Design-based sampling methods for environmental monitoring

Xin Zhao
Faculty of Forest Sciences
Department of Forest Resource Management
Umeå

Doctoral Thesis
Swedish University of Agricultural Sciences
Umeå 2021

Acta Universitatis agriculturae Sueciae
2021:51

ISSN 1652-6880
ISBN (print version) 978-91-7760-780-9
ISBN (electronic version) 978-91-7760-781-6

Print: SLU Service/Repro, Uppsala 2021

Design-based sampling methods for environmental monitoring

Abstract

Efficient strategies for environmental monitoring are proposed with an emphasis on the importance of using available information. In environmental monitoring, it is common to use area frames covering the assumed spread of the population of interest. By using such a frame, a sample unit is usually not a unit in the population, rather a point on a surface. The population unit of environmental surveys exits in a spatial context where nearby units often have similar values to each other. When this is the case, we can estimate the unknown population parameters more efficiently if the sample is well spread over the population. Spatially balanced sampling is sampling designs that employ available auxiliary variables to select well-spread samples. When applying such a design with equal inclusion probabilities, we match the sample distribution to the population distribution of the auxiliary variables, which can improve the estimation of the state of the population. Paper I presents a new sampling strategy for the Swedish national forest inventory using spatially balanced sampling designs for an area frame. When estimating change, we wish to update the sample at the following occasions using the most recently available information. When updating the sample, we also want to have a certain degree of overlap between the successive samples. By doing so, we can get more precise estimates for states and the change between two states simultaneously. Therefore, there is a demand for selecting well-spread and partially overlapping samples over time. In Papers II and III, the focus is on such samples, and more specifically, on positively coordinated and spatially balanced samples. In Paper II, a sampling strategy of selecting positively coordinated and spatially balanced samples is proposed for monitoring the change of environmental variables, while the objective of Paper III is to estimate the variance of an estimator of change using such samples. When a single survey does not provide sufficient quality of estimates for some domain, we can plan for a complementary survey or combine existing surveys to improve the quality. When multiple surveys are combined, there is a risk of introducing bias to the estimators. Combining several surveys to use all available information when estimating the population parameters thus becomes a challenge. In Paper IV, we investigate the possibility of producing less biased or unbiased estimators when combining several independent surveys of a finite population.

Keywords: auxiliary variables, sampling strategy, area frame sampling, inclusion probabilities, spatially balanced sampling designs, the local pivotal method, spatially correlated Poisson sampling, sample coordination, combining samples.

Author's address: Xin Zhao, SLU, Department of Forest Resource Management, Skogsmarksgränd, SE-901 83 Umeå, Sweden
E-mail: xin.zhao@slu.se

Similar things group together, similar people fit together.

Contents

List of Publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I Grafström, A., Zhao, X., Nylander, M. & Petersson, H. (2017). A new sampling strategy for forest inventories applied to the temporary clusters of the Swedish national forest inventory. *Canadian Journal of Forest Research* 47(9), 1161-1167.

II Zhao, X., & Grafström, A. (2020). A sample coordination method to monitor totals of environmental variables. *Environmetrics*, 31(6), e2625.

III Zhao, X., & Grafström, A. (2021). Estimation of change with partially overlapping and spatially balanced samples. (submitted)

IV Prentius, W., Zhao, X., & Grafström, A. (2021). Combining Environmental Area Frame Surveys of a Finite Population. *JABES*, 26(2), 250-266.

Paper I is reproduced with the permission of the publisher. Papers II and IV are published as open-source articles.

The contribution of Xin Zhao to the papers included in this thesis was as follows:

I Planned the study together with the coauthors. Wrote large parts of the paper. Performed the Monte-Carlo simulations.

II Planned the study and developed the new strategy together with the coauthor, performed the Monte-Carlo simulation, wrote most of the paper, was responsible for the review process.

III Planned the study together with the coauthor, performed the Monte-Carlo simulation, wrote most of the paper.

IV Planned the study together with the coauthors. Generated the population based on field data from the NFI used for the simulation, participated in writing the manuscript and in the review process.

Abbreviations

GRTS	Generalized Random Tessellation Stratified
HH	Hansen-Hurwitz
HT	Horvitz-Thompson
i.i.d	Independent and Identically Distributed
LPM	Local Pivotal Method
MC	Multiple Count
MSE	Mean Squared Error
NFIs	National Forest Inventories
PRNs	Permanent Random Numbers
RB	Relative Bias
RMSE	Root Mean Squared Error
RRMSE	Relative Root Mean Squared Error
SC	Single Count
SCPS	Spatial Correlated Poisson Sampling
SIR	Simple Random Sampling with Replacement

1 Introduction

Environmental monitoring should provide reliable information that is necessary for taking proper decisions on the management of natural resources. Since the population is often very large, it is unrealistic to observe the entire population in practice. A sample survey is usually applied for selecting some parts of a population instead. The sample will then be observed to obtain estimates of the unknown population characteristics or parameters. A variable of interest is often called a target variable. The unknown parameters are functions of the target variables. They can be, for example, totals or means of the target variables. The values of the target variables are not available before a survey.

A sample survey can be either design-based or model-based. The differences between these two approaches are given in, e.g. Särndal *et al*. (1992) and de Gruijter *et al*. (2006). For this thesis, the focus is on the design-based approach. A design-based approach requires probability sampling, where a sample should be selected by a random mechanism. This random mechanism is called a sampling design. In probability sampling, every unit in the population has a chance to be selected. It provides an objective way of sample selection and yields estimates whose uncertainty can be evaluated. To produce a proper estimate of the unknown population parameter, we also need an estimator, i.e., a formula that can be used to calculate the estimated value based on the observations in the sample.

The combination of a sampling design and an estimator is defined as a sampling strategy. The choice of sampling strategy should attempt to produce estimates as close as possible to the unknown parameters, taking into account the available information about the population. To know if a sampling strategy is efficient or not, we need to study the problem of choosing a sampling design and an estimator. The main challenge in constructing a good sampling strategy lies in using available information about the population wisely. This thesis deals with this issue and aims at proposing new sampling methods for environmental monitoring that take advantage of the available information.

There are times when information is available for the whole population. A variable that is known for every unit in the population is called an auxiliary variable. The auxiliary variables usually have some connection with the target variables, i.e. they can explain the variations in the target variables to some extent. Therefore, they can be employed to improve the sampling design or to enhance the estimation of the population parameters.

Sampling designs that apply auxiliary variables are, for example, *pps* sampling, Poisson sampling, stratified sampling, and balanced sampling, etc.

Including the auxiliary variables in the design normally reduces the need for including the same variables in the estimators and can then allow for more straightforward estimation and analysis. If the auxiliary variables explain some of the variation in the target variables, then we can improve the estimation of the population parameters by using these auxiliary variables in the sampling design.

Natural resources share a general feature of our environment, which is that nearby things often are more similar than distant things. This feature is also known as Tobler's first law of geography. As a result, when sampling an environmental population, the selection of nearby units should be avoided to ensure that we capture the variation of the population in the sample. The geographical position of the sample units is important, e.g., in agricultural and environmental surveys, because the units themselves are defined using spatial criteria. Spreading the sampling effort evenly across the study area is an effective strategy for environmental monitoring. One of the popular methods which take advantage of the spatial nature is called generalized random tessellation Stratified (GRTS) design proposed by Stevens & Olsen (2004).

Recently, it has been shown that it will be more efficient to make sure that the sample is well spread, both geographically and in other available auxiliary variables, see, e.g., Grafström & Schelin (2014). The local pivotal method (LPM) (Grafström *et al.*, 2012) and the spatially correlated Poisson sampling (SCPS) (Grafström, 2012) are two sampling designs that can select well-spread and representative samples in multiple dimensions. We need to notice that a representative sample does not need to be a probability sample, but only a probability sample allows for proper design-based inference. A sample that consists of approximately 80% spruces and 20% pines will be representative if we know that the population consists of 80 spruces and 20 pines. The selection of representative samples in multiple dimensions means that we match as closely as possible the sample distribution of a set of auxiliary variables to the population distribution of the set of auxiliary variables. According to Grafström & Schelin (2014), equal inclusion probabilities should be applied in the case of selecting a representative and well-spread probability sample. A sampling design that uses auxiliary variables to select representative samples is particularly useful for multipurpose environmental monitoring programs, such as National forest inventories(NFIs). When several target variables are of interest, there is a need to spread the samples evenly with respect to all target variables. Sampling designs such as GRTS, LPM, and SCPS are called spatially balanced sampling designs.

In environmental monitoring, a common focus is to track the change in a population over time. Therefore, a survey of the population commonly needs

to be repeated over time. For environmental surveys, a question of whether we should use independent samples, a permanent sample, or partially overlapping samples over time arises. To judge whether the strategy of monitoring change is efficient or not, we can compare the variance of an estimator of change for different strategies and different populations. It is well known that the variance of an estimator of change equals the sum of the variance of the two state estimators minus two times the covariance between them. For independent samples, we do not need to consider the covariance, thus simplifying the estimation problem. However, it will not be the best strategy to use independent samples when estimating change. This is because the variance of the change estimator becomes about twice the variance of the estimator of the state when using independent samples if the samples selected use the same design and sample size. When the time between surveys is short and the target variables have not changed much, a permanent sample might be employed to reduce the variance of an estimator of change. However, as the values of the target variables are likely to change over time, a permanent sample is not likely to be as representative as it used to be as the time between surveys increases. Thus, we then tend to have a larger variance of the state estimator at the second time occasion for permanent samples. Even if the covariance between the two state estimators becomes large by having fully overlapping samples, it is not guaranteed that the variance of change will be reduced. There is a need for updating the sample at the next time occasion to account for changes while retaining as many units as possible from the old sample. By doing so, we can improve the estimation of change.

Since spatially balanced sampling designs are often more efficient than, e.g., simple random sampling, it will be too conservative to apply the traditionally used variance estimators e.g., Berger 2004; Hájek 1964; Hartley & Rao 1962; Horvitz & Thompson 1952; Yates & Grundy 1953. Therefore, there is a need to develop variance estimators which are more suitable to be applied under spatially balanced sampling designs where no unbiased variance estimator already exists. Grafström & Schelin (2014) proposed a local mean variance estimator which was shown to perform well under such sampling designs. For repeated surveys, researchers have also paid a lot of attention to the estimation of covariance. Tam (1985) was one of the earliest studies that considered covariance estimation with overlapping samples. Qualité (2009, ch.5) derived covariance estimators based on two overlapping samples by considering sampling designs that can obtain rotating panels. Under such a design, only part of the sample in the previous occasion will be maintained in the sample at the next time occasion. New units will be selected in the sample at the next time occasion to replace the discarded units.

In paper III, inspired by the variance estimator in Grafström & Schelin (2014) and the covariance estimator in Qualité (2009, ch.5), we further develop a local mean covariance estimator which is suitable when estimating the change with spatially balanced and partially overlapping samples.

When merging available information to improve the estimation of population characteristics, there is also a need to produce unbiased or less biased estimators. An estimator is called an unbiased estimator if on average (over all possible samples) the estimator is equal to the population parameter. In presence of several independent probability samples from a finite population, if we estimate the population total by a linear combination of separate independent estimators, then the estimator can be biased as the separate estimators of the population totals can be highly correlated to their respective variance estimators. Thus, we need to find alternative ways to improve estimation when combining information from multiple samples. In Paper IV, we propose to estimate the population total based on either a pooled linear combination or on an unbiased estimator of the combined sample.

For a finite population, since we often have a list frame covering the population, the sample units are often the units in the population. It is straightforward to select and observe a sample that consists of the units of the population, thus making the inference easier. However, for environmental surveys, we seldom know the number of units in the population, which makes it impossible to construct a list of all units in the population.

Instead, an area frame that covers the population will be employed for the sample selection. One scenario may be to tessellate the area and construct a finite number of cells. Another way is to treat the area as a continuous population with an infinite number of points. For both scenarios, a sample unit is not a population unit anymore, it becomes an area or a point. Each sample unit may contain a different number of population units. In the first scenario, each population unit belongs to only one sample unit. Note that, in the second scenario, a point may represent a cluster, i.e. we select points within the area frame, then put one cluster centered on each selected point. The clusters can have different shapes and sizes, within the frame they have positions and may also have orientations (de Gruijter *et al*., 2006). In this case, each population unit may belong to an infinite number of points, because the clusters may overlap. All of this in general makes sampling natural resources very complex.

Traditionally, the design-based method focused on sampling from finite populations, and therefore the representation of the universe is discrete in this approach. To apply the methods for a finite population when sampling from an area frame, we can convert the continuous population to a finite popula-

tion. An area frame may consist of a finite number of area segments, e.g., the frame can be discretized by a fine grid of which the centers of the grid cells represent the possible sample locations. Since the population becomes the collection of the centers of these grid cells, we get a list frame. The methods for finite population sampling can thus be applied. Sampling from an area frame can also be accomplished by randomly selecting points and then observing the segments of an area with the points as the area center, which is also the method we use throughout the thesis. In Papers I and II, we applied a double (two-phase) sampling approach, where the samples are selected in two steps. In the first step, a very large initial sample with independent points is selected from an area frame, then auxiliary information is extracted from the initial sample. For the second step, methods for finite population sampling are applied. We used the auxiliary information and applied LPM or SCPS to select a well-spread subsample as our final sample. In Paper IV, sample points are selected directly by using three commonly used sampling designs. To combine several samples and find unbiased variance estimators, the sample properties are derived for the finite population units.

