airborne lidar and field plot data in a model-assisted sampling design. In Proceedings of the IUFRO Division 4 Conference: Extending Forest Inventory and Monitoring Over Space and Time. pp. 19–22.

Andersen, H.-E., Clarkin, T., Winterberger, K., and Strunk, J. 2009b. An accuracy assessment of positions obtained using survey-and recreational-grade global positioning system receivers across a range of forest conditions within the Tanana Valley of interior Alaska. West. J. Appl. For. **24**(3): 128–136.

Andersen, H.-E., Strunk, J., and McGaughey, R.J. 2022. Using high-performance global navigation satellite system technology to improve forest inventory and analysis plot coordinates in the pacific region. PNW-GTR-1000. USDA Forest Service, Pacific Northwest Research Station.

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.

Bechtold, W.A., and Patterson, P.L. 2005. The enhanced forest inventory and analysis program–national sampling design and estimation procedures. SRS-GTR-80. USDA Forest Service, Southern Research Station.

Breidt, F.J., and Opsomer, J.D. 2017. Model-assisted survey estimation with modern prediction techniques. Stat. Sci. **32**(2): 190–205.

Cahoon, S., and Baer, K. 2022. Forest resources of the Tanana Unit, Alaska: 2018. PNW-GTR-1005. USDA Forest Service, Pacific Northwest Research Station.

Cochran, W. 1977. Sampling techniques. Wiley, New York.

Cook, B.D., Corp, L.A., Nelson, R.F., Middleton, E.M., Morton, D.C., McCorkel, J.T., et al. 2013. NASA Goddard's LiDAR, Hyperspectral and Thermal (G-LiHT) airborne imager. Remote Sens. **5**(8): 4045–4066.

Dalponte, M., Ørka, H.O., Ene, L.T., Gobakken, T., and Næsset, E. 2014. Tree crown delineation and tree species classification in boreal forests using hyperspectral and als data. Remote Sens. Environ. **140**: 306–317. doi:10.1016/j.rse.2013.09.006.

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. Global Change, **5**. doi:10.3389/ffgc.2022.813569.

Ene, L., Gobakken, T., Andersen, H.-E., Næsset, E., Cook, B.D., Morton, D.C., et al. 2018. Large-area hybrid estimation of aboveground biomass in interior Alaska using airborne laser scanning data. Remote Sens. Environ. **204**: 741–755.

Ene, L., Næsset, E., Gobakken, T., Gregoire, T., Ståhl, G., and Nelson, R. 2012. Assessing the accuracy of regional LiDAR-based biomass estimation using a simulation approach. Remote Sens. Environ. **123**: 579–592. doi:10.1016/j.rse.2012.04.017.

EPA. 2023. Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990-2021. U.S. Environmental Protection Agency, 2023. Available from https://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-and-sinks-1990-2021 [accessed 24 September 2024].

FAO. 2016. Global forest resources assessment 2015: how are the world's forests changing? United Nations Food; Agricultural Organization. Available from https://www.fao.org/3/i4793e/i4793e.pdf [accessed 23 September 2024].

Gobakken, T., Næsset, E., Nelson, R., Bollandsås, O.M., Gregoire, T.G., Ståhl, G., et al. 2012. Estimating biomass in Hedmark County, Norway using national forest inventory field plots and airborne laser scanning. Remote Sens. Environ. **123**: 443–456. doi:10.1016/j.rse.2012.01.025.

Gregoire, T., and Valentine, H. 2007. Sampling strategies for natural resources and the environment. Chapman; Hall/CRS, New York.

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**(1): 83–95.

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. Remote Sens. **81**(5): 345–354.

Kangas, A., Astrup, R., Breidenbach, J., Fridman, J., Gobakken, T., Korhonen, K., et al. 2018. Remote sensing and forest inventories in Nordic countries–roadmap for the future. Scand. J. For. Res. **33**. doi:10.1080/02827581.2017.14166662018.

Knottnerus, P., and Scholtus, S. 2019. On a new estimator for the variance of the ratio estimator with small sample corrections. Surv. Methodol. **45**(3): 567–576. Available from https://www150.statcan.gc.ca/n1/pub/12-001-x/2019003/article/00003-eng.htm [accessed 23 September 2024].

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**(12): 1364.

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.

Marland, E., Domke, G., Hoyle, J., Marland, G., Bates, L., Helms, A., et al. 2017. Understanding and analysis: the California Air Resources Board Forest Offset Protocol. Springer, Cham. doi:10.1007/978-3-319-52434-4.

