Supplementary materials for this article are available at https://doi.org/10.1007/s13253-020-00425-z.

Combining Environmental Area Frame Surveys of a Finite Population

Wilmer PRENTIUS, Xin ZHAO, and Anton GRAFSTRÖM

New ways to combine data from multiple environmental area frame surveys of a finite population are being introduced. Environmental surveys often sample finite populations through area frames. However, to combine multiple surveys without risking bias, design components (inclusion probabilities, etc.) are needed at unit level of the finite population. We show how to derive the design components and exemplify this for three commonly used area frame sampling designs. We show how to produce an unbiased estimator using data from multiple surveys, and how to reduce the risk of introducing significant bias in linear combinations of estimators from multiple surveys. If separate estimators and variance estimators are used in linear combinations, there's a risk of introducing negative bias. By using pooled variance estimators, the bias of a linear combination estimator can be reduced. National environmental surveys often provide good estimators at national level, while being too sparse to provide sufficiently good estimators for some domains. With the proposed methods, one can plan extra sampling efforts for such domains, without discarding readily available information from the aggregate/national survey. Through simulation, we show that the proposed methods are either unbiased, or yield low variance with small bias, compared to traditionally used methods.

Key Words: Combining data sources; Combining estimators; Environmental monitoring; Linear combination estimator; Sample design properties.

1. INTRODUCTION

For a traditional finite population survey, one often think of some well-structured list frame covering the population of interest, from which a statistician can draw a sample according to some procedure, in order to produce an efficient and unbiased estimator of some population parameter. When conducting environmental surveys, however, this is often not the case.

Environmental surveys often lack well-structured, comprehensive list frames to sample from. In such settings, it is common to use area frames covering the assumed spread of

W. Prentius (✉)· X. Zhao · A. Grafström, Department of Forest Resource Management, Swedish University of Agricultural Sciences, 90183 Umeå, Sweden
(E-mail: *wilmer.prentius@slu.se*).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 26, Number 2, Pages 250–266
https://doi.org/10.1007/s13253-020-00425-z

the population of interest. Examples of environmental surveys using such area frames are national forest inventories (Axelsson et al. 2010), agricultural inventories (Fecso et al. 1986), landscape inventories (Allard 2017), among others. By using area frames, a sample unit becomes a point from a continuous population—the area surface—why there is a need to map the sample properties for the sampled points to the indirectly sampled units in the population of interest.

Other desirable outcomes in environmental surveys are domain estimates, or their counterparts, estimates created by aggregating domain estimates. In the first case, primary surveys are seldom planned with domain estimates in mind, why complementary surveys are often considered. The latter case may especially be considered when dealing with rare populations, or wanting to incorporate a previously conducted domain survey into an aggregate survey (Benedetti et al. 2015).

Scenarios like these, or when dealing with two samples with different designs, connect to the multiple-frame research area. When combining such samples, an optimal linearly combined estimator should be weighted by the variance (Lohr and Rao 2006). Since true variances are most likely not available, variance estimates are often used instead. However, environmental surveys conducted using area frames often have target variables with highly skewed distributions, since the units in the population of interest might be absent in large parts of the area frame. Under such circumstances, the estimators and the variance estimators are susceptible to correlation, which can introduce significant bias into linearly combined estimates using variance estimates as weights (Grafström et al. 2019).

In order to reduce the bias of a combined estimate, we propose two methods: The first approach is a generalization of the combining samples approach derived by Grafström et al. (2019), which combines unit sample properties from an arbitrary number of designs into design components for the combined design. The second approach uses a pooled variance estimator to estimate the variance of each survey's estimator by using all available information from the surveys.

The targeted applications are primarily environmental surveys and monitoring, where it is common to use area frames. Several countries have national landscape and forest monitoring programs that may not be enough to produce regional or domain level estimates, and thus need be complemented on some level to reach specific accuracy targets (Christensen and Ringvall 2013).

With the methodology presented in this paper, there might be a need to link surveys relating to different definitions of statistical units. Hence, this is something that should be planned for from start. We need be able to detect if the same population unit is included in more than one sample (or multiple times in the same sample). However, in most applications, the size of the area being sampled is likely to be very large compared to the area covered in the samples, which makes overlap not particularly common. In area-based surveys, we are likely to have geographical coordinates for at least the statistical unit. These coordinates can easily be used to detect possible overlap between different surveys. In the rare case of possible overlap, it may be difficult identify exactly which population unit that is included multiple times. If this is thought to be an issue, then it may be needed to use markings of coordinates and/or population units in the field to make such identification easier.

In some cases, e.g., for unbiased variance estimation using a combined sample, we need at least partial knowledge of the geographical coordinates of the sampled population units. Such knowledge can be included by the use of accurate satellite-based positioning systems, as is done, e.g., for permanent sample plots in the Swedish national forest inventory (Fridman et al. 2014).

In Sect. 2, we provide a general procedure to produce unit sample properties for a discrete population sampled using an area frame. Through Sect. 2.1, we show examples on unit sample properties for a discrete population sampled through three different, commonly used area frame designs. In Sect. 3, we recall the single and multiple count estimators that are used to estimate population totals. Then, in Sect. 4, we present the theory for combining samples, and for combining estimators using pooled variance estimators. In Sect. 5, we use a simulation to compare a naive linear combination with the combined sample and the linear combination using pooled variance estimates. Finally, we discuss the results in Sect. 6.

2. UNIT SAMPLE PROPERTIES FOR GENERAL DESIGNS

Assume that there is a finite, but unknown population U, represented by fixed points on an area of interest F_U, that has some measurable properties of interest. If a sample point $\mathbb{X}^{(k)}$, with probability density function (pdf) $f^{(k)}(\mathbf{x})$, falls within the inclusion zone $A_i^{(k)}$ of an unit $i \in U$, the unit is included in the sample.