2 Sampling from a finite population

Let $U = \{1, 2, ..., i, ..., N\}$ denote a population of size N. A variable of interest y has the value y_i for unit i. We want to select a probability sample S of size n from the population to estimate a population parameter. For a probability sample, every unit in the population has a nonzero probability of being selected, and the probability must be known for at least the selected units. The sample S can be selected with or without replacement. A sample selected without replacement is always a subset of U, whereas for a with replacement sampling design the same units may be selected multiple times in a sample. A possible realization of a with replacement sample can be, e.g., $s = (3, 1, 3, 6)$, where the unit 3 is selected two times in the sample. Denote the collection of all possible samples to obtain with a sampling procedure (with or without replacement) as the set $\mathbf{S} = \{s_1, s_2, ..., s_m\}$.

2.1 Sampling design and inclusion probabilities

A sampling design is a probability distribution on $\mathbf{S}$. Under probability sampling, there exists a function $p(\cdot)$ such that $p(s)$ is the probability of selecting the sample $s \in \mathbf{S}$, i.e. $\Pr(S = s) = p(s)$ for any $s \in \mathbf{S}$. The function $p(\cdot)$ is called the sampling design. It has the properties $p(s) > 0$ for any $s \in \mathbf{S}$ and $\sum_{s \in \mathbf{S}} p(s) = 1$.

Given a sampling design $p(\cdot)$, the inclusion of unit i in a probability sample S is indicated by a random variable I_i. We call I_i an inclusion indicator of unit i. It takes the value $I_i = 1$ if $i \in S$ and $I_i = 0$ otherwise. The probability that a unit i will be sampled is defined as

$$\pi_i = \Pr(I_i = 1) = E(I_i) = \sum_{s \in \mathbf{S}} I(i \in s) p(s).$$

Here π_i is called a first-order inclusion probability of unit i. The inclusion probabilities can be equal for all units in the population or proportional to an auxiliary variable. For sampling designs which select a sample with fixed size n, we have $\sum_{i \in U} \pi_i = n$. The probability that two units i and j are included in the sample is denoted as

$$\pi_{ij} = \Pr(I_i = 1, I_j = 1) = E(I_i I_j) = \sum_{s \in \mathbf{S}} I(i \in s) I(j \in s) p(s),$$

and π_{ij} is called a second-order inclusion probability.

2.2 Estimators and their properties

We can estimate the population parameter by using the observations of the units in a sample. The rule (function) of calculating an estimate of a population parameter based on the observations in a sample is known as an estimator. Suppose the population parameter that we want to estimate is θ, the estimator of θ is then denoted by $\widehat{\theta}$. The expected value of $\widehat{\theta}$ is $E(\widehat{\theta})$ and under a sampling design $p(\cdot)$ we have

$$E(\widehat{\theta}) = \sum_{s \in \mathbf{S}} \widehat{\theta}(s) p(s),$$

where $\widehat{\theta}(s)$ is the value of $\widehat{\theta}$ given the sample s. Under the same design the variance of $\widehat{\theta}$ can be expressed as

$$V(\widehat{\theta}) = \sum_{s \in \mathbf{S}} \left(\widehat{\theta}(s) - E(\widehat{\theta}) \right)^2 p(s).$$

Suppose the population parameter that we want to estimate is the population total $Y = \sum_{i \in U} y_i$. Then Y can be estimated by the Horvitz-Thompson (HT) estimator (Horvitz & Thompson, 1952)

$$\widehat{Y} = \sum_{i \in U} \frac{y_i}{\pi_i} I_i. \tag{1}$$

It is also named as the single-count (SC) estimator in paper IV. If the sampling design is without replacement, then (1) can be written as

$$\widehat{Y} = \sum_{i \in S} \frac{y_i}{\pi_i}.$$

The variance of (1) can be shown to be

$$V(\widehat{Y}) = \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}. \tag{2}$$

An estimator of the variance (2) is

$$\widehat{V}(\widehat{Y}) = \sum_{i \in S} \sum_{j \in S} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}. \tag{3}$$

Under a general sampling design, a unit may be included multiple times in a sample. Let S_i denote the number of inclusions of unit i in the sample S. For any sampling design, the population total Y can be estimated by Hansen-Hurwitz (HH) estimator (Hansen & Hurwitz, 1943)

$$\widehat{Y} = \sum_{i \in U} \frac{y_i}{\mu_i} S_i, \tag{4}$$

Where μ_i is the expected number of inclusions and $\mu_i = \mathrm{E}(S_i)$. When having a without-replacement sampling design, we have $\mu_i = \pi_i$. The estimator (4) is also called the multiple-count (MC) estimator in paper IV.

The variance of (4) is

$$V(\widehat{Y}) = \sum_{i \in U} \sum_{j \in U} (\mu_{ij} - \mu_i \mu_j) \frac{y_i}{\mu_i} \frac{y_j}{\mu_j}, \tag{5}$$

where $\mu_{ij} = \mathrm{E}(S_i S_j)$ is the second-order of expected inclusions. An estimator of (5) is

$$\widehat{V}(\widehat{Y}) = \sum_{i \in U} \sum_{j \in U} (\mu_{ij} - \mu_i \mu_j) \frac{y_i}{\mu_i} \frac{y_j}{\mu_j} \frac{S_i S_j}{\mu_{ij}}. \tag{6}$$

Sometimes we need to estimate the change of the population total between two time occasions $\Delta = Y_2 - Y_1$, where Y_t is the population total at time t. The estimator of Δ is given by $\widehat{\Delta} = \widehat{Y}_2 - \widehat{Y}_1$ where $\widehat{Y}_t$ represents the estimator of the population total at time t. To know the precision of the estimation, we also need to estimate the variance of the estimator of change. This variance is given by

$$V(\widehat{\Delta}) = V(\widehat{Y}_1) + V(\widehat{Y}_2) - 2C(\widehat{Y}_1, \widehat{Y}_2), \tag{7}$$

where $V(\widehat{Y}_t)$ is the variance of the estimator of the population total at time t and $C(\widehat{Y}_1, \widehat{Y}_2)$ is the covariance between the estimators. The covariance between two HT-estimators of two population totals is given by

$$C(\widehat{Y}_1, \widehat{Y}_2) = \sum_{i \in U} \sum_{j \in U} \left(\pi_{ij}^{12} - \pi_{i1}\pi_{j2}\right) \frac{y_{i1}}{\pi_{i1}} \frac{y_{j2}}{\pi_{j2}}, \tag{8}$$

where $\pi_{ij}^{12} = \Pr(i \in S_1, j \in S_2)$ is the probability of including unit i at time 1 and including unit j at time 2, S_t is the sample selected at time t, y_{it} is the value of y for unit i at time t.

Equation (7) implies to estimate the variance of the estimator of change, we need to estimate the variance of the separate state estimators and the covariance between the two estimators. The variance of the state estimator $V(\widehat{Y}_t)$ can be estimated by (3) or (6). The estimator of the covariance $C(\widehat{Y}_1, \widehat{Y}_2)$ can be expressed as

$$\widehat{C}(\widehat{Y}_1, \widehat{Y}_2) = \sum_{i \in S_1} \sum_{j \in S_2} \frac{\left(\pi_{ij}^{12} - \pi_{i1}\pi_{j2}\right)}{\pi_{ij}^{12}} \frac{y_{i1}}{\pi_{i1}} \frac{y_{j2}}{\pi_{j2}}. \tag{9}$$

Two common measures to assess the performance of an estimator are bias and mean squared error (MSE). Bias can be used to describe how much

an estimator deviates from the population parameter on average. For $\widehat{\theta}$ it is defined as

$$\text{Bias}(\widehat{\theta}) = E(\widehat{\theta}) - \theta.$$

When $\text{Bias}(\widehat{\theta}) = 0$ we say that $\widehat{\theta}$ is an unbiased estimator of θ. The HT- and the HH-estimators are such estimators. Under a without replacement sampling design, when $\pi_{ij} > 0$ for all pairs $\{i, j\} \in U$, the variance estimator (3) is an unbiased estimator of the variance of the HT-estimator (2). When $\mu_{ij} > 0$ for all pairs $\{i, j\} \in U$, the variance estimator (6) is also unbiased of (5). Similar to the variance estimators, the covariance estimator (9) is unbiased for (8) provided the π_{ij}^{12} are strictly positive for all i, j.

If $E(\widehat{\theta}) \neq \theta$, the estimator $\widehat{\theta}$ is said to be biased. $\text{Bias}(\widehat{\theta}) > 0$ means $\widehat{\theta}$ tends to overestimate θ and $\text{Bias}(\widehat{\theta}) < 0$ means $\widehat{\theta}$ tends to underestimate θ. When comparing different estimators, it is also useful to relate the size of the bias of an estimator to the value we are estimating. The ratio between the bias of an estimator and the value we are estimating is called relative bias (RB). For $\widehat{\theta}$, it is denoted as

$$\text{Relative Bias}(\widehat{\theta}) = \frac{\text{Bias}(\widehat{\theta})}{\theta} \cdot 100\%.$$

The MSE measures the average squared difference between the estimator and the true parameter value. For $\widehat{\theta}$ it is defined as

$$\text{MSE}(\widehat{\theta}) = E\left((\widehat{\theta} - \theta)^2\right) = V(\widehat{\theta}) + \left(\text{Bias}(\widehat{\theta})\right)^2.$$

If $\widehat{\theta}$ is unbiased for θ, we get $\text{MSE}(\widehat{\theta}) = V(\widehat{\theta})$. As we can see from the expression, the MSE incorporates both the variance and the bias of the estimator, thus it can be used to check the efficiency of an estimator. The smaller value of MSE implies a better estimator. As the MSE has a squared unit of measure, it is sometimes difficult to interpret. Instead, we can use the root mean squared error (RMSE) when interpreting the results, and we have

$$\text{RMSE}(\widehat{\theta}) = \sqrt{MSE(\widehat{\theta})}.$$

Similar to the relative bias, we may want to relate the size of the RMSE to the value we are estimating. The ratio between the RMSE of an estimator and the value we are estimating is called relative root mean squared error (RRMSE). It is defined as

$$\text{RRMSE}(\widehat{\theta}) = \frac{\text{RMSE}(\widehat{\theta})}{\theta} \cdot 100\%.$$

2.3 Sampling strategy and the nature of a good sampling strategy

A sampling strategy is the combination of a sampling design and an estimator. According to Royall (1970), a sampling strategy $\{p : \widehat{\theta}\}$ will be said to be better than strategy $\{p' : \widehat{\theta}'\}$, if

$$\mathrm{MSE}(\{p : \widehat{\theta}\}) < \mathrm{MSE}(\{p' : \widehat{\theta}'\}). \tag{10}$$

When the two sampling strategies employ unbiased estimators, the contribution to the MSE from the bias of the estimator will be zero in both sides in equation (10). Thus we say a sampling strategy $\{p : \widehat{\theta}\}$ is more efficient than strategy $\{p' : \widehat{\theta}'\}$, if

$$\frac{V(\{p : \widehat{\theta}\})}{V(\{p' : \widehat{\theta}'\})} < 1, \tag{11}$$

where $\widehat{\theta}$ and $\widehat{\theta}'$ are two unbiased estimators of θ. The ratio between the two variances is called the design effect. To ensure a fair comparison, we often require that the two strategies should select samples with the same expected sample size or same expected cost (Särndal *et al*., 1992).

3 Sampling from a continuous population

Suppose that a finite population U_t has its units scattered on a surface F_U, where F_U is a subset of the Euclidean plane R^2 with its surface area $\ell(F_U)$. An example of such a population can be trees in a forest stand. In such an area frame, we often do not have a list frame covering the population of trees. Thus, it is impossible, or at least not cost-efficient to sample the population units (trees) directly. For such a population, we can select a sample of points in the area frame. Then use an area with a fixed shape and size centered at the sample point in the field inventory.

The response of a target variable for a point $\mathbf{x} \in F_U$ can be denoted as $y(\mathbf{x})$, we have $y(\mathbf{x}) = 0$ if the point $\mathbf{x}$ is outside the frame F_U. The population total of the response for the target variable can hence be expressed as $Y = \int_{F_U} y(\mathbf{x})d\mathbf{x}$. In the area frame, we can introduce the sampling intensity function $\pi(\mathbf{x})$ which describes the expected number of sample points at location $\mathbf{x}$ (Cordy, 1993). We have $\int_{F_U} \pi(\mathbf{x})d\mathbf{x} = n$, $\pi(\mathbf{x}) > 0$ for any point $\mathbf{x}_k \in F_U$ and $\pi(\mathbf{x}) = 0$ for points outside F_U. Denote the random sample of n locations within F_U as $S = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$. The sampling intensity function is then given by $\pi(\mathbf{x}) = \sum_{k=1}^{n} f_k(\mathbf{x})$, where $f_k(\mathbf{x})$ is the marginal probability density function of $\mathbf{x}_k$. The second-order sampling intensity function can be expressed as $\pi(\mathbf{x}, \mathbf{x}') = \sum_{k=1}^{n} \sum_{l \neq k} f_{kl}(\mathbf{x}, \mathbf{x}')$, where $f_{kl}(\mathbf{x}, \mathbf{x}')$ is the joint probability density function for the pair of sample points $\{\mathbf{x}_k, \mathbf{x}_l\}$. The sampling intensity plays the same role as the inclusion probabilities play in finite population sampling.