McConville, K.S., Breidt, F.J., Lee, T., and Moisen, G.G. 2017. Model-assisted survey regression estimation with the lasso. J. Surv. Stat. Methodol. **5**(2): 131–158.

McGaughey, R.J., Ahmed, K., Andersen, H.-E., and Reutebuch, S.E. 2017. Effect of occupation time on the horizontal accuracy of a mapping-grade GNSS receiver under Dense Forest Canopy. Photogramm. Eng. Remote Sens. **83**(12): 861–868.

McRoberts, R.E. 2010. Probability-and model-based approaches to inference for proportion forest using satellite imagery as ancillary data. Remote Sens. Environ. **114**(5): 1017–1025.

McRoberts, R.E., Andersen, H.-E., and Næsset, E. 2014. Using airborne laser scanning data to support forest sample surveys. *In* Forestry applications of airborne laser scanning. Springer, Dordrecht. pp. 269–292.

Ringvall, A., Stahl, G., Ene, L., Næsset, E., Gobakken, T., and Gregoire, T. 2016. A poststratified ratio estimator for model-assisted biomass estimation in sample-based airborne laser scanning surveys. Can. J. For. Res. **46**: 1386–1395. doi:10.1139/cjfr-2016-0158.

Saarela, S., Andersen, H.-E., Grafström, A., Schnell, S., Gobakken, T., Næsset, E., et al. 2017. A new prediction-based variance estimator for two-stage model-assisted surveys of forest resources. Remote Sens. Environ. **192**: 1–11.

Saarela, S., Holm, S., Grafström, A., Schnell, S., Næsset, E., Gregoire, T.G., et al. 2016. Hierarchical model-based inference for forest inventory utilizing three sources of information. Ann. For. Sci. **73**(4): 895–910.

Särndal, C.-E., Swensson, B., and Wretman, J. 2003. Model assisted survey sampling. Springer Science & Business Media.

Shoot, C., Andersen, H.-E., Moskal, L.M., Babcock, C., Cook, B.D., and Morton, D.C. 2021. Classifying forest type in the national forest inventory context with airborne hyperspectral and lidar data. Remote Sens. **13**(10). doi:10.3390/rs13101863.

Ståhl, G., Holm, S., Gregoire, T.G., Gobakken, T., Næsset, E., and Nelson, R. 2011. Model-based inference for biomass estimation in a LiDAR sample survey in Hedmark County, Norway. Can. J. For. Res. **41**(1): 96–107.

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): 1–11.

Strîmbu, V.F., Ene, L.T., Gobakken, T., Gregoire, T.G., Astrup, R., and Næsset, E. 2017. Post-stratified change estimation for large-area forest biomass using repeated ALS strip sampling. Can. J. For. Res. **47**(6): 839–847.

Strunk, J.L., Temesgen, H., Andersen, H.-E., and Packalen, P. 2014. Prediction of forest attributes with field plots, landsat, and a sample of lidar strips. Photogramm. Eng. Remote Sens. **80**(2): 143–150.

USDA Forest Service. 2016. Forest inventory and analysis strategic plan. *In* FS-1079. USDA Forest Service Washington, DC, USA.

Van den Boogaart, K.G., and Tolosana-Delgado, R. 2013. Analyzing compositional data with R. Springer, Berlin. doi:10.1007/978-3-642-36809-7.

Westfall, J.A., Coulston, J.W., Moisen, G.G., and Andersen, H.-E. 2022. Sampling and estimation documentation for the enhanced forest inventory and analysis program. NRS-GTR-207. USDA Forest Service, Northern Research Station. doi:10.2737/nrs-gtr-2072022.