Let P be the set of independent but not necessarily equally distributed sample points. For any sample point $\mathbb{X}^{(k)} \in P$, and units $\{i, j\} \in U$, we make the following definitions:

$$S_i^{(k)} := \mathrm{I}\left(\mathbb{X}^{(k)} \in A_i^{(k)}\right), \tag{1}$$

$$\pi_i^{(k)} := \Pr\left(S_i^{(k)} > 0\right) = \int_{A_i^{(k)}} f^{(k)}(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{2}$$

$$\pi_{ij}^{(k)} := \Pr\left(S_i^{(k)} > 0, S_j^{(k)} > 0\right) = \int_{A_i^{(k)} \cap A_j^{(k)}} f^{(k)}(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{3}$$

$$E_i^{(k)} := \mathrm{E}\left[S_i^{(k)}\right] = \pi_i^{(k)}, \tag{4}$$

$$E_{ij}^{(k)} := \mathrm{E}\left[S_i^{(k)} S_j^{(k)}\right] = \pi_{ij}^{(k)}, \tag{5}$$

where $\mathrm{I}(\cdot)$ denotes the indicator function, $S_i^{(k)}$ is the number of inclusions of unit i by sample point $\mathbb{X}^{(k)}$, $\pi_i^{(k)}$ is the first-order inclusion probability of unit i by sample point $\mathbb{X}^{(k)}$, i.e., the probability of unit i being included into the sample by a sample point $\mathbb{X}^{(k)}$, $\pi_{ij}^{(k)}$ is the second-order inclusion probability for units i, j to be included in the sample simultaneously by sample point $\mathbb{X}^{(k)}$, $E_i^{(k)}$ is the (first-order) expected number of inclusions of unit i by $\mathbb{X}^{(k)}$, and $E_{ij}^{(k)}$ is the second-order expected number of inclusions of units i, j by $\mathbb{X}^{(k)}$.

For the set of independent sample points P, we extend the definition in (1) to

$$S_i^{(P)} := \sum_{\mathbb{X}^{(k)} \in P} S_i^{(k)}. \tag{6}$$

Expanding the definition of (4) to the first-order expected number of inclusions for unit i by the set of sample points P, we have

$$E_i^{(P)} := \mathrm{E}\left[S_i^{(P)}\right] = \sum_{\mathbb{X}^{(k)} \in P} E_i^{(k)}, \tag{7}$$

while it can be shown (see "Appendix" for further details), that the expected number of inclusions of the second-order for units i, j by the set of sample points P can be extended from (5) to

$$E_{ij}^{(P)} := \mathrm{E}\left[S_i^{(P)} S_j^{(P)}\right] = E_i^{(P)} E_j^{(P)} + \sum_{\mathbb{X}^{(k)} \in P} \left(E_{ij}^{(k)} - E_i^{(k)} E_j^{(k)}\right). \tag{8}$$

Moreover, the inclusion probabilities of the first and second-order of units i, j by the set of sample points P can be expressed similarly to (2) and (3) as

$$\pi_i^{(P)} := \Pr\left(S_i^{(P)} > 0\right) = 1 - \prod_{\mathbb{X}^{(k)} \in P} \left(1 - \pi_i^{(k)}\right), \tag{9}$$

$$\pi_{ij}^{(P)} := \Pr\left(S_i^{(P)} > 0, S_j^{(P)} > 0\right) = \pi_i^{(P)} + \pi_j^{(P)} - \left(1 - \prod_{\mathbb{X}^{(k)} \in P} \left(1 - \pi_i^{(k)} - \pi_j^{(k)} + \pi_{ij}^{(k)}\right)\right). \tag{10}$$

For any set of sample points P to be used to make an unbiased estimator of a parameter of U, we require that all units in the population have positive inclusion probabilities, equivalent to ensuring that a sampling design satisfies

$$\forall i \in U \; \exists \mathbb{X}^{(k)} \in P : \pi_i^{(k)} > 0. \tag{11}$$

For an unbiased estimator of variance by any set of sample points P, we require that all pairs of units $\{i, j\} \in U$ have positive second-order inclusion probabilities, equivalent to ensuring that a sampling design satisfies

$$\forall \{i, j\} \in U \; \exists \{\mathbb{X}^{(k)}, \mathbb{X}^{(k')}\} \in P, k \neq k' : \pi_{ij}^{(k)} + \pi_i^{(k)} \pi_j^{(k')} > 0. \tag{12}$$

While the requirements in (11) and (12) are necessary and sufficient for positive inclusion probabilities of the first and second-order, they are in reality often not assessable if the units in U are unknown. Instead, sufficient counterparts with respect to F_U can be formulated as

$$\forall \mathbf{x} \in F \; \exists \mathbb{X}^{(k)} \in P : f^{(k)}(\mathbf{x}) > 0, \tag{13}$$

$$\forall \{\mathbf{x}, \mathbf{x}'\} \in F \; \exists \{\mathbb{X}^{(k)}, \mathbb{X}^{(k')}\} \in P, k \neq k' : f^{(k)}(\mathbf{x}) f^{(k')}(\mathbf{x}') > 0, \tag{14}$$

where F, the sample frame, is connected to F_U so that $\int_{F_U \setminus F} \mathrm{d}\mathbf{x} = 0$, assuming reasonably defined inclusion zones. It holds that (14) is sufficient for (13).

2.1. SAMPLE PROPERTIES FOR THREE COMMON DESIGNS

Provided the derived sample properties, it is easy to show the sample properties for three common designs—i.i.d., one point per stratum stratified, and systematic—given uniform sample point distributions. Assuming that unit i's inclusion zones are identical for all sample points within a specific design, i.e., $A_i^{(k)} = A_i$ for all $\mathbb{X}_d^{(k)}$, we define F as the area enclosing all possible inclusion zones, a_F as the area of F, a_i as the area of A_i, and a_{ij} as the area of $A_i \cap A_j$.