According to Cordy (1993), the continuous version of the unbiased HT estimator of Y is given by

$$\widehat{Y} = \sum_{\mathbf{x} \in S} \frac{y(\mathbf{x})}{\pi(\mathbf{x})}.$$

When sample size is fixed, the variance of the estimator in Sen-Yates-Grundy form is

$$V_{SYG}(\widehat{Y}) = \frac{1}{2} \iint_{F_U} \left(\pi(\mathbf{x})\pi(\mathbf{x}') - \pi(\mathbf{x}, \mathbf{x}') \right) \left(\frac{y(\mathbf{x})}{\pi(\mathbf{x})} - \frac{y(\mathbf{x}')}{\pi(\mathbf{x}')} \right)^2 d\mathbf{x} d\mathbf{x}'. \quad (12)$$

An unbiased estimator of (12) is

$$\widehat{V}_{SYG}(\widehat{Y}) = \frac{1}{2} \sum_{\mathbf{x} \in S} \sum_{\mathbf{x}' \neq \mathbf{x}} \frac{\pi(\mathbf{x})\pi(\mathbf{x}') - \pi(\mathbf{x}, \mathbf{x}')}{\pi(\mathbf{x}, \mathbf{x}')} \left(\frac{y(\mathbf{x})}{\pi(\mathbf{x})} - \frac{y(\mathbf{x}')}{\pi(\mathbf{x}')} \right)^2,$$

provided that $y(\mathbf{x})$ is bounded, $\int_{F_U} 1/\pi(\mathbf{x})d\mathbf{x} < \infty$ and $\pi(\mathbf{x}, \mathbf{x}') > 0$ for all points $\mathbf{x}' \neq \mathbf{x}$ on F_U.

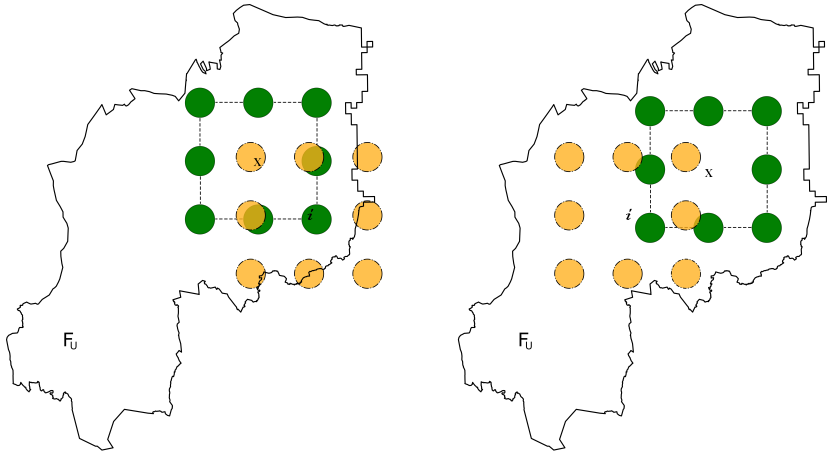


Figure 1: Example of inclusion zone. The green circles connected with dots represent a cluster of eight circular plots. The inclusion zone K_i for the unit i consists of the yellow circles. Any tract with its center $\mathbf{x}$ within K_i, such as the one in (a), includes unit i in one of the plots. If the point $\mathbf{x}$ does not fall into the inclusion zone K_i of the unit, as the case in (b), the tract will not include the unit.

As mentioned previously, the sample points may represent an area with a fixed size and shape. Such an area is often called a tract. When sampling tracts, each unit in the population has an inclusion zone. The inclusion zone K_i of a unit i is the collection of all points that lead to the inclusion of the unit. The unit will be included if a point falls within its inclusion zone. In the case of fixed area plots, the inclusion zones are usually of the same shape and size as the sample plot but centered at the units' locations. Figure 1 shows the inclusion zone of a unit for fixed area tracts. The inclusion zone is related to the response function and the shape of tract determines the response function. Denote the area of the inclusion zone for a population unit i as $\ell(K_i)$, the density function $y(\mathbf{x})$ of the target variable at point $\mathbf{x}$ is then given by

$$y(\mathbf{x}) = \sum_{i \in U_t} \frac{I_i(\mathbf{x}) y_i}{\ell(K_i)}, \tag{13}$$

where $I_i(\mathbf{x}) = 1$ if $\mathbf{x}$ is within the inclusion zone of element i and $I_i(\mathbf{x}) = 0$ otherwise. By using this expression, the continuous total is then corresponding to the finite population total. Because

$$Y = \int_{F_U} y(\mathbf{x}) d\mathbf{x} = \int_{F_U} \sum_{i \in U_t} \frac{I_i(\mathbf{x}) y_i}{\ell(K_i)} d\mathbf{x} = \sum_{i \in U_t} \frac{y_i}{\ell(K_i)} \int_{F_U} I_i(\mathbf{x}) d\mathbf{x} = \sum_{i \in U_t} y_i.$$

The above equality means we can get an unbiased estimator of Y by applying the response function (13).

As we can see from Figure 1, units near the border of F_U have parts of their inclusion zone outside F_U. This makes the estimation complicated. To simplify the problem, we can add a buffer zone around F_U, such that each population unit has an identical inclusion zone. The width of the buffer should be at least as big as the radius of the tract. By doing so, the sampling frame expands from F_U to F. Ideally, the sampling frame should provide full coverage and can uniquely identify all target population units. According to Särndal *et al*. (1992), there is a problem of frame imperfection under some circumstances. In this thesis, we assume that the area frames are always perfect.

We can convert the problem of sampling from a continuous population to sampling from a finite population. When using sampling designs that select samples of independent points, like the three designs in Paper IV, we can derive the inclusion probabilities and the expected number of inclusions of the units by a set of sample points and derive an unbiased estimator of Y as sampling from the finite population. However, it becomes more difficult to select samples with independent sample points, if we want to capture the spatial features of the environmental populations by using some auxiliary variables. In this case, we can apply a double sampling approach to select samples in two steps, see Papers I and II.

4 Some commonly used sampling designs

A probability sample can be selected in different ways and the sampling design describes the randomness included in the selection. Various sampling designs can be applied when monitoring environmental variables. Among others I will give details of the designs that are applied in this thesis.

4.1 Simple random sampling with replacement (SIR)

In SIR, we make independent selections of the units in the population and each unit has the same probability to be selected each time. Thus, it was also called i.i.d. sampling design in Paper IV. To use the sampling design in the sampling frame, we generate points uniformly on a rectangle that contains the area frame, if a point falls into the frame, we will accept it as a sample point. Note that, SIR is a special case of i.i.d when using equal inclusion probability $\pi_i = n/N$ (or constant sampling intensity $\pi(\mathbf{x}) = n/\ell(F)$).

4.2 Systematic sampling

In systematic sampling, there is an ordered list frame and a predefined sampling interval k. The first unit in the sample is drawn from a uniform distribution (randomly and with equal probability) among the first k elements in the list frame. The rest of the sample is decided by systematically choosing every kth unit from the list. There is only one random draw in the sample selection, which makes the design simple and thus easy to apply in practice. Note that, the sample size is in general not fixed under systematic sampling. When applying the design in F, we should define fixed distances d_x and d_y between sample points on two mutually perpendicular axes. We can select a fixed point, then generate two random numbers $r_x \sim U(0,d_x)$ and $r_y \sim U(0,d_y)$ and add the numbers to the x and y coordinates of the fixed point to randomize the position. The rest of the sample points will be given by the distances d_x and d_y. By doing so, the entire grid is randomly shifted with the uniform distribution.

4.3 Stratified sampling

Stratified sampling is a sampling method that employs auxiliary information. Because of its efficiency, it is widely used in practice. By stratification, the population is partitioned into several nonoverlapping strata according to one or some auxiliary variables. From each stratum, a random sample is selected independently. Often the same sampling design and estimation methods are

used in all strata. Nevertheless, different sampling and estimation methods can also be applied within strata which makes the method flexible.

5 Use of available information

Two ways of using available information to improve the estimation are considered in this thesis. In Papers I, II and III, the efficiency of spatially balanced sampling designs which employ auxiliary variables to spread the sample are studies. The focus of Paper IV is to merge several available independent samples to produce more efficient estimators.

5.1 Spatially balanced sampling designs

A spatially balanced sampling design is a sampling design that selects spatially balanced samples. A spatially balanced sample is a sample which is well spread in some auxiliary variables. Well spread means we avoid the selection of units that have similar values on the auxiliary variables. The intuition behind spatial balancing is that the auxiliary variables are related to the target response to be assessed. By spreading the sample in the auxiliary variables, we can select samples whose empirical distribution matches the population distribution of the auxiliary variables. Such samples are spatially balanced in the auxiliary space, leading to an approximate balance for any target y well explained by those auxiliary variables (Grafström & Lundström, 2013). Thus, for such targets we achieve $\widehat{Y} \approx Y$, and hence can get more precise estimates than designs that do not produce spatially balanced samples.

5.1.1 Spatial balance

Spatial balance is a measure to check the spread of a spatial sample. The measure is based on Voronoi polytopes (Stevens & Olsen, 2004). For a sample of size n, we need to construct n polytopes. For each $i \in S$, the polytope ρ_i includes all units in the population closer to i than to any other sample unit $j \in S$, $j \neq i$. The distance used when we construct the polytopes is the Euclidean distance on standardized variables. Ideally, if the sample is well spread, the total probability mass within ρ_i equals to 1. The spatial balance of a sample is expressed as

$$B = \frac{1}{n} \sum_{i \in S} (v_i - 1)^2, \tag{14}$$

where $v_i = \sum_{j \in \rho_i} \pi_j$ is the total probability mass within the polytope. Because B is a measure of the variance of the total probability mass within the polytopes, the smaller the value of B the better the spread of the sample is. To measure how well a design succeeds in selecting spatially balanced samples, simulation to find the expected value of B under the design is needed.

5.1.2 The Local Pivotal Method

The local pivotal method (LPM) is a spatially balanced sampling method proposed by Grafström *et al*. (2012), it has been shown to be one of the most effective methods for spreading the sample in the auxiliary space, e.g. (Benedetti *et al*., 2015, ch.7). When applying the LPM, spatial balance is achieved by successively updating the inclusion probabilities for nearby units until they become inclusion indicators, i.e. 0's and 1's, where the 0's indicate the exclusion of the units and 1's indicate the inclusion of the units. In one step of the LPM, we randomly select one unit i and find its nearest neighbor j. The pair of nearby units will compete with the (possibly updated) inclusion probabilities $0 < \pi_i < 1$ and $0 < \pi_j < 1$. The winner takes as much inclusion probability as possible from the loser. Thereafter, the winner has an updated inclusion probability $\pi_W = \min(1, \pi_i + \pi_j)$ while the loser has the new inclusion probability $\pi_L = \pi_i + \pi_j - \pi_W$. Thus, if $\pi_i + \pi_j \geq 1$, then $\pi_W = 1$, and the winner is included in the sample. If $\pi_i + \pi_j \leq 1$, then the loser is excluded from the sample. A final decision is made for at least one unit each step. The procedure for the competition is given by

$$(\pi_i', \pi_j') = \begin{cases} (\pi_W, \pi_L) & \text{with probability } \frac{\pi_W - \pi_j}{\pi_W - \pi_L} \\ (\pi_L, \pi_W) & \text{with probability } \frac{\pi_W - \pi_i}{\pi_W - \pi_L} \end{cases}, \tag{15}$$

where (π_i', π_j') denotes the new and updated probabilities for the pair. When nearby units compete for inclusion they are unlikely to be included simultaneously, which forces the sample becoming well spread. Figure 2 shows an example of the competition procedure for one step in a two-dimensional space.