Appendix A: Variance of *ratio-of-ratios* estimator of the domain density

In deriving the variance estimation for the *ratio-of-ratios* (*RoR*) estimator of domain density without poststratification, we note that a first-order Taylor approximation leads to

$$\text{(A1)}\quad V\left(\widehat{D}_{RoR}\right) = V\left[\frac{\sum_{i=1}^{m}\widehat{t}_{ri}}{\sum_{i=1}^{m}\widehat{a}_{ri}}\right] \approx \frac{1}{E\left(\left(\sum_{i=1}^{m}\widehat{a}_i\right)^2\right)} V\left[\sum_{i=1}^{m}\left(\widehat{t}_{ri} - D\widehat{a}_{ri}\right)\right] = \frac{1}{E\left(\left(\sum_{i=1}^{m}\widehat{a}_i\right)^2\right)} V(\cdot)$$

Conditioning of the first stage sample leads to: $V(\cdot) = V_1(E_2(\cdot)) + E_1(V_2(\cdot))$. This means that we separate the variance into (1) a first term which is the variance, among first stage sampling units, of their expected values from second stage sampling and (2) a second term which is the variance from second stage sampling. This is a common technique applied for deriving variances following multistage or multiphase sampling (e.g., Cochran 1977). We start by looking at $V(\cdot) = V_1(E_2(\cdot))$, that is $V_1\left(E_2\left(\sum_{i=1}^{m}\widehat{t}_{ri} - D\widehat{a}_{ri}\right)\right)$. Because the second stage expectations are (at least approximately) the true values of t and a for all first stage units, we obtain $V_1\left(\sum_{i=1}^{m}\widehat{t}_{ri} - D\widehat{a}_{ri}\right) = m\left(1 - \frac{m}{M}\right)S_u^2 = m^2\left(\frac{1}{m} - \frac{1}{M}\right)S_u^2$ where $D = D_{RoR}$ and $S_u^2 = \frac{1}{M-1}\sum_{i=1}^{M}(u_i - \bar{u})^2$.

Thus

$$\begin{aligned}\text{(A2)}\quad V_1\left(E_2\left(\widehat{D}_{RoR}\right)\right) &= \frac{1}{E\left(\left(\sum_{i=1}^{m}\widehat{a}_i\right)^2\right)} m^2\left(\frac{1}{m} - \frac{1}{M}\right)S_u^2 \\ &= \frac{1}{\frac{M^2}{m^2}E\left(\left(\sum_{i=1}^{m}\widehat{a}_i\right)^2\right)} M^2\left(\frac{1}{m} - \frac{1}{M}\right)S_u^2 \\ &= \frac{1}{E\left(\widehat{A^*}^2\right)} M^2\left(\frac{1}{m} - \frac{1}{M}\right)S_u^2\end{aligned}$$

This is the first term of the variance. An estimator is obtained by substituting $E\left(\widehat{A^*}^2\right)$ by $\widehat{A^*}^2$ and S_u^2 by s_u^2, i.e., the estimator based on the sample.

We then turn to the second term of the Taylor approximation, i.e., $E_1(V_2(\cdot))$, which is the expectation across first stage units of the second stage variances. Thus, $E_1(V_2(\cdot)) = E_1\left(V_2\left(\sum_{i=1}^{m}\widehat{t}_{ri} - D\widehat{a}_{ri}\right)\right)$. Variability in the second stage is due to subsampling field plots, where

$$\begin{aligned}\text{(A3)}\quad & V_2\left(\sum\widehat{t}_{ri} - D\widehat{a}_{ri}\right) \\ &= V_2\left(\sum\left(\sum C_1 + \frac{N_i}{n_i}\sum e_{ik}^{y}\right) - D\left(\sum C_2 + \frac{N_i}{n_i}\sum e_{ik}^{a}\right)\right) \\ &= V_2\left(\sum^{m}\left(\frac{N_i}{n_i}\left(\sum\left(e_{ik}^{y} - De_{ik}^{a}\right)\right)\right)\right) \\ &= V_2\left(\sum^{m}\left(\frac{N_i}{n_i}\left(\sum v_{ik}\right)\right)\right) \\ &= \sum^{m}\frac{N_i^2}{n_i^2}V_2\left(\sum_{k} v_{ik}\right) \\ &= \sum^{m}\frac{N_i^2}{n_i^2}n_i\left(1 - \frac{n_i}{N_i}\right)S_{vi}^2 \\ &= \sum^{m}N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_{vi}^2\end{aligned}$$

where $v_{ik} = e_{ik}^{y} - De_{ik}^{a}$, C_1 are predictions of y_i (with error terms e_{ik}^{y}), C_2 are predictions of a_i (with error terms e_{ik}^{a}), and C_1 and C_2 are constants that do not affect the variance.

Next, we should take the expectation over first stage samples:

$$\begin{aligned}\text{(A4)}\quad & E_1\left(\sum^{m}N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_{vi}^2\right) \\ &= \sum^{M}\pi_i N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_{vi}^2 \\ &= \sum^{M}\frac{m}{M}N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_{vi}^2 \\ &= \frac{m}{M}\sum^{M}N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_{vi}^2\end{aligned}$$

where π_i is the inclusion probability of the ith *PSU*.