An i.i.d. design defined by P_1 implies that $f_1^{(k)}(\mathbf{x}) = f_1^{(k')}(\mathbf{x})$ for every pair of sample points $\mathbb{X}_1^{(k)}, \mathbb{X}_1^{(k')}$. The inclusion probabilities for units i, j by a single sample point $\mathbb{X}_1^{(k)}$ can thus be described as

$$\pi_i^{(k)} = \int_{A_i} f_1^{(k)}(\mathbf{x})\mathrm{d}\mathbf{x} = \frac{a_i}{a_F},$$
$$\pi_{ij}^{(k)} = \int_{A_i \cap A_j} f_1^{(k)}(\mathbf{x})\mathrm{d}\mathbf{x} = \frac{a_{ij}}{a_F}.$$

From this, it follows that the first-order sample properties for unit i are

$$\pi_i^{(P_1)} = 1 - \left(1 - \frac{a_i}{a_F}\right)^{n_1}, \qquad E_i^{(P_1)} = n_1 \frac{a_i}{a_F},$$

with the second-order sample properties for units i, j

$$\pi_{ij}^{(P_1)} = \pi_i^{(P_1)} + \pi_j^{(P_1)} - \left(1 - \left(1 - \frac{a_i + a_j - a_{ij}}{a_F}\right)^{n_1}\right),$$
$$E_{ij}^{(P_1)} = \frac{n_1(n_1 - 1)}{a_F a_F} a_i a_j + \frac{n_1 a_{ij}}{a_F},$$

where n_1 denotes the cardinality of P_1, i.e., the number of sample points in the design.

A systematic design with uniform pdf's, and a repeating pattern in the inclusion zones defined by the stratification (exemplified in Fig. 1), is a special case of the i.i.d. design where only one point is sampled. Thus, for the systematic design, the sample properties for units i, j are $\pi_i^{(P_2)} = E_i^{(P_2)} = a_i/a_F$ and $\pi_{ij}^{(P_2)} = E_{ij}^{(P_2)} = a_{ij}/a_F$.

The final example is the one point per stratum stratified design defined by P_3, where one point is sampled from each of a fixed number of disjoint strata. Let the stratum for sample point $\mathbb{X}_3^{(k)}$ be given as $F^{(k)} = \{\mathbf{x} : f_3^{(k)}(\mathbf{x}) > 0\}$, $a_F^{(k)}$ be the area of $F^{(k)}$, $a_i^{(k)}$ denote the area of $A_i \cap F^{(k)}$, and let $a_{ij}^{(k)}$ denote the area of $A_i \cap A_j \cap F^{(k)}$. The inclusion probabilities for units i, j by $\mathbb{X}_3^{(k)}$, given uniform pdf's, can then be described as

$$\pi_i^{(k)} = \int_{A_i} f_3^{(k)}(\mathbf{x})\mathrm{d}\mathbf{x} = \frac{a_i^{(k)}}{a_F^{(k)}},$$
$$\pi_{ij}^{(k)} = \int_{A_i \cap A_j} f_3^{(k)}(\mathbf{x})\mathrm{d}\mathbf{x} = \frac{a_{ij}^{(k)}}{a_F^{(k)}},$$

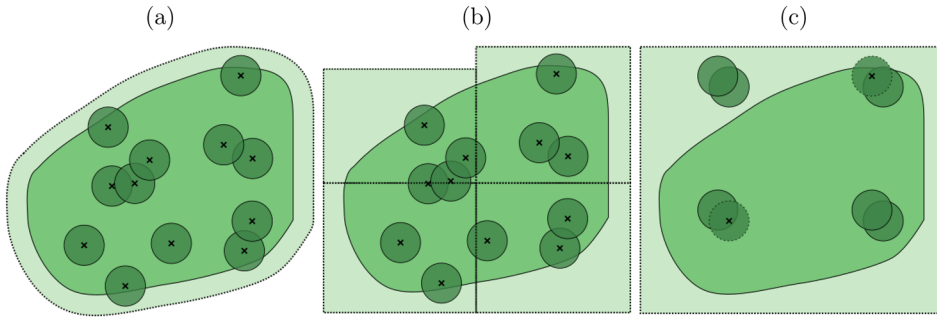

Figure 1. Examples of **a** i.i.d., **b** stratified, and **c** systematic frames and inclusion zones. The outer areas represent the sample frames (F), the inner areas represents the areas of interest (F_U), and the circles represents the inclusion zones (A) for units. In both **a** and **b**, the sample frame expands around the area of interest so that the largest of the inclusion zones will always be fully within the area frame. In **b** four disjoint strata of unequal sizes and shapes are exemplified through the dashed lines. **c** shows inclusion zones for two units, where dashed circles and x'es indicate the units' positions. These types of inclusion zones would exemplify systematic plot sampling.

from which the results in (7), (8), (9), and (10) follows. In the case of equally sized and disjoint strata, $a_F^{(k)} = a_F/n_3$, where n_3 represent the number of strata/sample points.

3. SINGLE AND MULTIPLE COUNT ESTIMATORS

The sample properties derived in Sect. 2 are needed for two common estimators used when estimating the population total $Y = \sum_{i\in U} y_i$ of a finite population U . The first of these two estimators is the single-count (SC) Horvitz–Thompson estimator (Horvitz and Thompson 1952), defined as

$$\hat{Y}_{SC} = \sum_{i\in U} \frac{y_i}{\pi_i} \mathrm{I}\left(S_i > 0\right),$$

where S_i denotes the number of inclusions of unit i, $\pi_i = \Pr\left(S_i > 0\right)$ denotes the inclusion probability for unit i, i.e., the probability for unit i to be included in the sample, and $\mathrm{I}\left(\cdot\right)$ denotes the indicator function. The variance of $\hat{Y}_{SC}$ can be shown to be

$$\mathrm{V}\left(\hat{Y}_{SC}\right) = \sum_{i\in U}\sum_{j\in U} \frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\left(\pi_{ij} - \pi_i\pi_j\right),$$

where $\pi_{ij} = \Pr\left(S_i > 0, S_j > 0\right)$ denotes the second-order inclusion probability, i.e., the probability for units i, j to be included in the sample simultaneously. Given that the second-order inclusion probabilities are strictly positive for all pairs $\{i, j\} \in U$, an unbiased variance estimator for $\hat{Y}_{SC}$ is

$$\hat{\mathrm{V}}\left(\hat{Y}_{SC}\right) = \sum_{i\in U}\sum_{j\in U} \frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\left(\pi_{ij} - \pi_i\pi_j\right) \times \frac{\mathrm{I}\left(S_i > 0\right)\mathrm{I}\left(S_j > 0\right)}{\pi_{ij}}.$$