5.1.3 Spatially correlated Poisson sampling

SCPS derived by Grafström (2012) is a list-sequential sampling method of selecting spatially balanced samples. It is a fixed size πps design that achieves a good spread of the selected samples by using auxiliary variables. Same as the LPM, it is guaranteed by creating a strong negative correlation between the inclusion indicators of nearby units. It is assumed that we have a list U of the units to be sampled. The sampling outcome is first decided for the first unit in the list and then for the second, etc. After each sampling decision, the inclusion probabilities for the remaining units in the list are updated. Denote the prescribed inclusion probability of each unit i as π_i, $i = 1, 2, ..., N$, with $\sum_{i=1}^{N} \pi_i = n$. Then, we have a starting vector of inclusion probabilities $(\pi_1, ..., \pi_N)$. In the end of the algorithm, we will get a vector of inclusion indicators by gradually updating the vector of inclusion probabilities in a max-

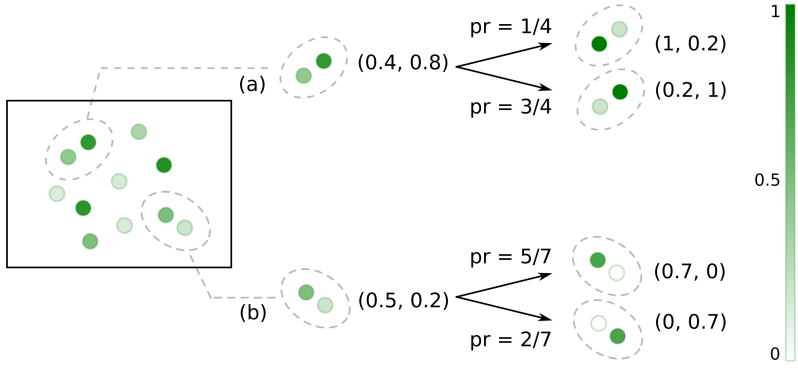


Figure 2: One step in the LPM for a pair of nearby units i and j. The intensity of the colour correlates with the inclusion probability. If $\pi_i + \pi_j \geq 1$ (case a), then the winner receives probability 1 and will definitely be included. If $\pi_i + \pi_j \leq 1$ (case b), then the loser receives probability 0 and will definitely be excluded.

imum of N steps. The updating can be illustrated as

$$\begin{array}{lccccccc}
\boldsymbol{\pi}^{(0)}: & \pi_1 & \pi_2 & \pi_3 & \pi_3 & \cdots & \pi_N \\
\boldsymbol{\pi}^{(1)}: & I_1 & \pi_2^{(1)} & \pi_3^{(1)} & \pi_4^{(1)} & \cdots & \pi_N^{(1)} \\
\boldsymbol{\pi}^{(2)}: & I_1 & I_2 & \pi_3^{(2)} & \pi_4^{(2)} & \cdots & \pi_N^{(2)} \\
\boldsymbol{\pi}^{(3)}: & I_1 & I_2 & I_3 & \pi_4^{(3)} & \cdots & \pi_N^{(3)} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\boldsymbol{\pi}^{(N)}: & I_1 & I_2 & I_3 & I_4 & \cdots & I_N
\end{array}. \tag{16}$$

The first unit is included with probability $\pi_1^{(0)} = \pi_1$. If the first unit is included, we set $I_1 = 1$, otherwise $I_1 = 0$. The sampling outcome at each step is decided by comparing the inclusion probability of the step unit with a random number associated with the unit. Denote the random number associated with the unit $i \in U$ as r_i, with $r_1, r_2, ..., r_N$ i.i.d. $U(0,1)$. When the values for $I_1, ..., I_{j-1}$ have been decided for the first $j-1$ units in the list, the step unit j is included in the sample S, i.e. $I_j = 1$, if $r_j < \pi_j^{(j-1)}$, and $I_j = 0$ otherwise. The inclusion probabilities for the rest of the units in the list are updated according to

$$\pi_i^{(j)} = \pi_i^{(j-1)} - \left(I_j - \pi_j^{(j-1)}\right) w_i^{(j)}, \tag{17}$$

where $i = j+1, ..., N$ and $i \geq 2$, $\pi_i^{(0)} = \pi_i$, $w_i^{(j)}$ is the weight received by unit i from the step unit j, $\sum_{i=j+1}^{N} w_i^{(j)} = 1$. The weight $w_i^{(j)}$, depends on the

sampling outcomes of the first $j-1$ units. To make sure that $0 \leq \pi_i^{(j)} \leq 1$ holds, the weights need to satisfy the following restrictions

$$-\min\left(\frac{1-\pi_i^{(j-1)}}{1-\pi_j^{(j-1)}}, \frac{\pi_i^{(j-1)}}{\pi_j^{(j-1)}}\right) \leq w_i^{(j)} \leq \min\left(\frac{\pi_i^{(j-1)}}{1-\pi_j^{(j-1)}}, \frac{1-\pi_i^{(j-1)}}{\pi_j^{(j-1)}}\right). \quad (18)$$

In the SCPS, the decision (including or not) is always made for the step unit j, i.e. the outcome of I_j is decided in step j. From (18), we can see that it is possible for the weights to be negative. However, to achieve spatial balance, the weights need to be positive. When updating the inclusion probability for unit i, the weight it receives depends on the distance between i and the step unit j. If the inclusion indicator for the step unit is 1, then based on the inclusion probability in step $j-1$, the unit j needs to "steal" more probability from its nearby units. The closer the unit to the step unit, the more probability mass will be "stolen" by the step unit until the updated inclusion probability of the step unit becomes 1, i.e. the nearest unit to j will receive as much weight as possible from j, then as much as possible weight will received by the second nearest unit from j, etc. By contrast, if the step unit has an inclusion indicator equal to 0, it will give away all its probability mass to its neighbours in a similar way. The above strategy is called maximal weight strategy. Thus, with the restriction of the maximum weight of each unit and the sum of the weights equal to 1, we will have a sample of a fixed size.

5.2 Double sampling approach

Double sampling is also called two-phase sampling. It is a sampling method exploiting the correlation between auxiliary variables and the target variables. As the name implies, by applying two-phase sampling, the sample will be selected in two steps. In the first phase, a relatively large first phase sample is taken, in which only the auxiliary variables are derived. In the second phase, a subsample is taken from the first phase sample with the help of the auxiliary variable(s) derived in the first phase, and the target variable is measured only on this subsample. It can be applied to convert a continuous population to a finite population and apply the available auxiliary information to select well-spread samples, see papers I and II.

5.3 Sample coordination

Sample coordination is a method to statistically control the overlap of successive samples. Various sampling strategies have been developed based on sample coordination (e.g. Patterson, 1950; Keyfitz, 1951; Kish & Scott, 1971; Ernst, 1999). There are two kinds of coordination: positive coordination and

negative coordination. They intend to maximize or minimize the overlap of two or more samples, respectively. The focus is only on positive coordination in order to produce good estimates of change for this thesis. The positive coordination method is based on permanent random numbers (PRNs). These PRNs are used to decide the the inclusion or exclusion of a unit at each time point. We assign a uniformly distributed random number on the interval [0,1] to each unit in the frame (as in the algorithm of SCPS), and these numbers remain with the units over time. Such a number is called a PRN.

The positive sample coordination method was applied together with SCPS in Papers II and III to select partially overlapping and spatially balanced samples when monitoring change over time. By doing so, we update the sample at a second time occasion by using the updated auxiliary information and meanwhile keep a certain degree of overlap with the sample selected on the previous time occasion.

5.4 Combining environmental surveys

Sometimes an existing survey is not sufficient to produce good estimators. An example is when applying a national environmental survey to obtain estimators in the regional level. In such a case, one or several complementary surveys are often needed to improve the estimation to reach specific accuracy (Christensen & Ringvall, 2013). We should not discard the national survey when planning an extra sampling effort. It will be beneficial to use all available information to produce the best possible estimators.

When combining different surveys, an important consideration is the variance estimation. It is well known that, the linear combination of independent estimators $\widehat{Y}_1, \widehat{Y}_2, ..., \widehat{Y}_k$ with the smallest variance is

$$\widehat{Y}_l = \alpha_1 \widehat{Y}_1 + \alpha_2 \widehat{Y}_2 + ... + \alpha_k \widehat{Y}_k, \tag{19}$$

where $\alpha_i = V^{-1}(\widehat{Y}_i) / \sum_{j=1}^{k} V^{-1}(\widehat{Y}_j)$ and it is common that variance estimates are used when calculating the α_i. Since the separate estimators of the population total can be highly correlated to their respective variance estimators, the linear combination estimator based on the estimated variances can produce large bias. Alternative methods are needed to overcome the problem of linear combination with estimated variances.

6 Swedish national forest inventory

The Swedish NFI started in 1923 and it is a sample-based multipurpose inventory. It provides regional and national level statistics, and in addition also produces national full-coverage forest maps by combining field and remote sensing data (Fridman *et al*., 2014). It is an annual sample inventory of the country's forests carried out by the Department of Forest Resource Management at SLU. The main purpose of the NFI is to describe the state and change of the Swedish forest. The inventory covers all types of land, but it is on productive forest land as the most comprehensive description is made.

In the Swedish NFI, Sweden was divided into five strata with decreasing sampling intensity towards the north. Within each stratum, clusters of circular plots are sampled. The clusters are quadratic or rectangular in shape, with a side length varying from 300 to 1800 m between different parts of the country. Along the sides of the clusters, circular plots were located with fixed distance between plots. The within-stratum distance between plots is fixed and it increases with latitude. The design was motivated by the assumed autocorrelation for relevant forest variables such as stem volume. In other words, the landscape changes more rapidly in the south compared to the north. In the south we have an overall higher anthropological influence coupled with higher variation in species and fertility, while the coniferous forests in the north are more homogenous. Thus, longer distances between plots was needed in the north to obtain new information. Two kinds of clusters are used: temporary ones and permanent ones. The temporary clusters are mainly intended to capture the current state of the forest and are only surveyed once, whereas permanent clusters primarily aim to capture changes and are resurveyed regularly (Tomppo *et al*., 2010). The sample selection in different strata are independent, and the estimation of target variables is required at the stratum level. A sample of the survey clusters distributed over the whole country is measured annually. A five-year inventory cycle is used, using five consecutive yearly inventories, and the estimates are calculated as a five-year moving average. Separate estimators are used for each year and each cluster type, and a weighting is used to calculate the averages of both cluster types. Details about the estimators used in the Swedish NFI can be found in e.g. Fridman *et al*. (2014).

7 Summary of the papers

Four papers are included in this thesis with a focus on design-based sampling methods for environmental monitoring. A short summary of each paper is included in this section.

7.1 Paper I: A new sampling strategy for forest inventories applied to the temporary clusters of the Swedish national forest inventory

In this paper, a new sampling strategy for forest inventories is presented. The strategy is developed for an area frame and the sample is selected through the double sampling approach. In the first phase, a very large number N of clusters is selected by randomly and independently placing cluster centers in the sampling frame F with the sampling intensity $N/\ell(F)$. For each cluster, the auxiliary information of the cluster mean is derived. According to the Glivenko-Cantelli theorem and its multivariate generalisations, the empirical distribution of the auxiliary variables in the first phase sample converges uniformly almost surely to the population distribution as the size of the sample increases. In the second phase, the LPM is applied with equal inclusion probabilities n/N. Then we achieve a representative and well spread sample with a sampling intensity $n/\ell(F)$.

The potential of implementing the new strategy for temporary clusters within the Swedish national forest inventory is evaluated with five auxiliary variables derived from remote sensing data: geographical coordinates, elevation, predicted tree height, and predicted basal area. By comparing with two reference strategies (independent observations and geographically well-spread observations), through a Monte-Carlo simulation, we show that the new strategy succeeds in producing precise or improved estimates. For this reason, we conclude that the new strategy has a great potential to achieve large improvements in estimation of many important forest attributes. The encouraging results of this study have led to a decision to implement this sampling strategy in all regions for the selection of temporary tracts within the Swedish NFI since 2018.

7.2 Paper II: A sample coordination method to monitor totals of environmental variables

We propose a design-based sampling strategy for environmental monitoring. Two concepts are combined in the strategy, spatially balanced sampling and

coordination of samples over time. By employing this strategy, the selected samples are partially overlapping and spatially balanced.

We show how sample coordination can be applied within a continuous population framework by using a double sampling approach. Where the sample selected at the first phase is treated as a permanent but dynamic population. The second phase sample is selected by SCPS. The algorithm for the new strategy, where the SCPS and sample coordination based on PRNs are also illustrated by examples.

The strategy was evaluated by the same NFI example as Paper I (with added another time occasion). Compared with four reference strategies $\big($(i) a strategy that uses permanent geographical-spread samples; (ii) a strategy that employs permanent well-spread samples; (iii) a strategy with independent well-spread samples; (iv) a strategy that applies a split-panel design$\big)$, the superiority of the new strategy is illustrated by a forest inventory application. By using Monte-Carlo simulations, we show that the new strategy can outperform the reference strategies for both state and change estimators of the auxiliary variables. This implies that we can potentially produce good estimates for the target variables that are related to the auxiliary variables.

7.3 Paper III: Estimation of change with partially overlapping and spatially balanced samples

This paper is a follow-up to Paper II. The advantage of a strategy which employs partially overlapping and spatially balanced samples is verified in Paper II. However, the problem of estimating the variance of the estimator of change under the strategy was left open. In this paper, we try to solve the problem by proposing a local mean estimator for partially overlapping and spatially balanced samples for the variance of an estimator of the change. The estimator is derived by modifying an estimator for partially overlapping samples with independent observations. Through simulations, we illustrate that for partially overlapping and spatially balanced samples the proposed local mean estimator is a viable option, compared to a reference estimator.

7.4 Paper IV: Combining Environmental Area Frame Surveys of a Finite Population

In this paper, new ways to combine data from multiple environmental surveys of a finite population are introduced. We derive two methods to reduce the bias of the combined estimator. The first approach is to derive design components for the combined design. The second approach uses a pooled variance

estimator to estimate the variance of each separate estimator by using information from all surveys. We derive the design components needed at unit level of the finite population to combine multiple surveys for the sampling designs in Section 4. 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, then there is a risk of introducing negative bias. By using pooled variance estimators, the bias of a linear combination estimator can be reduced. Through simulation, we show that the proposed methods are either unbiased or yield small bias, compared to traditionally used methods. Our results can be used to combine data from different surveys with improved accuracy and efficiency. If an existing survey did not provide sufficiently good estimators, then the results can also be used when planning a complementary survey that can be combined with the existing survey.

8 Conclusions and remarks

In this thesis, the main focus is the design stage of the sampling strategies. This is why the HT/HH-estimator is always applied for the estimations. For multi-purpose surveys, such as environmental monitoring, it is difficult to find an optimal sampling strategy. The optimization of the sample for one variable may decrease the precision of other variables. In such a situation, the model-assisted approach may be more efficient than the pure design-based approach if the models used in the estimation are adequate. Because it can borrow strength from the auxiliary variables twice, once at the design stage and once in the estimation. But, if the information has been used to spread the sample, then the improvement will be small by using the same information again in the estimation. However, if new information has become available after the sample has been selected, then by using the new auxiliary information in the estimation we may get a larger improvement.

For spatially balanced sampling designs, the distance measure we used to verify if units are nearby or not is important. It is a way of describing what it means for elements in some space to be “close to", or “far away from" each other. We applied the standardized Euclidean distance in the LPM and SCPS, i.e., all auxiliary variables are equally important when spreading the sample. The reason of standardization is to balance out the contributions of different variables. This is essential for multipurpose environmental surveys. Without standardization, it may happen that a single variable dominates in the calculation of the distances.