Can. J. For. Res. 54: 1425–1442 (2024) | dx.doi.org/10.1139/cjfr-2024-0039

Lastly, we should multiply this term by $\frac{1}{E\left(\sum^m \widehat{a}_{ri}\right)^2}$ leading to:

$$\text{(A5)} \quad E_1V_2\left(\widehat{D}_{RoR}\right) = \frac{1}{E\left(\sum^m \widehat{a}_{ri}\right)^2}\frac{m}{M}\sum^M N_i^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_{vi}^2$$
$$= \frac{1}{\frac{M^2}{m^2}E\left(\sum^m \widehat{a}_{ri}\right)^2}\frac{M}{m}\sum^M N_i^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_{vi}^2$$
$$= \frac{1}{E\left(\widehat{A^*}\right)^2}\frac{M}{m}\sum^M N_i^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_{vi}^2$$

This is the second term in the variance. To obtain an estimator, we substitute $E\left(\widehat{A^*}\right)^2$ by $\widehat{A^*}^2$ and S_{vi}^2 by s_{vi}^2 for the sampled *PSU*s. Also, we change the summation from M to m. This might appear to lead to an underestimation of the variance, but this underestimation if compensated by using estimated rather than the true values when S_u^2 is estimated (e.g., see Cochran 1977 or Särndal et al. 2003). The resulting variance estimator is therefore:

$$\text{(A6)} \quad \widehat{V}\left(\widehat{D}_{RoR}\right) = \frac{1}{\widehat{A^*}^2}\left[M^2\left(\frac{1}{m}-\frac{1}{M}\right)s_u^2 + \frac{M}{m}\sum_{S_1} N_i^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right)s_{vi}^2\right]$$

Appendix B: Variance of *ratio-of-ratios* estimator of the domain density with poststratification

$$\text{(B1)} \quad V\left(\widehat{D}_{RoR,PS}\right) \approx \frac{1}{E\left(\widehat{A}_{R,PS}\right)^2}V\left(\widehat{t}_{R,PS} - D_{RoR,PS}\widehat{A}_{R,PS}\right) = \frac{1}{E\left(\widehat{A}_{R,PS}\right)^2}\left[V\left(\widehat{t}_{R,PS}\right) + D^2{}_{RoR,PS}V\left(\widehat{A}_{R,PS}\right) - 2D_{RoR,PS}Cov\left(\widehat{t}_{R,PS},\ \widehat{A}_{R,PS}\right)\right]$$

We expand each of the terms inside the square brackets, reflecting the two-stage sampling procedure, i.e.,$V\left(\widehat{t}_{R,PS}\right) = V_1\left(E_2\left(\widehat{t}_{R,PS}\right)\right) + E_1\left(V_2\left(\widehat{t}_{R,PS}\right)\right)$, and similarly for $V\left(\widehat{A}_{R,PS}\right)$ and $Cov\left(\widehat{t}_{R,PS}, \widehat{A}_{R,PS}\right)$. In the following this will be done for $V\left(\widehat{t}_{R,PS}\right)$ and $Cov\left(\widehat{t}_{R,PS}, \widehat{A}_{R,PS}\right)$, noting that the formula for $V\left(\widehat{A}_{R,PS}\right)$ will be similar to the formula for $V\left(\widehat{t}_{R,PS}\right)$. Thus,

$$\text{(B2)} \quad V\left(\widehat{t}_{R,PS}\right) = V\left(\sum_h \widehat{t}_{Rh}\right) = \sum_h V\left(N_h\frac{\sum_{i=1}^m \widehat{t}_{rhi}}{\sum_{i=1}^m N_{hi}}\right) + \sum_h\sum_{g\neq h} Cov\left(N_h\frac{\sum_{i=1}^m \widehat{t}_{rhi}}{\sum_{i=1}^m N_{hi}}, N_g\frac{\sum_{i=1}^m \widehat{t}_{rgi}}{\sum_{i=1}^m N_{gi}}\right)$$