The second estimator to be used in this paper is the multiple-count (MC), or Hansen–Hurwitz, estimator (Hansen and Hurwitz 1943), defined as

$$\hat{Y}_{MC} = \sum_{i \in U} \frac{y_i}{E_i} S_i,$$

where $E_i = \mathrm{E}[S_i]$ denotes the expected number of inclusions for an unit i. The variance of $\hat{Y}_{MC}$ is

$$\mathrm{V}\left(\hat{Y}_{MC}\right) = \sum_{i \in U} \sum_{j \in U} \frac{y_i}{E_i} \frac{y_j}{E_j} \left(E_{ij} - E_i E_j\right),$$

where $E_{ij} = \mathrm{E}\left[S_i S_j\right]$ denotes the second-order expected number of inclusions for two units i, j. Given that the second-order expected number of inclusions are strictly positive for all pairs $\{i, j\} \in U$, an unbiased variance estimator of $\hat{Y}_{MC}$ is

$$\hat{\mathrm{V}}\left(\hat{Y}_{MC}\right) = \sum_{i \in U} \sum_{j \in U} \frac{y_i}{E_i} \frac{y_j}{E_j} \left(E_{ij} - E_i E_j\right) \frac{S_i S_j}{E_{ij}}.$$

As by the requirements in (13) and (14), the variance estimators presented here are not applicable when using a one-per-stratum stratified or systematic sample design such as those presented in Sect. 2.1. However, when combining two or more independent samples, these criteria will be evaluated on the combined sample.

4. COMBINING SAMPLES

Let $\mathcal{D} = \{P_d\}_d$ denote a combined sample, i.e., a set of independent sets of sample points P_d. By extending the definition of (6) to the number of inclusions by the combined sample as

$$S_i^{(\mathcal{D})} := \sum_{P_d \in \mathcal{D}} S_i^{(P_d)}, \tag{15}$$

the inclusion probability of unit i by a combined sample $\mathcal{D}$ becomes

$$\pi_i^{(\mathcal{D})} = 1 - \prod_{P_d \in \mathcal{D}} \left(1 - \pi_i^{(P_d)}\right), \tag{16}$$

similar to (9). Comparable to (7), (8), and (10), the rest of the necessary sample properties for units i, j by a combined sample $\mathcal{D}$ follows as

$$\begin{aligned} \pi_{ij}^{(\mathcal{D})} =& \pi_i^{(\mathcal{D})} + \pi_j^{(\mathcal{D})} \\ & - \left(1 - \prod_{P_d \in \mathcal{D}} \left(1 - \pi_i^{(P_d)} - \pi_j^{(P_d)} + \pi_{ij}^{(P_d)}\right)\right), \end{aligned} \tag{17}$$

$$E_i^{(\mathcal{D})} = \sum_{P_d \in \mathcal{D}} E_i^{(P_d)}, \tag{18}$$

$$E_{ij}^{(\mathcal{D})} = E_i^{(\mathcal{D})} E_j^{(\mathcal{D})} + \sum_{P_d \in \mathcal{D}} \left(E_{ij}^{(P_d)} - E_i^{(P_d)} E_j^{(P_d)} \right). \tag{19}$$

By using these combined sample properties, the estimators in Sect. 3 can be applied directly.

When combining samples, for example in a multiple frame setting, the individual designs' sample frames do not need to be identical, nor do they need to individually cover the area of interest. The requirements in (11) and (12) needs to be fulfilled with respect to the sample points in $\cup_d P_d$, i.e., the necessary condition for positive second-order inclusion probabilities and positive expected number of inclusions for all pairs in the combined sample $\mathcal{D}$ is

$$\begin{aligned} &\forall \{i, j\} \in U \; \exists \{\mathbb{X}_d^{(k)}, \mathbb{X}_{d'}^{(k')}\} \in \cup_d P_d, \\ &(k, d) \neq (k', d') : \pi_{ij}^{(k)} + \pi_i^{(k)} \pi_j^{(k')} > 0, \end{aligned} \tag{20}$$

with sufficient counterpart

$$\begin{aligned} &\forall \{\mathbf{x}, \mathbf{x}'\} \in F \; \exists \{\mathbb{X}_d^{(k)}, \mathbb{X}_{d'}^{(k')}\} \in \cup_d P_d, \\ &(k, d) \neq (k', d') : f_d^{(k)}(\mathbf{x}) f_{d'}^{(k')}(\mathbf{x}') > 0, \end{aligned} \tag{21}$$

both of which imply positive first-order inclusion probabilities and positive expected number of inclusions for all units by the combined sample $\mathcal{D}$.

If sample frames are extended in ways similar to those in Fig. 1, or if combining multiple frames, there will be some oversampling. In such cases, it will be required to be able to identify objects not part of the population of interest.

These results are not limited to area frames. As per an example in Lohr and Rao (2006), it is possible to combine, for example, a sample taken from an area frame with full coverage of the population of interest, and a list frame with unknown coverage of the population of interest, as long as it is possible to identify units in the list frame that are not part of the population of interest, and units sampled from the area frame that are also present in the list frame.

4.1. COMBINING ESTIMATORS BY LINEAR COMBINATIONS

When combining a set of unbiased estimates formed of the samples in $\mathcal{D}$ by linear combinations, the form

$$\hat{Y}_L^{(\mathcal{D})} = \sum_{P_d \in \mathcal{D}} \alpha^{(P_d)} \hat{Y}^{(P_d)}$$

is often considered, since it will yield an unbiased result. Often the inverse variance proportion is used as the weight in order to increase accuracy. However, as described by Grafström et al. (2019), if true variances are not available, using variance estimates may in certain cases introduce bias to such a linear combination, especially when the variance estimator is correlated with the estimator of the population parameter. We denote a linear combination

using variance estimates as

$$\hat{Y}_{L*}^{(\mathcal{D})} = \sum_{P_d \in \mathcal{D}} \hat{\alpha}_*^{(P_d)} \hat{Y}_*^{(P_d)}, \qquad \hat{\alpha}_*^{(P_d)} = \frac{\hat{\mathrm{V}}\left(\hat{Y}_*^{(P_d)}\right)^{-1}}{\sum_{P_{d'} \in \mathcal{D}} \hat{\mathrm{V}}\left(\hat{Y}_*^{(P_{d'})}\right)^{-1}},$$

with $*$ for either SC (single-count) or MC (multiple-count).