When domain estimates are required in large national environmental surveys, there is a need for adequate resources of the domains to get estimates with acceptable precision. There is sometimes a demand for additional samples of such domains to complement the sparse national level samples. By combining different samples as we did in Paper IV, the restriction is that the samples are independent probability samples selected from the same population. It basically means that the samples should be observed at the same time point. However, it is seldom the case in practise. As a result, when applying the methods proposed, we need to assume that the population does not change within the time that the samples are observed.

The proposed sampling methods have a great potential to be applied in large-scale environmental monitoring programs. In fact, some of the methods developed in this thesis have already been applied within the Swedish NFI and within the National Inventory of Landscapes in Sweden. When using the spatially balanced samples, we can increase the efficiency of current state estimation. When using partially overlapping and spatially balanced samples,

the precision for both state and change estimators can be improved. We no longer need to compromise between having good estimates of the state and good estimates of change. The methods proposed in Paper IV can be applied, if we want to produce more efficient estimators when combining a large-scale environmental survey with one or several complementary surveys.

References

Berger, Y. G. (2004). A simple variance estimator for unequal probability sampling without replacement. *Journal of Applied Statistics*, 31(3), 305-315.

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

Brewer, K.R.W., Early, L.J. & Joyce, S.F. (1972). Selecting several samples from a single population. *Australian Journal of Statistics*, 14(3), 231-239. doi:10.1111/j.1467-842X.1972.tb00899.x.

Christensen, P., & Ringvall, A. H. (2013). Using statistical power analysis as a tool when designing a monitoring program: experience from a large-scale Swedish landscape monitoring program. *Environmental monitoring and assessment*, 185(9), 7279-7293.

Cordy, C.B. (1993). An extension of the Horvitz—Thompson theorem to point sampling from a continuous universe. *Statistics & Probability Letters*, 18(5), 353-362.

de Gruijter, J.J., Brus, D.J., Bierkens, M.F.P. & Knotters, M. (2006). *Sampling for natural resource monitoring*. Springer-Verlag, Berlin.

Ernst, L.R. (1999). The maximization and minimization of sample overlap: A half century of results. *Bulletin of the International Statistical Institute*, 57, 293-296.

Fridman, J., Holm, S., Nilsson, M., Nilsson, P., Ringvall, A. H. & Ståhl, G. (2014). Adapting national forest inventories to changing requirements–the case of the Swedish national forest inventory at the turn of the 20th century. *Silva Fennica*, 48(3), 1-29.

Grafström, A. (2012). Spatially correlated Poisson sampling. *Journal of Statistical Planning and Inference*, 142(1), 139-147. https://doi.org/10.1016/j.jspi.2011.07.003.

Grafström, A. Lundström, N.L.P. & Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, 68(2), 514-520. doi:10.1111/j.1541-0420.2011.01699.x.

Grafström, A. & Lundström, N.L.P. (2013). Why well spread probability samples are balanced. *Open Journal of Statistics*, 3(1), 36-41.

Grafström, A. & Schelin, L. (2014). How to select representative samples. *Scandinavian Journal of Statistics*, 41(2), 277-290. https://doi.org/10.1111/sjos.12016.

Grafström, A., Zhao, X., Nylander, M. & Petersson, H. (2017b). A new sampling strategy for forest inventories applied to the temporary clusters of the Swedish NFI. *Canadian Journal of Forest Research*, 47(9), 1161-1167.

https://doi.org/10.1139/cjfr-2017-0095.
Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35(4), 1491–1523. https://doi.org/10.1214/aoms/1177700375
Hansen, M.H. & Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14(4), 333-362.
Hartley, H.O. & Rao, J.N.K. (1962). Sampling with unequal probabilities and Without replacement. *The Annals of Mathematical Statistics*, 33(2), 350 –374.
Horvitz, D.G. & Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663–685.
Keyfitz, N. (1951). Sampling with probabilities proportional to size: adjustment for changes in the probabilities. *Journal of the American Statistical Association*, 46(253), 105–109.
Kish, L. & Scott, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66(335), 461–470.
Mandallaz, D. (2007). *Sampling techniques for forest inventories*. CRC Press.
Patterson, H. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Series B (Methodological)*, 12(2), 241-255.
Qualité, L. (2009). *Unequal probability sampling and repeated surveys* (Doctoral dissertation, Université de Neuchâtel).
Royall, R.M. (1970). On finite population sampling theory under certain linear regression models.*Biometrika*, 57(2), 377-387.
Stevens, D.L. & Olsen, A.R. (2003). Variance estimation for spatially balanced samples of environmental resources. *Environmetrics*, 14(6), 593-610.
Stevens, D.L. & Olsen, A.R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, 99(465), 262-278. doi:10.1198/016214504000000250.
Särndal, C.-E., Swensson, B. & Wretman, J. (1992). *Model assisted survey sampling*. Springer Verlag, New York.
Tam, S.M. (1984). On covariance in finite population sampling. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 34(4), 429–433.
Tomppo, E., Gschwantner, T., Lawrence, M. & McRoberts, R.E. (2010). *National forest inventories. Pathways for common reporting. European Science Foundation*, Dordrecht: Springer Netherlands.
Yates, F. & Grundy, P.M. (1953.) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical*

Society, Series B (Methodological), 15(2), 253–261.
Zhao, X., & Grafström, A. (2020). A sample coordination method to monitor totals of environmental variables. *Environmetrics*, 31(6). https://doi.org/10.1002/env.2625.

Acknowledgements

Many people were included in the process of completing this thesis. I would like to extend my sincere thanks to all of them.

First and foremost, I express my heartfelt gratitude to my main supervisor, Anton Grafström, for sharing knowledge and ideas, for providing guidance and feedback throughout these years. Without consistent support and overall insight from you, this thesis would not exist. I appreciate all things you have done and I could not have imagined having a better supervisor. Also, my deepest thanks to my co-supervisors: Anna-Lena Axelsson, Cornelia Roberge and Hans Petersson for your encouragement and help. Thank you Hans, also for the time you spent in the administration for my studies.

I would like to offer my special thanks to Johan Fransson. I appreciate all your support and encouragement during my hard time. Thank you Annica de Groote, Magnus Ekström and Wilmer Prentius, for your questions and suggestions during and after the final seminar. Thank you Wilmer, also for the nice company and collaboration. Jonas Fridman, thank you for providing me with the NFI related information.

Inka and Peder, thanks for the nice company during these years, it has been a great help during my bad days. Alex, Anders M., Ann-Helen, Anne-Maj, Arne, Emma S., Gun, Jaime, Jonas B., Jörgen, Kenneth N., Langning, Mats N., Torgny, Ylva M., Ylva J., it is your kind support that has made my studies at the department a wonderful time. To other people from the Department of Forest Resource Management, whose names are not mentioned but have helped me during this long journey, thank you!

My dear friend Yuli, can you imagine that eighteen years have passed since the first time we met? I look forward to our adventures for another eighteen years! Chun, Jia, Wen, Ye, and other members in the “Tomtebo alliance", thank you for the good food and those wonderful family activities we have had together. Elisabeth Svensson, as my master thesis supervisor and as a friend, you have motivated me a lot.

To my parents in law, thanks for the unlimited support you gave me. To my parents, thank you for raising me up and for letting me be myself. To Siyi and Siqi, my lovely sons and troublemakers, without you I could have had this thesis finished much earlier, but it would never be as good as it is now! Finally, I would like to acknowledge with gratitude, the support and love of my husband, Xijia. Although we always have different views on things, we take the attributes of each other, and in the end, we will become the same and a better person!

Umeå, July 2021

I

NRC Research Press

ARTICLE

A new sampling strategy for forest inventories applied to the temporary clusters of the Swedish national forest inventory

Anton Grafström, Xin Zhao, Martin Nylander, and Hans Petersson

Abstract: A new sampling strategy for forest inventories is presented. The most important difference from the traditional sampling strategies is that auxiliary variables from remote sensing are incorporated into the sampling design. The sample is selected to match population distributions of the auxiliary variables as well as possible. This is achieved by a double sampling approach, where auxiliary variables are extracted for a large first-phase sample. The second selection is done by the local pivotal method and produces an even thinning of the first-phase sample. Thus, we make sure that the selected second-phase sample becomes much more representative of the population than what is possible by the use of traditional designs. The potential of implementing the new strategy for the temporary clusters within the Swedish national forest inventory is evaluated with five auxiliary variables: the geographical coordinates, elevation, predicted tree height, and predicted basal area. The increased representativity that we achieve with the new strategy induces up to 95% reduction of the variance of the sample means of the remote sensing auxiliary variables compared with traditional designs. For this reason, we conclude that the new strategy that will be implemented in the forthcoming Swedish national forest inventory has a great potential to achieve large improvements in estimation of many important forest attributes.

Key words: continuous population, double sampling, local pivotal method, remote sensing, sampling design.

Résumé : Nous présentons une nouvelle stratégie d'échantillonnage pour les inventaires forestiers. La différence la plus importante par rapport aux stratégies d'échantillonnage traditionnelles est l'incorporation dans le plan d'échantillonnage de variables auxiliaires de télédétection. L'échantillon est sélectionné de manière à correspondre autant que possible à la distribution de la population des variables auxiliaires. Cela est accompli grâce à une méthode de double échantillonnage, où les variables auxiliaires sont extraites pour un grand échantillon lors de la première phase. La deuxième sélection est effectuée avec la méthode du pivot local et produit une réduction uniforme de l'échantillon de la première phase. Ainsi, nous nous assurons que l'échantillon sélectionné lors de la deuxième phase devient beaucoup plus représentatif de la population que le permet l'utilisation des modèles traditionnels. Le potentiel de mise en œuvre de la nouvelle stratégie pour les grappes temporaires de l'inventaire forestier national suédois est évalué à l'aide de cinq variables auxiliaires : les coordonnées géographiques, l'altitude, la hauteur prédite des arbres et la surface terrière prédite. La représentativité accrue, que nous obtenons avec la nouvelle stratégie, entraîne jusqu'à 95 % de réduction de la variance des moyennes d'échantillonnage des variables auxiliaires de télédétection par rapport aux modèles traditionnels. Pour cette raison, nous concluons que la nouvelle stratégie, qui sera mise en œuvre dans le prochain inventaire forestier national suédois suédois, a de fortes chances d'améliorer grandement l'estimation de nombreux attributs forestiers importants. [Traduit par la Rédaction]

Mots-clés : population continue, double échantillonnage, méthode du pivot local, télédétection, plan d'échantillonnage.

Introduction

National forest inventories (NFIs) have evolved and developed, in some cases more than 100 years, and the need for accurate national-level information is more requested than ever (Tomppo et al. 2010, chap. 1). Still the NFI designs normally rest on traditional area-based sampling, which spreads the sample units over the landscape. Often the sample units are systematically distributed and organised in clusters of circular plots. NFIs in general have a very low sampling intensity due to the large areas that need to be covered. In such a situation, it is inevitable that forest attributes vary rapidly across the landscape with respect to the low sampling intensity. This means that spreading the sample only geographically is not sufficient to ensure that the sample is representative of the population. With the intention of providing a more effective sampling design and thereby increasing the precision of estimates of forest attributes, we present a strategy for obtaining a more representative sample by using auxiliary information from remote sensing in the planning phase of a forest inventory. In recent years, for example, assessments using LiDAR techniques (light detection and ranging) can provide quite up to date wall-to-wall coverage of remote sensing data. In some countries, such data are available even at the national scale and may be used for distributing sample units efficiently for NFIs.

Even though NFIs have been well developed overtime, it is still imperative for NFIs to adopt new strategies to be cost-efficient and increase the precision of estimates (Fridman et al. 2014). Despite the fact that auxiliary variables from remote sensing are becoming increasingly available, they are rarely used in the sampling designs. In the Swedish NFI, for example, clusters have been distributed more or less evenly across the landscape without the use of additional auxiliary variables.

Received 6 March 2017. Accepted 12 May 2017.

A. Grafström, X. Zhao, M. Nylander, and H. Petersson. Department of Forest Resource Management, Swedish University of Agricultural Sciences SLU, Skogsmarksgränd, SE-901 83 Umeå, Sweden.

Corresponding author: Anton Grafström (email: anton.grafstrom@slu.se).

Auxiliary variables can be used in different ways in a sampling design. Common use includes stratification (e.g., Särndal et al. 2003, chaps. 3 and 12), balancing (e.g., Deville and Tillé 2004), and using unequal probabilities or achieving a good spread of the sample (e.g., Stevens and Olsen 2004). Including the auxiliary variables in the design normally reduces the need for including the same variables in the estimators and can allow for a simpler analysis. A sampling design that uses auxiliary variables to spread the sample is particularly useful for multipurpose inventories, such as NFIs (Grafström and Schelin 2014). When a multipurpose inventory is planned, the choice of a robust design is especially important. Tillé and Wilhelm (2017) discussed principles for choice of sampling design and stated that "Indeed, if the response variable is correlated with the auxiliary variable, then spreading the sample on the space of auxiliary variables also spreads the sampled response variable. It also induces an effect of smooth stratification on any convex set of the space of variables. The sample is thus stratified for any domain, which can be interpreted as a property of robustness." As demonstrated by for example, Grafström and Ringvall (2013), use of auxiliary variables in an estimator can only partly compensate for neglecting the use of the same variables in the design.