In eq. (B2), each of the terms $V\left(N_h\frac{\sum_{i=1}^m \widehat{t}_{rhi}}{\sum_{i=1}^m N_{hi}}\right)$ can be further developed as

$$\text{(B3)} \quad V\left(N_h\frac{\sum_{i=1}^m \widehat{t}_{rhi}}{\sum_{i=1}^m N_{hi}}\right) \approx \frac{N_h^2}{\left(E\sum_{i=1}^m N_{hi}\right)^2}V\left(\sum_{i=1}^m \left(\widehat{t}_{rhi} - R_hN_{hi}\right)\right)$$

In eq. (B3), the R_h-terms are the stratum-specific ratios $\frac{\sum_{i=1}^m t_{rhi}}{\sum_{i=1}^m N_{hi}}$. Addressing $V\left(\sum_{i=1}^m \left(\widehat{t}_{rhi} - R_hN_{hi}\right)\right)$ through $V_1(E_2(.)) + E_1(V_2(.))$, the first term will be $V\left(\sum_{i=1}^m \left(t_{rhi} - R_hN_{hi}\right)\right)$ and the second term $\frac{m}{M}V\left(\sum_{i=1}^M \left(\widehat{t}_{rhi}\right)\right)$. Again, we use $r_h = t_{rhi} - R_hN_{hi}$. Using simple random sampling without replacement the variance of the first term is $m^2\left(\frac{1}{m}-\frac{1}{M}\right)S_{rh}^2$ and the variance of the second term $\frac{1}{M^2}\sum_M N_{hi}^2\left(\frac{1}{n_{hi}}-\frac{1}{N_{hi}}\right)\ S_{e,h}^2$. Here, $S_{e,h}^2 = \frac{1}{N_{hi}-1}\sum_{k=1}^{N_{hi}}\left(e_{k,h}-\bar{e}_{hi}\right)^2$ with $\bar{e}_{hi} = \frac{\sum e_{k,h}}{N_{hi}}$. Thus, each of the variance terms in eq. (B3) can be expressed as

$$\text{(B4)} \quad V\left(N_h\frac{\sum_{i=1}^m \widehat{t}_{rhi}}{\sum_{i=1}^m N_{hi}}\right) = \frac{N_h^2}{\left(\frac{m}{M}\sum_{i=1}^M N_{hi}\right)^2}\left[m^2\left(\frac{1}{m}-\frac{1}{M}\right)S_{rh}^2 + \frac{m^2}{M^2}\sum_M N_{hi}^2\left(\frac{1}{n_{hi}}-\frac{1}{N_{hi}}\right)S_{e,h}^2\right]$$
$$= \frac{N_h^2}{\left(\frac{1}{M}\sum_{i=1}^M N_{hi}\right)^2}\left[\left(\frac{1}{m}-\frac{1}{M}\right)S_{rh}^2 + \frac{1}{M^2}\sum_M N_{hi}^2\left(\frac{1}{n_{hi}}-\frac{1}{N_{hi}}\right)S_{e,h}^2\right]$$
$$= \frac{N_h^2}{\bar{N}_h^2}\left[M^2\left(\frac{1}{m}-\frac{1}{M}\right)S_{rh}^2 + \sum_M N_{hi}^2\left(\frac{1}{n_{hi}}-\frac{1}{N_{hi}}\right)S_{e,h}^2\right]$$

where $\bar{N}_h$ is the average across all strips.

A variance estimator corresponding to the variance in eq. (B4) would be

$$\text{(B5)}\quad \hat{V}\left(N_h \frac{\sum_{i=1}^{m} \hat{t}_{rhi}}{\sum_{i=1}^{m} N_{hi}}\right) = \frac{N_h^2}{\hat{N}_h^2}\left[M^2\left(\frac{1}{m} - \frac{1}{M}\right)s_{rh}^2 + \frac{M}{m}\sum_{S1} N_{hi}^2\left(\frac{1}{n_{hi}} - \frac{1}{N_{hi}}\right)s_{e,h}^2\right]$$

The variance $V\left(\hat{A}_{R,PS}\right)$ in eq. (B1) and its estimator would be similar, using "area" instead of biomass in s_{rh}^2 and $s_{e,h}^2$.