To overcome the issue with biased variance estimators, we propose a pooled variance estimator, using all available information to estimate the separate variances. We denote the linear combination estimator using such pooled variance estimates as

$$\hat{Y}_{LP*}^{(\mathcal{D})} = \sum_{P_d \in \mathcal{D}} \hat{\alpha}_{P*}^{(P_d)} \hat{Y}_*^{(P_d)}, \qquad \hat{\alpha}_{P*}^{(P_d)} = \frac{\hat{\mathrm{V}}_P\left(\hat{Y}_*^{(P_d)}\right)^{-1}}{\sum_{P_{d'} \in \mathcal{D}} \hat{\mathrm{V}}_P\left(\hat{Y}_*^{(P_{d'})}\right)^{-1}}, \tag{22}$$

where

$$\begin{aligned}
\hat{\mathrm{V}}_P\left(\hat{Y}_{SC}^{(P_d)}\right) &= \sum_{i \in U} \sum_{j \in U} \frac{y_i}{\pi_i^{(P_d)}} \frac{y_j}{\pi_j^{(P_d)}} \left(\pi_{ij}^{(P_d)} - \pi_i^{(P_d)} \pi_j^{(P_d)}\right) \\
&\quad \times \frac{\mathrm{I}\left(S_i^{(\mathcal{D})} > 0\right) \mathrm{I}\left(S_j^{(\mathcal{D})} > 0\right)}{\pi_{ij}^{(\mathcal{D})}}, \\
\hat{\mathrm{V}}_P\left(\hat{Y}_{MC}^{(P_d)}\right) &= \sum_{i \in U} \sum_{j \in U} \frac{y_i}{E_i^{(P_d)}} \frac{y_j}{E_j^{(P_d)}} \left(E_{ij}^{(P_d)} - E_i^{(P_d)} E_j^{(P_d)}\right) \\
&\quad \times \frac{S_i^{(\mathcal{D})} S_j^{(\mathcal{D})}}{E_{ij}^{(\mathcal{D})}},
\end{aligned}$$

are both unbiased estimators of the variances of the single and multiple count estimators, given $\forall\{i, j\} \in U, \pi_{ij}^{(\mathcal{D})} > 0$ and $\forall\{i, j\} \in U, E_{ij}^{(\mathcal{D})} > 0$. Note that the final fractions for both variance estimators for a design P_d assures that all available information are used through $S_i^{(\mathcal{D})}$, $\pi_{ij}^{(\mathcal{D})}$ and $E_{ij}^{(\mathcal{D})}$, as defined in (15), (17) and (19). However, if many second-order design properties are positive, but small, the variance estimators might produce negative and unstable estimates, making them unsuitable for combinations.

5. SIMULATION

In order to evaluate the proposed combinations of samples and estimates, a simulation study was performed. The simulation sampled 10,000 times from a simulated population generated from the SLU (Swedish University of Agricultural Sciences) Forest Map (Reese et al. 2003). The SLU Forest Map, previously known as kNN-Sweden, has extensive information about Swedish forest land and is based on satellite and field data from the Swedish national forest inventory (NFI). The map contains information about age, height, species

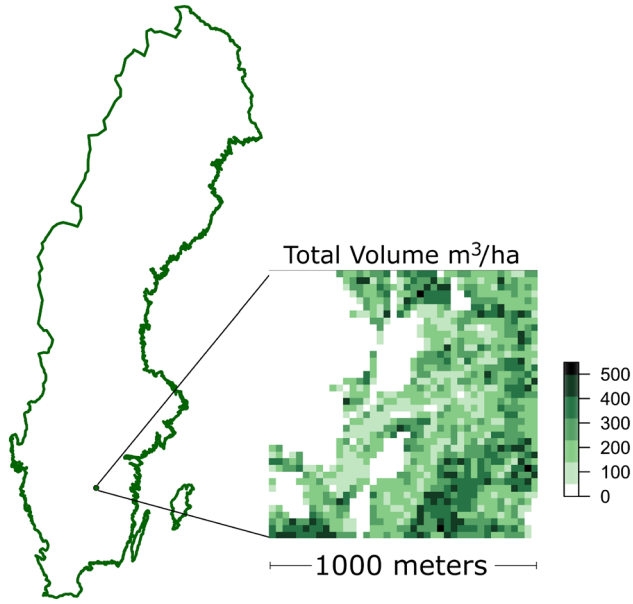

Figure 2. Location and the total biomass volume (m^3/ha) for the area used as a boilerplate for simulating the population. Darker colors indicate higher volumes (Color figure online).

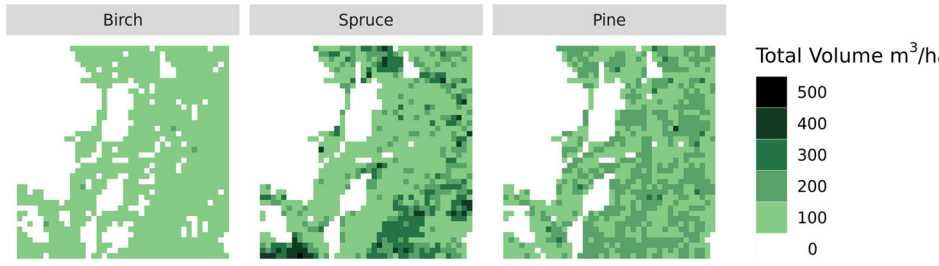

Figure 3. Total biomass volume (m^3/ha) per species for the simulated population. Darker colors indicate higher volumes (Color figure online).

of wood and woodland for the country's forest land. The basic format is raster data with a resolution of 25 × 25 square meters.

From the SLU Forest map, an area of 1000 × 1000 square meters of southern Sweden was cropped to represent the area of interest. Figure 2 illustrates the location as well as the total volume of the stand for the cropped area. Using individual tree data variables from the Swedish NFI, the three dominating tree species—birch, pine, and spruce—were randomly added to the population according to species-specific volume maps of the cropped area. In the resulting population, the number of trees for each species is 7411 (13%), 24,428 (41%) and 27,212 (46%), respectively. The resulting population is presented in Fig. 3, color-coded by volume intensity.