Grafström and Ringvall (2013) and Grafström et al. (2014) have recently introduced different sampling designs for forest inventories that are able to select spatially balanced samples, which means that the samples are well spread in some space. We have now developed this theoretical framework further to meet the specific needs of forest inventories. Our framework includes using the continuous population approach, which was first proposed for forest inventories by Mandallaz (1991); see also Eriksson (1995), Barabesi (2003, 2004), Mandallaz (2007, chap. 4), and Gregoire and Valentine (2008, chap. 10). Following Cordy (1993), we can in this framework use a general sampling design for selection of clusters of any shape and with any prescribed sampling intensity function. However, we focus on the selection of representative samples, which means that we match as closely as possible the sample distribution of a set of auxiliary variables to the population distribution. This is achieved through a double (or two-phase) sampling, where auxiliary responses are extracted for a very large first-phase sample of clusters. For the second-phase sample selection, we use the local pivotal method (LPM) by Grafström et al. (2012) to spread the sample. When using a constant sampling intensity, the LPM produces representative samples (Grafström and Schelin 2014). Different implementations of the LPM can be found in the R package 'BalancedSampling' (Grafström and Lisic 2016).

The new strategy is illustrated with an application, where we select the temporary clusters for the Swedish NFI. As auxiliary variables, we use a digital elevation model and a recent nationwide forest attribute map of Sweden predicted using airborne laser scanning data and field data from the NFI (Nilsson et al. 2017). When compared with two reference strategies (independent observations and geographically well-spread observations), through a Monte-Carlo simulation, it is evident that the new strategy succeeds in producing representative samples.

The new sampling strategy

For the new sampling strategy, a continuous population approach with double sampling is employed. In the first-phase sample, a very large number N of clusters is selected by randomly and independently placing cluster centers in the region. For each cluster, the auxiliary information of the cluster mean is derived. According to the Glivenko–Cantelli theorem and its multivariate generalisations, the empirical distribution of the auxiliary variables in the first-phase sample converges uniformly almost surely to the population distribution as the size of the sample increases (Wolfowitz 1954; Dehardt 1971). Then, a smaller sample of size n is selected from the N clusters by the LPM in such a way that the distribution of the auxiliary variables in the second-phase sample matches the distribution in the large first-phase sample very closely. Thus, by using a very large first-phase sample, we make sure that the distribution of the auxiliary variables in the second-phase sample is very close to the corresponding distribution in the population, which means that we obtain a sample that is representative of the auxiliary variable space. In this section, the new strategy as well as an example to illustrate the superiority of the new strategy to the reference strategies are presented. The general framework and the notation of a sampling strategy for continuous populations are provided. The subsequent subsections show the framework and the notation of using auxiliary information in a double sampling approach, introduce the definitions of spatial balance, focus on the LPM that we emply for the second-phase sample selection, and finally, provide an illustrative example of the proposed strategy.

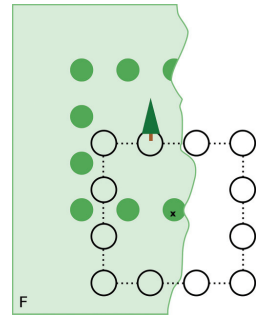


Fig. 1. An example of an inclusion zone. The inclusion zone K for the tree consists of the darker circles intersected by the surface of the forest; the circles connected with dots represent a cluster. Any cluster $\mathcal{C}(\mathbb{X})$ with its center $\mathbb{X}$ within K, such as the one in the figure, includes the tree in one of the plots. [Colour online.]

A sampling strategy for continuous populations

Consider a surface F that is assumed to be a subset of the Euclidean plane $\mathbb{R}^2$ with its surface area $\ell(F)$. For a finite population consisting of N_T objects (e.g., trees) located in F, the N_T objects are represented by points. Let $U = \{1, \ldots, i, \ldots, N_T\}$ be the identifiers for the N_T objects, and let $S_T \subset U$ denote the probability sample of identifiers for the selected objects. The inclusion probability of object i to be sampled is defined as $\pi_i = \Pr(i \in S_T)$. The variable of interest, which is generally nonnegative and bounded, is denoted by y_i. An important objective of a forest inventory is the estimation of the population total $Y = \sum_{i \in U} y_i$. For forest inventories, since the sampling frame is indeterminable for the units in U, the objects cannot be sampled directly. Instead, we select our sample from a continuous population on F as described in, e.g., Mandallaz (2007).

A sampling design on F is defined by a joint distribution of n random variables. Denote the random sample of n locations within F as $S_F = \{\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_n\}$. The (prescribed) sampling intensity is $\pi(\mathbb{X}) = \sum_{i=1}^{n} f_i(\mathbb{X})$, where $f_i(\mathbb{X})$ is the marginal probability density function of $\mathbb{X}_i$ and moreover, $\pi(\mathbb{X}) > 0$ for $\mathbb{X} \in F$ and $\pi(\cdot) = 0$ outside F. The sampling intensity plays the same role as the inclusion probabilities play in finite population sampling. We have $n = \int_F \pi(\mathbb{X}) \mathrm{d}\mathbb{X}$ for a design of a fixed size n.

When using clusters with a given configuration and a fixed orientation, the inclusion zone $K_i \subset F$ for a tree i on location $\mathbb{X}_i$ can be expressed as $K_i = K(\mathbb{X}_i) = \{\mathbb{X} \in F : \mathbb{X}_i \in \mathcal{C}(\mathbb{X})\}$, where $\mathcal{C}(\mathbb{X})$ is a cluster centered on $\mathbb{X}$. Figure 1 shows an example of the inclusion zone of a tree close to the forest boundary.

There exist several ways to formulate the density function $Y(\mathbb{X})$ of the target variable. For this article, we define the density function as a weighted sum of y_is over the objects that are selected:

$$(1) \qquad Y(\mathbb{X}) = \sum_{i \in U} \frac{I_i(\mathbb{X}) y_i}{\ell(K_i)}$$

where the weight is the inverse of the area of the inclusion zone of the tree, $I_i(\mathbb{X}) = 1$ if $\mathbb{X} \in K_i$ and 0 otherwise. The density function 1 has been used by, e.g., Mandallaz (2007). The density function is constructed in such a way that $Y = \int_F Y(\mathbb{X}) \mathrm{d}\mathbb{X}$ is identical to the corresponding finite population total $Y = \sum_{i \in U} y_i$, which follows from

$$(2) \qquad Y = \int_F Y(\mathbb{X}) \mathrm{d}\mathbb{X} = \int_F \sum_{i \in U} \frac{I_i(\mathbb{X}) y_i}{\ell(K_i)} \mathrm{d}\mathbb{X} = \sum_{i \in U} \frac{y_i}{\ell(K_i)} \int_F I_i(\mathbb{X}) \mathrm{d}\mathbb{X} = \sum_{i \in U} y_i$$

Cordy (1993) proposed a continuous version of the Horvitz—Thompson estimator of the population total Y as well as the variance of the estimator in Sen–Yates–Grundy form. They are given by

$$\hat{Y} = \sum_{\mathbb{X} \in S_F} \frac{Y(\mathbb{X})}{\pi(\mathbb{X})}$$

$$V_{\mathrm{SYG}}(\hat{Y}) = \frac{1}{2} \int_F \int_F [\pi(\mathbb{X})\pi(\mathbb{X}') - \pi(\mathbb{X}, \mathbb{X}')] \times \left[\frac{Y(\mathbb{X})}{\pi(\mathbb{X})} - \frac{Y(\mathbb{X}')}{\pi(\mathbb{X}')} \right]^2 \mathrm{d}\mathbb{X} \mathrm{d}\mathbb{X}'$$

where $\pi(\mathbb{X}, \mathbb{X}')$ is the second-order sampling intensity for a pair of points $(\mathbb{X}, \mathbb{X}')$.

Double sampling approach to achieve spatial balance and select representative samples

If $Y(\mathbb{X})$ is well explained by the auxiliary variables, then it is efficient to select a sample whose empirical distribution of the auxiliary variables matches the population distribution of the auxiliary variables. By well explained, we mean that points with a small distance in auxiliary space in general have more similar values on the target variable than points farther apart.

Normally, auxiliary information from remote sensing is available at a grid-cell level with different resolutions. To utilize such auxiliary information for the selection of spatially balanced samples, we need to implement double sampling.

To obtain the prescribed sampling intensity $\pi(\mathbb{X}) = n/\ell(F)$ and a spatially balanced second-phase sample of size n, we first select a large sample S_{F_1} of size N with independent observations over F, where $N \gg n$, with the sampling intensity $\pi_1(\mathbb{X}) = N/\ell(F)$. Then we extract the auxiliary variables for each cluster. For the second selection, we propose the use of the LPM with equal probabilities n/N. Then we achieve a representative and well-spread second-phase sample with the prescribed sampling intensity $\pi(\mathbb{X})$.

Suppose we have p auxiliary variables available from any source that provides wall-to-wall data. They are defined as $\mathbb{Z}'(\mathbb{X}) = [Z'_1(\mathbb{X}), \ldots, Z'_p(\mathbb{X})]^T \in \mathbb{R}^p$. Let $\mathbb{Z}'(\mathbb{X})$ be the single point response for the auxiliary variables (i.e., the value for the grid cell that contains the point). Thus, all single point responses within one grid cell have the same value for the auxiliary variable. To preserve the relationship between the auxiliary and the target variables, it is ideal to derive the auxiliary response in a similar way as $Y(\mathbb{X})$.

The point response of the cluster $\mathcal{C}(\mathbb{X})$ is here defined as

$$(3) \qquad \mathbb{Z}^*(\mathbb{X}) = \int_{\mathbb{X}' \in F} \frac{I[\mathbb{X}' \in \mathcal{C}(\mathbb{X})] \mathbb{Z}'(\mathbb{X}')}{\ell[K(\mathbb{X}')]} \mathrm{d}\mathbb{X}'$$

Then, in a similar way as for the target variable (e.g., see eq. 2), we obtain

$$(4) \qquad \int_{\mathbb{X} \in F} \mathbb{Z}^*(\mathbb{X}) \mathrm{d}\mathbb{X} = \int_{\mathbb{X} \in F} \int_{\mathbb{X}' \in F} \frac{I[\mathbb{X}' \in \mathcal{C}(\mathbb{X})] \mathbb{Z}'(\mathbb{X}')}{\ell[K(\mathbb{X}')]} \mathrm{d}\mathbb{X}' \mathrm{d}\mathbb{X}$$

$$= \int_{\mathbb{X}' \in F} \frac{\mathbb{Z}'(\mathbb{X}')}{\ell[K(\mathbb{X}')]} \int_{\mathbb{X} \in F} I[\mathbb{X}' \in \mathcal{C}(\mathbb{X})] \mathrm{d}\mathbb{X} \mathrm{d}\mathbb{X}' = \int_{\mathbb{X}' \in F} \mathbb{Z}'(\mathbb{X}') \mathrm{d}\mathbb{X}'$$

Equation 4 means that the total of the cluster response equals the total of the single point response.

Measuring the spatial balance for continuous populations

When the auxiliary space is multidimensional, spatial balance can be used as a measure to check if the empirical distribution of a sample fits the sampling distribution. Stevens and Olsen (2004) proposed to use a statistic based on Voronoi polytopes to describe the spatial balance. The polytope p_i for a point $\mathbb{X}_i$ in the sample includes all points in the population closer to $\mathbb{X}_i$ than to any other sample point $\mathbb{X}_j$, $j \neq i$. If a sample is well spread, there should be an approximately equal amount of probability mass in each polytope. This implies that if a constant intensity is applied, then all polytopes should optimally be of equal size. The spatial balance of a sample from a continuous population can be expressed as

$$B = \frac{1}{n} \sum_{i \in s} (v_i - 1)^2$$

where $v_i = \int_{p_i} \pi(\mathbb{X}) \mathrm{d}\mathbb{X}$ is the total probability mass within the polytope p_i. Additionally, all the v_is should be close to 1 for a spatially balanced sample. Hence, B is a measure of the variance of the total probability mass within the polytopes. Obviously, the smaller the value of B is, the better the sample fits the sampling distribution. A simulation to find the expected value of B under a design reveals how well the design succeeds in producing spatially balanced samples.

Local pivotal method

The LPM has been shown to be one of the most effective methods in regards to spreading the sample in auxiliary space (e.g., Benedetti et al. 2015, chap. 7). By employing the LPM, we can select samples whose empirical distribution matches the population distribution of the auxiliary variables. Such samples are spatially balanced in the auxiliary space, leading to an approximate balance for any target $Y(\mathbb{X})$ well explained by those auxiliary variables (Grafström and Lundström 2013). Thus, for such targets, we achieve $\hat{Y} \approx Y$. When applying the LPM, spatial balance is achieved by successively updating the inclusion probabilities for nearby units until they become inclusion indicators, i.e., 0′s and 1′s, where the 0′s indicate exclusions of the units and the 1′s indicate inclusions of the units.