Next, we address the covariance terms in eq. (B2), i.e., $Cov\left(N_h \frac{\sum_{i=1}^{m} \hat{t}_{rhi}}{\sum_{i=1}^{m} N_{hi}}, N_g \frac{\sum_{i=1}^{m} \hat{t}_{rgi}}{\sum_{i=1}^{m} N_{gi}}\right)$. The second term of $Cov(.,.) = Cov_1(E_2(.,.)) + E_1(Cov_2(.,.)$ should be zero, since no covariance should arise due to the second stage subsampling. Thus, the covariance would be limited to the first term, i.e., to $Cov\left(N_h \frac{\sum_{i=1}^{m} t_{rhi}}{\sum_{i=1}^{m} N_{hi}}, N_g \frac{\sum_{i=1}^{m} t_{rgi}}{\sum_{i=1}^{m} N_{gi}}\right)$. Using the Taylor approximation this covariance can be re-written as

$$\text{(B6)}\quad Cov\left(N_h \frac{\sum_{i=1}^{m} t_{rhi}}{\sum_{i=1}^{m} N_{hi}}, N_g \frac{\sum_{i=1}^{m} t_{rgi}}{\sum_{i=1}^{m} N_{gi}}\right) \approx \frac{N_h N_g}{E\left(\sum_i^m N_{hi}\right) E\left(\sum_i^m N_{gi}\right)}$$

$$Cov\left(\sum_{i=1}^{m} t_{rhi} - R_h N_{hi}, \sum_{i=1}^{m} t_{irg} - R_g N_{ig}\right) = \frac{N_h N_g}{E\left(\sum_i^m N_{hi}\right) E\left(\sum_i^m N_{gi}\right)} Cov\left(\sum_{i=1}^{m} r_{hi}, \sum_{i=1}^{m} r_{gi}\right)$$

Thus, under simple random sampling without replacement the covariance would be

$$\text{(B7)}\quad Cov\left(N_h \frac{\sum_{i=1}^{m} t_{rhi}}{\sum_{i=1}^{m} N_{hi}}, N_g \frac{\sum_{i=1}^{m} t_{rgi}}{\sum_{i=1}^{m} N_{gi}}\right) = \frac{N_h N_g}{E\left(\sum_i^m N_{hi}\right) E\left(\sum_i^m N_{gi}\right)} m \left(1 - \frac{m}{M}\right) Cov\left(r_h, r_g\right)$$
$$= \frac{N_h N_g}{E\left(\hat{N}_h\right) E\left(\hat{N}_g\right)} M^2 \left(1 - \frac{m}{M}\right) \frac{Cov\left(r_h, r_g\right)}{m}$$

Again, a covariance estimator would use the estimates of $E\left(\hat{N}_h\right)$ and $E\left(\hat{N}_g\right)$ and $Cov(r_h, r_g)$ estimated in the usual way across the m survey strips. The covariance for the terms related to $\hat{A}_{R,PS}$ should be computed in a similar way, with "area" inputs rather than biomass inputs.

What remains from eq. (B1) is the term $Cov\left(\hat{t}_{R,PS}, \hat{A}_{R,PS}\right) = Cov\left(\sum_h \hat{t}_{Rh}, \sum_h \hat{A}_{Rh}\right) = \sum_h \sum_g Cov\left(\hat{t}_{Rh}, \hat{A}_{Rg}\right)$. Like before, it is not likely that the second stage subsampling leads to covariance and thus only the covariance conditional on the stage 2 expectations should be addressed. This leads to

$$\text{(B8)}\quad Cov\left(N_h \frac{\sum_{i=1}^{m} \hat{t}_{rhi}}{\sum_{i=1}^{m} N_{hi}}, N_g \frac{\sum_{i=1}^{m} \hat{a}_{rgi}}{\sum_{i=1}^{m} N_{gi}}\right) \approx \frac{N_h N_g}{E\left(\sum_i^m N_{hi}\right) E\left(\sum_i^m N_{gi}\right)} Cov\left(\sum_{i=1}^{m} t_{rhi} - R_h N_{hi}, \sum_{i=1}^{m} a_{rgi} - R_{a,g} N_{gi}\right)$$
$$= \frac{N_h N_g}{E\left(\sum_i^m N_{hi}\right) E\left(\sum_i^m N_{gi}\right)} Cov\left(\sum_{i=1}^{m} r_{hi}, \sum_{i=1}^{m} r_{a,gi}\right)$$
$$= \frac{N_h N_g}{E\left(\hat{N}_h\right) E\left(\hat{N}_g\right)} M^2 \left(1 - \frac{m}{M}\right) \frac{Cov\left(r_h, r_{a,g}\right)}{m}$$

An estimator for $V\left(\hat{A}_{R,PS}\right)$ would be derived similarly.