For each of the 10,000 simulation runs, four samples were generated from the sample frame using uniform densities—two i.i.d. samples, one systematic sample, and one stratified sample. Each design used circular inclusion zones of common sizes per design, correspond-

Table 1. Sample designs used in the simulation study

Design	n	Radius (m)	Sample frame (m^2)	Stratum size (m^2)	Sampled area (m^2)
i.i.d. 1	10	10	1020 × 1020		3142
i.i.d. 2	40	5	1010 × 1010		3142
Systematic	16	8	1016 × 1016	254 × 254	3217
Stratified	16	8	1016 × 1016	254 × 254	3217

n Sample size; *Radius* Radius of inclusion zones

ing to plot sampling. In order to have equal first-order expected number of inclusions for all units, the sample frames were expanded around the area of interest in each direction by the size of the inclusion zone radius, guaranteeing that all inclusion zones are fully within the sample frames. In Table 1, the designs are described in further detail.

For each sample and combination, single (SC) and multiple count (MC) estimates were calculated. To show the effect of different ways of combining data, we compared the estimators using combined samples, with sample properties derived through (16), (17), (18) and (19), with the estimators based on linear combinations of estimates using estimated variances and pooled variance estimates as in (22).

As mentioned in Sect. 3, for variance estimators to be unbiased, we require positive second-order sample properties for all pairs in the population. While the systematic and stratified designs fulfills the requirements in (20) and (21) in combination with each other or any of the i.i.d. designs, they do not fulfill (12) and (14) individually, while also being prone to negative and unstable pooled variance estimates due to small second-order design properties, making them unsuitable to use in a linear combination. In environmental surveys, one often deal with this by using a more conservative variance estimator, for example by using the i.i.d. variance estimator (Benedetti et al. 2015). However, using the i.i.d. variance estimator might be too conservative, i.e., reducing the assumed efficiency of the stratified and systematic designs.

For this simulation, second-order design properties were calculated as if they were sampled using a i.i.d. design, when calculating the linear combination of estimates using pooled variances. For the naive combination, plot variance estimates in the linear combination

$$\hat{\mathrm{V}}_{Plot}\left(\hat{Y}_{MC}^{(P_d)}\right) = \frac{1}{n_d(n_d-1)} \sum_{\mathbb{X}_d^{(k)} \in P_d} \left(y_d^{(k)} - \hat{y}_d\right)^2,$$

$$\hat{y}_d = \frac{1}{n_d} \sum_{\mathbb{X}_d^{(l)} \in P_d} y_d^{(l)},$$

were used, where $y_d^{(l)}$ is the plot l estimate of the total. In order to reduce the efficiency impact of the stratified and systematic designs, plot variances were calculated using a variant of the local mean variance estimator proposed by Grafström and Schelin (2013)

Table 2. Results from 10,000 simulations for the i.i.d. 1 (i), systematic (sy), and stratified (st) designs showing [empirical relative bias] and relative root-mean-squared error (RRMSE) for birches and all species in percent

	SC	MC	LPlot		LPSC		LPMC	
Birches								
i	50.22	50.14	–	–	[–]	–	[–]	–
sy	42.79	42.79	[–]	–	[–]	–	[–]	–
st	41.76	41.76	[–]	–	[–]	–	[–]	–
i / sy	32.77	32.83	[-13.92]	36.79	[-0.70]	32.21	[-0.76]	32.19
i / st	32.49	32.55	[-13.92]	36.36	[-0.90]	31.90	[-0.96]	31.88
sy / st	30.01	30.05	[-12.32]	33.65	[-0.26]	30.05	[-0.27]	30.05
i / sy / st	25.95	26.01	[-18.98]	33.81	[-0.69]	25.64	[-0.73]	25.63
All species								
i	28.53	28.49	[–]	–	[–]	–	[–]	–
sy	21.62	21.62	[–]	–	[–]	–	[–]	–
st	19.69	19.69	[–]	–	[–]	–	[–]	–
i / sy	17.88	17.91	[-2.48]	18.83	[-0.83]	17.44	[-0.89]	17.44
i / st	17.23	17.25	[-2.40]	17.46	[-0.78]	16.55	[-0.84]	16.54
sy / st	14.71	14.69	[-2.44]	15.95	[-0.35]	14.69	[-0.35]	14.69
i / sy / st	13.63	13.65	[-3.32]	14.91	[-0.70]	13.25	[-0.74]	13.25

SC Single-count estimator; *MC* Multiple-count estimator; *LPlot* Linear combination weighted by plot variances; *LPSC* Linear combination weighted by pooled SC-variances; *LPMC* Linear combination weighted by pooled MC-variances

$$\hat{\mathrm{V}}_{Plot}\left(\hat{Y}_{MC}^{(P_d)}, n^*\right) = \frac{n^*}{n^* - 1} \sum_{\mathbb{X}_d^{(k)} \in P_d} \left(y_d^{(k)} - \hat{y}_d^*(k, n^*)\right)^2,$$

$$\hat{y}_d^*(k, n^*) = \frac{1}{n^*} \sum_{\mathbb{X}_d^{(l)} \in P_d^*(k)} y_d^{(l)},$$

where $P_d^*(k)$ is the set of n^* sample points of design d closest to $\mathbb{X}_d^{(k)}$. For this simulation, the fixed number of neighbors was set to $n^* = 4$.

The results, presented in Table 2, show that while any combination reduced the variance in the estimator, the combination based on plot variance estimates introduced bias at least three times of that generated by the pooled variance estimates. Because of the relatively small probability of two sample points sampling the same tree, the SC and MC estimators perform similarly.

In Table 3, bias, MSE, and variance estimates are presented for the i.i.d. 1 and 2 designs, and the combinations of the two. Comparing the combined samples versus the combined estimates, one can observe the trade-off between unbiased estimates and estimates with reduced variances.