In one step of the LPM, we randomly select one unit i and find its nearest neighbour j. The pair of nearby units will compete with the (possibly updated) inclusion probabilities $0 < \pi_i < 1$ and $0 < \pi_j < 1$. The winner takes as much inclusion probability as possible from the loser. Thereafter, the winner has an updated inclusion probability $\pi_W = \min(1, \pi_i + \pi_j)$, while the loser has the new inclusion probability $\pi_L = \pi_i + \pi_j - \pi_W$. Thus, if $\pi_i + \pi_j \geq 1$, then $\pi_W = 1$ and the winner is included in the sample. If $\pi_i + \pi_j < 1$, then $\pi_L = 0$ and the loser is excluded from the sample. A final decision is made for at least one unit each step. The procedure for the competition is given by

$$\left(\pi_i', \pi_j'\right) = \begin{cases} (\pi_W, \pi_L) & \text{with probability} \dfrac{\pi_W - \pi_j}{\pi_W - \pi_L} \\ (\pi_L, \pi_W) & \text{with probability} \dfrac{\pi_W - \pi_i}{\pi_W - \pi_L} \end{cases}$$

where (π_i', π_j') denote the new and updated probabilities for the pair. When nearby units compete for inclusion, they are unlikely to be included simultaneously, which forces the sample becoming well spread. Figure 2 shows an example of the competition procedure for one step in a two-dimensional space.

Example for a one-dimensional auxiliary space

To illustrate the proposed strategy, we provide an example for a one-dimensional auxiliary variable space. Let the auxiliary variable distribution be $Z \sim N(0,1)$. We perform a simulation of 1000 random samples of size $n = 350$ with independent observations and compare with 1000 first-phase samples of size $N = 100\,000$ with independent observations followed by a selection of second-phase samples of size $n = 350$ using the LPM with probabilities $\pi_i = n/N$, $i = 1, 2, \ldots, N$.

The results of the comparisons are presented in Fig. 3 for variation of sample mean, spatial balance, and maximum distance. The maximum distance is the maximum distance between the empirical distribution function and the reference distribution, which was calculated by employing the one-sample Kolmogorov–Smirnov test.

For the LPM with a second-phase sample of size 350, the variance of the sample mean corresponded approximately to the variance of the sample mean of 35 000 independent observations. Thus, for the mean of the auxiliary variables, such balanced samples of size 350 are as good samples of size 35 000 with independent observations. The mean of the spatial balance of the LPM was 0.065 and the mean of the maximum distance was 0.007 compared with 0.499 and 0.046 for independent random sampling (IRS), respectively.

As we can see from Fig. 3, the sampling method that has a lower value of spatial balance also has a lower value of maximum distance. In fact, for the 1000 selected samples, even the "worst" samples resulting from the LPM fit the sampling distribution much better than the "best" samples selected by IRS. When the auxiliary variable space is multidimensional, we can use the spatial balance to measure how well a sample represents the sampling distribution (and hence the population in the case of a constant sampling intensity).

An approximate variance estimator of the LPM was derived by Grafström and Schelin (2014). The continuous version of the estimator can be expressed as

$$\hat{V}_{\text{LPM}}(\hat{Y}) = \frac{1}{2} \sum_{\mathbb{X} \in S_F} \left[\frac{Y(\mathbb{X})}{\pi(\mathbb{X})} - \frac{Y(\mathbb{X}')}{\pi(\mathbb{X}')} \right]^2$$

In the auxiliary space, $\mathbb{X}'$ is the nearest neighbour to $\mathbb{X}$ in the random sample with n locations S_F. The nearest neighbours are identified by the Euclidean distance on standardized variables.

Swedish NFI and the current sampling strategy

The current Swedish NFI follows the strategy developed by Ranneby et al. (1987). The country was divided into five strata with decreasing sampling intensities towards the north. Within each stratum, clusters of circular plots are sampled. The clusters were quadratic or rectangular in shape, with a side length varying from 300 to 1800 m between different parts of the country. The circular plots were located along the sides of the cluster with fixed distance between plots within stratum. The within-stratum fixed distance between plots increased by latitude. The design was motivated by assumed autocorrelation for relevant forest variables such as stem volume. In other words, the landscape changes more rapidly in the south with mixed species forests, while the boreal conifer forests in the north are more homogenous and often dominated by one species. Thus, longer distances between plots was needed in the north to obtain new information.

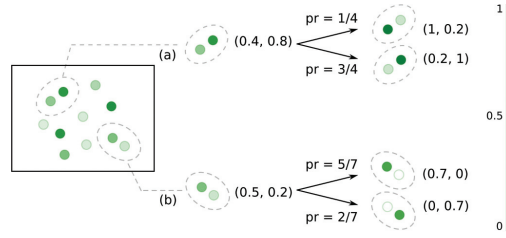


Fig. 2. One step in the local pivotal method for a pair of nearby units i and j. The intensity of the colour correlates with the inclusion probability. (*a*) If $\pi_i + \pi_j > 1$, then the winner receives probability 1 and will definitely be included. (*b*) If $\pi_i + \pi_j < 1$, then the loser receives probability 0 and will definitely be excluded. [Colour online.]

Two kinds of clusters are used: temporary ones and permanent ones. The temporary clusters are mainly intended to capture the current state of the forest and are only surveyed once, whereas permanent clusters primarily aim to capture changes and are resurveyed regularly (Tomppo et al. 2010, chap. 35). The selections in different strata are independent, and the estimation for target variables is required at the stratum level. A sample of the survey clusters, systematically distributed over the whole country, is measured annually from early May to mid-October. A 5 year inventory cycle is used, using five consecutive yearly inventories, and the estimates are calculated as a 5 year moving average. Separate estimators are used for each year and each cluster type, and a weighting is used to calculate averages of both cluster types. Details about the estimators used in the Swedish NFI can be found in Ranneby et al. (1987) and Fridman et al. (2014, appendices A–C).

The current sampling strategy (2013–2017) of temporary clusters is based on the R Package "spsample" using an unaligned systematical sampling design. This specific systematic design is used mainly to spread the sample geographically and thus also avoid the risk of overlapping sample units.

Implementation of the new strategy in Sweden

To evaluate the potential improvement in efficiency by introducing the new sampling strategy in Sweden, a simulation was performed for selecting the positions of temporary clusters of the Swedish NFI. The efficiency of alternatively using two reference sampling strategies was compared with the new sampling strategy. The new sampling strategy, denoted LPM-5 (LPM using five auxiliary variables), is in many ways similar to the previous strategy. We use the same geographical stratification and the same number of clusters. The main difference is that the new strategy uses auxiliary information in the sampling design to ensure that the selected clusters are more representative. As the first reference sampling strategy, we use IRS where the clusters are randomly and independently distributed over the area. The second reference sampling strategy (LPM-*xy*) is the LPM with geographical spread, which represents a proxy for the current strategy. The reason for including IRS is that we then can see also the effect of geographical spread.

We selected Region 3 in the middle of Sweden as our study region (see Fig. 4). In this region, the clusters consist of 12 circular plots of 7 m radius. The plots in a cluster are placed along a square

Fig. 3. Results for the one-dimensional example. Box plots for sample mean, spatial balance, and maximum distance for independent random sampling and the local pivotal method, respectively. All of the results are based on a simulation of 1000 samples of size 350, and for the local pivotal method, we used a first-phase sample of size $N = 100\,000$. [Colour online.]

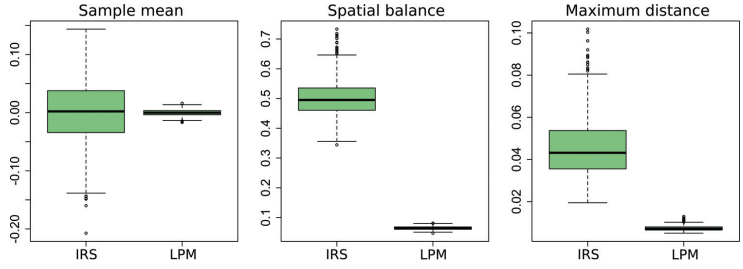


Fig. 4. Illustration of the selected region. [Colour online.]

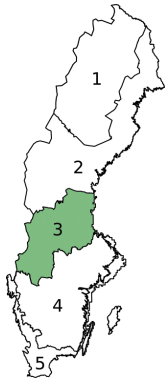
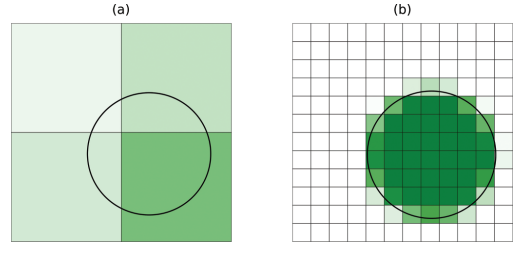


Fig. 5. Illustration of how we derive the plot total of auxiliaries for a 7 m radius plot. Each cell receives a weight proportional to the area of its intersection with the plot, which correlates with the intensity of the colour in the figure. (*a*) An example for the tree height and the basal area, which are available on a 12.5 m × 12.5 m grid. (*b*) An example for elevation, which is available on a 2 m × 2 m grid. [Colour online.]



formation with a side length of 1500 m and with 500 m between plots. Five auxiliary variables were used simultaneously with equal weights to spread the sample for the new strategy. These variables were geographical coordinates of the cluster center, the mean elevation of the cluster, the cluster mean tree height, and the mean basal area. Elevation was derived from a digital elevation model, while tree height and basal area were derived from remote sensing information from airborne laser scanning data, which were collected between 2009 and 2015. The forest variables were estimated by regression models combining NFI plot data with airborne laser scanning data metrics and were available on a nationwide map (Nilsson et al. 2017).

For the first-phase sample, a 100 000 clusters were independently selected. For each such cluster of plots, the cluster response of the five auxiliary variables was derived. Then a subset of size 360 of clusters was selected by the LPM-5 and the two reference designs, respectively. Spatial balance, design effects, and estimators for the auxiliary variables were compared by a Monte-Carlo simulation.

Equation 3 can be employed to calculate the value of auxiliaries for the point response of a cluster. However, it is unpractical to use the expression of $\mathbb{Z}^*(\mathbb{X})$ directly, since it is difficult to integrate the function in the equation. As we match the distribution of the derived auxiliary response, we are free to introduce any approximation to the auxiliary response.

The inclusion zones for a point within a plot vary less than they vary within a cluster. Hence, it is natural to set an equal value of the area of the inclusion zone for all points in the same plot. Then, the response of the cluster can be calculated by a weighted sum

Table 1. Design effect for five auxiliary variables with respect to reference designs.

	Design effect		
Auxiliary variable	$\hat{V}_{\text{LPM-5}}/V_{\text{IRS}}$	$\hat{V}_{\text{LPM-5}}/\hat{V}_{\text{LPM-}xy}$	$\hat{V}_{\text{LPM-}xy}/V_{\text{IRS}}$
x-coordinate	0.030	5.104	0.006
y-coordinate	0.032	5.107	0.006
Elevation	0.036	0.303	0.121
Tree height	0.036	0.061	0.589
Basal area	0.035	0.059	0.603

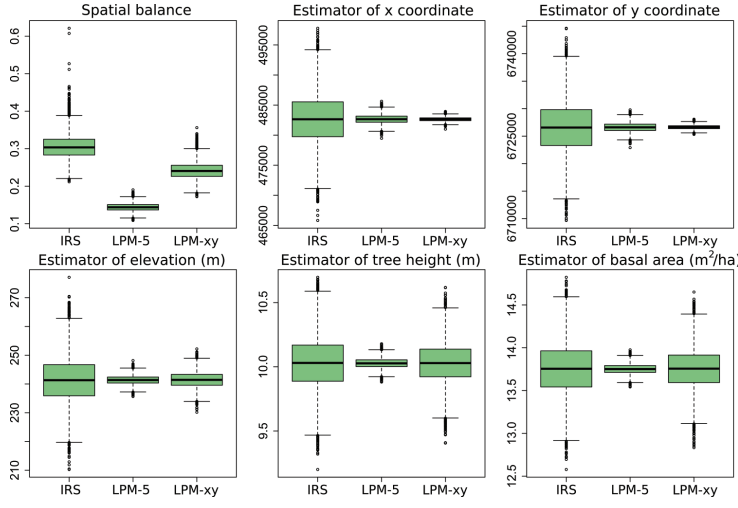
Note: First-phase sample size is 100 000, second-phase sample size is 360, and 10 000 samples were generated. LPM-5, local pivotal method with all five auxiliary variables; LPM-*xy*, local pivotal method with only *xy*-coordinates; IRS, independent random sampling. The variance ratios presented are called design effects.

over the plots. To achieve this, we introduce an approximation by assuming all points in a plot have the same inclusion zone as the plot center. The cluster response 3 can then be approximated as

$$\mathbb{Z}^*(\mathbb{X}) = \sum_{i=1}^{n_C} \int_{\mathbb{X}' \in \mathcal{C}_i(\mathbb{X}) \cap F} \frac{\mathbb{Z}'(\mathbb{X}')}{\ell[\mathrm{K}(\mathbb{X}')]} \mathrm{d}\mathbb{X}' \approx \sum_{i=1}^{n_C} \frac{1}{\ell_i(\mathbb{X})} \int_{\mathbb{X}' \in \mathcal{C}_i(\mathbb{X}) \cap F} \mathbb{Z}'(\mathbb{X}') \mathrm{d}\mathbb{X}' = \mathbb{Z}(\mathbb{X})$$

where $\mathcal{C}_i(\mathbb{X})$ is plot i in the cluster centered at $\mathbb{X}$, n_C is the number of plots in a cluster, and $\ell_i(\mathbb{X})$ is the surface area of the inclusion zone of the center point of plot i in the cluster. The integral

Fig. 6. Box plots of spatial balance and estimators for the five auxiliary variables. LPM-5, local pivotal method with all five auxiliary variables; LPM-*xy*, local pivotal method with only *xy*-coordinates; IRS, independent random sampling. [Colour online.]



$$\int_{\mathbb{X}' \in C_i(\mathbb{X}) \cap F} \mathbb{Z}'(\mathbb{X}')\mathrm{d}\mathbb{X}' \quad (5)$$

is the total of the single point response on plot *i* in the cluster. We obtain this plot total if we multiply cell values with respect to intersected area of the plot. Figure 5 is an example of how we weight the grid cells to calculate equation 5 of auxiliary variables derived from airborne laser scanning and digital elevation model, respectively. The values of auxiliary variables for each grid cell were available beforehand (e.g., see Nilsson et al. 2017). The resolution of the grid cell is 12.5 m × 12.5 m for the airborne laser scanning data and 2 m × 2 m for the elevation. The radius of each plot is 7 m.