6. DISCUSSION

In Table 2, we showed that combined samples and linear combinations based on pooled variances (pooled combination) will probably always be preferable to linear combinations

Table 3. Results from 10,000 simulations for the i.i.d. 1 and 2 designs showing [empirical relative bias] in percent, mean variance estimates, and empirical mean-squared error (MSE) for birches and all species

	Estimator	Rel. bias	Mean var. (10^4)	MSE (10^4)
Birches				
i.i.d. 1	SC	[–]	26.08	26.02
	MC	[–]	26.16	25.95
i.i.d. 2	SC	[–]	13.91	14.25
	MC	[–]	13.96	14.21
i.i.d. 1 / 2	SC	[–]	9.93	10.07
	MC	[–]	9.99	10.12
	LMC	[-12.61]	6.63	12.15
	LPSC	[-3.83]	8.71	9.08
	LPMC	[-3.97]	8.74	9.07
All species				
i.i.d. 1	SC	[–]	1675.85	1716.50
	MC	[–]	1671.77	1711.94
i.i.d. 2	SC	[–]	640.74	646.99
	MC	[–]	639.36	645.09
i.i.d. 1 / 2	SC	[–]	573.51	589.58
	MC	[–]	573.24	591.06
	LMC	[-2.03]	437.48	538.30
	LPSC	[-2.03]	454.02	506.76
	LPMC	[-2.19]	453.07	507.65

SC Single count estimator; *MC* Multiple count estimator; *LMC* Linear combination weighted by estimated variances; *LPSC* Linear combination weighted by pooled SC-variances; *LPMC* Linear combination weighted by pooled MC-variances

based on individual variances (naive combination), given that the target variable has a skewed distribution. Even if no correlation exists between the estimator and its variance estimator, the pooled combination should be more efficient than the naive combination, as more information is used. The main drawback of the pooled combination is the need to compute additional second-order design properties, which may be difficult if positional data is not available or accurate enough to map the sample properties of the designs. Furthermore, for some designs the pooled variance estimator might be unstable, which makes it an unsuitable choice for such designs. However, the combined samples approach will function sufficiently in most cases, as its estimate is not dependent on second-order design properties, why the impact of absence of reliable positional data should be small, for most designs.

While the results from the simulation are conditional to the simulated population, we expect the bias to be proportional to the heterogeneity of the population, why we may draw some general conclusions. We believe both of these methods to be useful for domain estimates. For the domain estimate of a primary survey, the target variable will have a skewed distribution, even if the target variable over the domain is not. It is thus expected that significant bias will be introduced by using the naive combination.

Another scenario where both presented methods might be useful are when combining designs like those used in the simulation here, where it is not possible to get an unbiased variance estimator for one or more of the individual designs. The pooled combination is unbiased if the combined second-order sample properties are positive for all units in the

population, whereas the naive combination needs positive second-order sample properties for all units and all designs. Furthermore, the combined samples approach has none of these restrictions and is also more relaxed in terms of first-order sample properties.

Table 3 provides results regarding MSE and variance estimates for i.i.d. designs. These results highlight the bias–variance trade-off between the pooled combination and the combined sample approaches. The combined samples approach produces unbiased estimators, however, in the simulation, with larger empirical mean-squared errors than the pooled combinations. A statistician deciding between these two approaches should thus know to what extent the end product needs to be accurate or reliable.

In Tables 2 and 3, we see that the bias is, as expected, more apparent when dealing with skewed target variables, as the volume of birch. It is not uncommon to reach acceptable MSE's for some dominant or aggregate target variable in a primary survey, here represented by the total wood volume, while needing complementary surveys to study some target variable with a more skewed distribution. The results of the simulation show that different methods of combination will affect the reliability of the combined estimates.

Further research would study the effects of errors in the positioning of units, to see how previously described mismatching would affect the estimates. For plot sampling procedures, that are commonly used in forest inventories, one can assume two types of mismatching to be common: One where there is a difference between the location of the studied plot and the sampled location, and one where the positioning of units within a plot are inaccurate. Depending on designs, these errors will have different effects.

ACKNOWLEDGEMENTS

We would like to thank the two anonymous reviewers and the associate editor for their helpful comments.

Funding Open access funding provided by Swedish University of Agricultural Sciences.

[Received February 2020. Accepted November 2020. Published Online January 2021.]

APPENDIX: UNIT DESIGN PROPERTIES

Let U be a finite, unknown population, representable by fixed points on an area of interest F_U. If a sample point $\mathbb{X}^{(k)}$, with probability density function (pdf) $f^{(k)}(\mathbf{x})$, falls within the inclusion zone $A_i^{(k)}$ of unit $i \in U$, the unit is included in the sample.

Let P be the set of independent sample points. For any sample point $\mathbb{X}^{(k)} \in P$, and units $\{i, j\} \in U$, we make the following definitions:

$$S_i^{(k)} := \mathrm{I}\left(\mathbb{X}^{(k)} \in A_i^{(k)}\right), \tag{23}$$

$$\pi_i^{(k)} := \Pr\left(S_i^{(k)} > 0\right) = \int_{A_i^{(k)}} f^{(k)}(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{24}$$

$$\pi_{ij}^{(k)} := \Pr\left(S_i^{(k)} > 0, S_j^{(k)} > 0\right) = \int_{A_i^{(k)} \cap A_j^{(k)}} f^{(k)}(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{25}$$

$$E_i^{(k)} := \mathrm{E}\left[S_i^{(k)}\right] = \pi_i^{(k)}, \tag{26}$$

$$E_{ij}^{(k)} := \mathrm{E}\left[S_i^{(k)} S_j^{(k)}\right] = \pi_{ij}^{(k)}, \tag{27}$$

where $\mathrm{I}(\cdot)$ denotes the indicator function, $S_i^{(k)}$ is the number of inclusions of unit i by sample point $\mathbb{X}^{(k)}$, $\pi_i^{(k)}$ is the first-order inclusion probability of unit i by sample point $\mathbb{X}^{(k)}$, i.e., the probability of unit i being included into the sample by a sample point $\mathbb{X}^{(k)}$, $\pi_{ij}^{(k)}$ is the second-order inclusion probability for units i, j by sample point $\mathbb{X}^{(k)}$, $E_i^{(k)}$ is the (first-order) expected number of inclusions of unit i by $\mathbb{X}^{(k)}$, and $E_{ij}^{(k)}$ is the second-order expected number of inclusions of units i, j by $\mathbb{X}^{(k)}$.

For a set of independent but not necessarily equally distributed sample points P, we extend the definitions to

$$S_i^{(P)} := \sum_{\mathbb{X}^{(k)} \in P} S_i^{(k)}, \tag{28}$$

$$\pi_i^{(P)} := \Pr\left(S_i^{(P)} > 0\right), \tag{29}$$

$$\pi_{ij}^{(P)} := \Pr\left(S_i^{(P)} > 0, S_j^{(P)} > 0\right), \tag{30}$$

$$E_i^{(P)} := \mathrm{E}\left[S_i^{(P)}\right], \tag{31}$$

$$E_{ij}^{(P)} := \mathrm{E}\left[S_i^{(P)} S_j^{(P)}\right]. \tag{32}$$

It follows quite clearly from (31), (28), and (26) that

$$E_i^{(P)} = \sum_{\mathbb{X}^{(k)} \in P} E_i^{(k)} = \sum_{\mathbb{X}^{(k)} \in P} \pi_i^{(k)},$$

and by expanding (29), we can express it in terms of (24)

$$\begin{aligned}\pi_i^{(P)} &= 1 - \Pr\left(S_i^{(P)} = 0\right) = 1 - \Pr\left(\bigcap_{\mathbb{X}^{(k)} \in P} S_i^{(k)} = 0\right) \\ &= 1 - \prod_{\mathbb{X}^{(k)} \in P} \left(1 - \pi_i^{(k)}\right).\end{aligned}$$

Through some work, we can get the second-order expected number of inclusions for units i, j by the set of sample points P

$$\begin{aligned} E_{ij}^{(P)} &= \mathrm{E}\Big[\sum_{\mathbb{X}^{(k)}\in P} S_i^{(k)} \sum_{\mathbb{X}^{(k')}\in P} S_j^{(k')}\Big] = \sum_{\mathbb{X}^{(k)}\in P} \mathrm{E}\left[S_i^{(k)} S_j^{(k)}\right] \\ &\quad + \sum_{\substack{\mathbb{X}^{(k)}\in P,\ \mathbb{X}^{(k')}\in P \\ k\neq k'}} \sum \mathrm{E}\left[S_i^{(k)} S_j^{(k')}\right] \\ &= \sum_{\mathbb{X}^{(k)}\in P} E_{ij}^{(k)} + \sum_{\substack{\mathbb{X}^{(k)}\in P,\ \mathbb{X}^{(k')}\in P \\ k\neq k'}} \sum E_i^{(k)} E_j^{(k')} = E_i^{(P)} E_j^{(P)} \\ &\quad + \sum_{\mathbb{X}^{(k)}\in P} \left(E_{ij}^{(k)} - E_i^{(k)} E_j^{(k)}\right), \end{aligned}$$

due to the independence of sample points in P. For the second-order inclusion probability for units i, j by the set of sample points P, we start by showing that

$$\begin{aligned} \pi_{ij}^{(P)} &= \Pr\left(S_i^{(P)} > 0\right) + \Pr\left(S_j^{(P)} > 0\right) \\ &\quad - \Pr\left(S_i^{(P)} > 0 \cup S_j^{(P)} > 0\right) \\ &= \pi_i^{(P)} + \pi_j^{(P)} - \left(1 - \Pr\left(S_i^{(P)} = 0, S_j^{(P)} = 0\right)\right). \end{aligned} \tag{33}$$

Through the independence between sample points in P, the following equality holds

$$\Pr\left(S_i^{(P)} = 0, S_j^{(P)} = 0\right) = \prod_{\mathbb{X}^{(k)}\in P} \Pr\left(S_i^{(k)} = 0, S_j^{(k)} = 0\right),$$

and conversely, apparent from (33), we have

$$\Pr\left(S_i^{(k)} = 0, S_j^{(k)} = 0\right) = 1 + \pi_{ij}^{(k)} - \pi_i^{(k)} - \pi_j^{(k)},$$

leading to

$$\pi_{ij}^{(P)} = \pi_i^{(P)} + \pi_j^{(P)} - \left(1 - \prod_{\mathbb{X}^{(k)}\in P} \left(1 - \pi_i^{(k)} - \pi_j^{(k)} + \pi_{ij}^{(k)}\right)\right).$$

REFERENCES

Allard A (2017) NILS—a nationwide inventory program for monitoring the conditions and changes of the Swedish landscape. In: Diaz-Delgado R, Lucas R, Hurford C (eds) The roles of remote sensing in nature conservation. Springer International Publishing, Cham, pp 79–90

Axelsson A, Ståhl G, Söderberg U, Petersson H, Fridman J, Lundström A (2010) Sweden. In: Tomppo E, Gschwantner T, Lawrence M, McRoberts R (eds) National forest inventories: pathways for common reporting. Springer, Dordrecht, pp 541–553

Benedetti R, Piersimoni F, Postiglione P (2015) Sampling spatial units for agricultural surveys. Springer, Berlin

Christensen P, Ringvall AH (2013) Using statistical power analysis as a tool when designing a monitoring program: experience from a large-scale Swedish landscape monitoring program. Environ Monit Assess 185(9):7279–7293

Fecso R, Tortora RD, Vogel FA (1986) Sampling frames for agriculture in the United States. J Off Stat 2(3):279–292

Fridman J, Holm S, Nilsson M, Nilsson P, Ringvall AH, Ståhl G (2014) Adapting National Forest Inventories to changing requirements - the case of the Swedish National Forest Inventory at the turn of the twentieth century. Silva Fenn 48(3):1–29

Grafström A, Ekström M, Jonsson BG, Esseen P-A, Ståhl G (2019) On combining independent probability samples. Surv Methodol 45(2):349–364

Grafström A, Schelin L (2013) How to select representative samples. Scand J Stati 41(2):277–290. https://doi.org/10.1111/sjos.12016

Hansen MH, Hurwitz WN (1943) On the theory of sampling from finite populations. The Ann Math Stat 14(4):333–362. https://doi.org/10.1214/aoms/1177731356

Horvitz D, Thompson D (1952) A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 47(260):663–685. https://doi.org/10.2307/2280784

Lohr S, Rao JK (2006) Estimation in multiple-frame surveys. J Am Stat Assoc 101(475):1019–1030. https://doi.org/10.1198/016214506000000195

Reese H, Nilsson M, Pahlén TG, Hagner O, Joyce S, Tingelöf U, Egberth M, Olsson H (2003) Countrywide estimates of forest variables using satellite data and field data from the National Forest Inventory. AMBIO A J Hum Environ 32(8):542–548. https://doi.org/10.1579/0044-7447-32.8.542