Table 1 and Fig. 6 demonstrate variance for the estimator of the five auxiliary variables with respect to the three designs. Compared with IRS, the reduction of the variance was more than 95% for all five auxiliary variables when using LPM-5. We have also reduced variance by more than 90% for mean tree height and mean basal area, even compared with the design that spreads geographically (LPM-*xy*). We can clearly see from the table, if we just spread the samples geographically, that the reduction of the variance was less than 45% of mean tree height and mean basal area compared with IRS. The mean of the spatial balance was 0.144, 0.242, and 0.306 for LPM-5, LPM-*xy*, and IRS, respectively.

Conclusion and discussion

We proposed a new sampling strategy that uses auxiliary information in the sampling design in a continuous frame. Based on a simulation study, we illustrated that the new strategy performed better than the reference strategies for selecting the temporary clusters within the Swedish NFI. For the new NFI design (LPM-5), each selected sample is representative of the auxiliary space. The spatial balance indicates a very good fit of the multivariate distribution, and as a consequence, the variances for the sample means of the auxiliary variables are significantly reduced (which implies the potential to reduce the variances for the target variables related to the auxiliary variables).

The approximation $\mathbb{Z}(\mathbb{X})$ introduces only very slight disturbance to the auxiliary response (and only for the response close to the forest borders). Far enough from the boundary, all points in a plot have the same inclusion zone, which means that there is no approximation for such a cluster, i.e., $\mathbb{Z}(\mathbb{X}) = \mathbb{Z}^*(\mathbb{X})$. The overall approach is purely design based and provides unbiased estimators for the target variables, no matter how the auxiliary variables are derived. We want to derive them in a similar way as the targets to not lose strength in the possible relationship and thus maximize the efficiency for estimation of target variables related to the auxiliary variables.

For the application study of the new strategy in Sweden, the auxiliary variables that we used for the sampling design are related to most of the target variables of NFIs. Therefore, adapting the NFI to the proposed strategy will lead to visible improvements for the estimation of the related target variables. If a variable is not related to the auxiliaries, the new strategy will not make their estimation worse.

The observed potential of using the new sampling strategy confirms the claims from earlier studies. In the article by Grafström and Ringvall (2013), another sampling design called the local cube method confirmed the advantages of selecting spatially balanced samples. However, the LPM tends to produce slightly better spread than the local cube method, and we chose to prioritize a better spread due to the multipurpose nature of NFIs.

According to Henttonen and Kangas (2015), the optimal sampling strategy depends heavily on the purpose of the inventory; thus, prioritizing the forest characteristics is also needed if an optimal strategy is to be determined. For multipurpose forest inventories, when the number of characteristics of interest is large, the task becomes more complicated. To choose a proper sampling strategy while using the auxiliary variables in the design, we need to consider the relationship between the auxiliary variables and the target variables, e.g., balanced samples are optimal for linear relationships and spatially balanced samples perform better for nonlinear relationships (Grafström and Lundström 2013). The encouraging results of this study have led to a decision to implement

this sampling strategy in all regions for the selection of temporary tracts within the Swedish NFI, starting from 2018.

Acknowledgements

The authors are grateful to Jonas Jonzén, Henrik Persson, and Mats Högström for their contributions in providing the raster data and for technical support. We are thankful to the Swedish NFI for good cooperation and for partly funding this research. We would also like to thank two anonymous reviewers and an Associate Editor for suggestions that improved the paper.

References

Barabesi, L. 2003. A Monte Carlo integration approach to Horvitz–Thompson estimation in replicated environmental designs. Metron, **61**(3): 355–374.

Barabesi, L. 2004. Replicated environmental sampling design and Monte Carlo integration methods: two sides of the same coin. *In* Proceedings of the XLII Conference of the Italian Statistical Society, Bari, Italy, 9–11 June 2004.

Benedetti, R., Piersimoni, F., and Postiglione, P. 2015. Sampling spatial units for agricultural surveys. Springer, Berlin Heidelberg. doi:10.1007/978-3-662-46008-5.

Cordy, C.B. 1993. An extension of the Horvitz–Thompson theorem to point sampling from a continuous universe. Stat. Probab. Lett. **18**(5): 353–362. doi:10.1016/0167-7152(93)90028-H.

Dehardt, J. 1971. Generalizations of the Glivenko–Cantelli Theorem. Ann. Math. Stat. **42**(6): 2050–2055. doi:10.1214/aoms/1177693073.

Deville, J.-C., and Tillé, Y. 2004. Efficient balanced sampling: the cube method. Biometrika, **91**(4): 893–912. doi:10.1093/biomet/91.4.893.

Eriksson, M. 1995. Design-based approaches to horizontal-point-sampling. For. Sci. **41**(4): 890–907.

Fridman, J., Holm, S., Nilsson, M., Nilsson, P., Ringvall, A.H., and Ståhl, G. 2014. Adapting National Forest Inventories to changing requirements — the case of the Swedish National Forest Inventory at the turn of the 20th century. Silva Fenn. **48** (3): 1095. doi:10.14214/sf.1095.

Grafström, A., and Lisic, J. 2016. BalancedSampling: balanced and spatially balanced sampling [online]. R package version 1.5.2. Available from http://www.antongrafstrom.se/balancedsampling/.

Grafström, A., and Lundström, N.L.P. 2013. Why well spread probability samples are balanced. Open J. Stat. **3**(1): 36–41. doi:10.4236/ojs.2013.31005.

Grafström, A., and Ringvall, A.H. 2013. Improving forest field inventories by using remote sensing data in novel sampling designs. Can. J. For. Res. **43**(11): 1015–1022. doi:10.1139/cjfr-2013-0123.

Grafström, A., and Schelin, L. 2014. How to select representative samples. Scand. J. Stat. **41**(2): 277–290. doi:10.1111/sjos.12016.

Grafström, A., Lundström, N.L.P., and Schelin, L. 2012. Spatially balanced sampling through the pivotal method. Biometrics, **68**(2): 514–520. doi:10.1111/j.1541-0420.2011.01699.x.

Grafström, A., Saarela, S., and Ene, L.T. 2014. Efficient sampling strategies for forest inventories by spreading the sample in auxiliary space. Can. J. For. Res. **44**(10): 1156–1164. doi:10.1139/cjfr-2014-0202.

Gregoire, T.G., and Valentine, H.T. 2008. Sampling strategies for natural resources and the environment. CRC Press, Boca Raton, Fla.

Henttonen, H.M., and Kangas, A. 2015. Optimal plot design in a multipurpose forest inventory. For. Ecosyst. **2**: 31. doi:10.1186/s40663-015-0055-2.

Mandallaz, D. 1991. A unified approach to sampling theory for forest inventory based on infinite population and superpopulation models. Ph.D. thesis, ETH Zürich, Zürich. doi:10.3929/ethz-a-000585900.

Mandallaz, D. 2007. Sampling techniques for forest inventories. CRC Press, Boca Raton, Fla.

Nilsson, M., Nordkvist, K., Jonzén, J., Lindgren, N., Axensten, P., Wallerman, J., Egberth, M., Larsson, S., Nilsson, L., Eriksson, J., and Olsson, H. 2017. A nationwide forest attribute map of Sweden predicted using airborne laser scanning data and field data from the National Forest Inventory. Remote Sens. Environ. **194**: 447–454. doi:10.1016/j.rse.2016.10.022.

Ranneby, B., Cruse, T., Björn, H., Härje, J., and Johan, S. 1987. Designing a new national forest survey for Sweden. Stud. For. Suec. 177.

Särndal, C.-E., Swensson, B., and Wretman, J. 2003. Model assisted survey sampling. Springer, New York.

Stevens, D.L., and Olsen, A.R. 2004. Spatially balanced sampling of natural resources. J. Am. Stat. Assoc. **99**(465): 262–278. doi:10.1198/016214504000000250.

Tillé, Y., and Wilhelm, M. 2017. Probability sampling designs: principles for choice of design and balancing. Stat. Sci. **32**(2): 176–189. doi:10.1214/16-STS606.

Tomppo, E., Gschwantner, T., Lawrence, M., and McRoberts, R.E. 2010. National Forest Inventories: pathways for common reporting. Springer, Dordrecht, Netherlands. doi:10.1007/978-90-481-3233-1.

Wolfowitz, J. 1954. Generalization of the Theorem of Glivenko–Cantelli. Ann. Math. Stat. **25**(1): 131–138. doi:10.1214/aoms/1177728852.

II

Received: 11 March 2019 | Revised: 4 February 2020 | Accepted: 8 February 2020

DOI: 10.1002/env.2625

RESEARCH ARTICLE

WILEY

A sample coordination method to monitor totals of environmental variables

Xin Zhao | **Anton Grafström**

Department of Forest Resource Management, Swedish University of Agriculture Sciences, Uppsala, Sweden

Correspondence
Xin Zhao, Department of Forest Resource Management, Swedish University of Agriculture Sciences, Umeå, Sweden.
Email: xin.zhao@slu.se

Abstract

A new sampling strategy for design-based environmental monitoring is proposed. It has the potential to produce superior estimators for totals of environmental variables and their changes over time. In the strategy, we combine two concepts known as spatially balanced sampling and coordination of samples. Spatially balanced sampling can provide superior estimators of totals, while coordination of samples over time is often used to improve estimators of change. Compared with reference strategies, we show that the new strategy can improve the precision of the estimators of totals and their change simultaneously. A forest inventory application is used to illustrate the new strategy and the results can be summarized as (i) using auxiliary information to spread the sample can improve the estimators of totals; (ii) a positive coordination of the samples reduced the variance of the estimator of change by more than 37% compared with independent samples; and (iii) a high overlap between successive samples does not guarantee a good estimator of change. The presented strategy can be used to develop more efficient environmental monitoring programs.

KEYWORDS

positive sample coordination, spatially balanced samples, spatially correlated Poisson sampling

1 | INTRODUCTION

Environmental monitoring is defined as the observation and study of the environment (Awange, 2012, ch. 1). The approach for environmental monitoring is to collect and analyze a subset that represents the environment in space and time (Artiola, Pepper, & Brusseau, 2004, ch. 2). In the whole process, sampling is usually employed as a tool to select a representative portion from the population in order to do the analysis. As an important component of environmental monitoring, it provides the foundation of data required for assessments of environmental variables.

In this article, we propose a design-based sampling strategy for monitoring totals of environmental variables. Two concepts are combined in the strategy, spatially balanced sampling and coordination of samples over time. Spatially balanced sampling designs can provide representative samples, and sample coordination is a method to statistically control the overlap of successive samples. Spatially correlated Poisson sampling (SCPS) and sample coordination based on permanent random numbers (PRNs), introduced by Brewer, Early, and Joyce (1972), are used in the algorithm for the new strategy. SCPS was first presented by Grafström (2012) as a spatial sampling method for selecting well-spread samples

from finite populations. We show how sample coordination can be applied within a continuous population framework by using a double (two-phase) sampling approach. Auxiliary information is derived for a very large first-phase sample, and from the finite first-phase sample we select well-spread and positively coordinated samples using SCPS with the aid of the auxiliary variables.

Two common objectives of environmental monitoring are to characterize the current state of some resource as well as the change or trend in the state over time and space (Marker & Stevens Jr., 2009). Here, we consider only states that can be expressed as totals of environmental variables. Good estimators of such states can be achieved by using representative samples of the population at different time occasions, which means that states are generally best detected by updating the samples according to the current population. A high degree of overlap level between successive samples can produce more precise estimators of change over time (see Qualité & Tillé, 2008; Sen, 1973). Permanent samples have traditionally been used to address this issue. The samples are then often systematically distributed over the landscape (Scott, 1998).

Huge amount of financial resources are spent on environmental monitoring all over the world and it is very important to apply more efficient sampling strategies that can increase the quality of estimates while potentially saving a considerable amount of costs. McDonald (2003) provided a very detailed review of different survey designs for large-scale environmental monitoring programs, and the split panel designs proposed by Kish (1983, 1986) were recommended among many other designs.

One feature of environmental populations is that they exist in a spatial context. Commonly, the responses of nearby locations tend to be more similar than the responses of locations which are farther apart. Thus, the spatial distribution of these populations can be used as important information when designing the sample. Several efficient spatial sampling methods have recently been developed for sampling from georeferenced populations. One of the first and the most widely used method is called Generalized Random Tessellation Stratified (GRTS) design proposed by Stevens and Olsen (2004). This method uses a random mapping to map the two-dimensional locations into one dimension while preserving some spatial relationship. Then the systematic πps sampling is applied and the sample is mapped back to the two-dimensional original space. Using this design the samples can be spread evenly over the geographical space. Robertson, Brown, McDonald, and Jaksons (2013) extended the idea of GRTS to a new design called balanced acceptance sampling (BAS). To select a spatially balanced sample using BAS, we need first to specify a d-dimensional hyperrectangular box that encloses the population. For example, it could be a rectangle that encloses a two-dimensional geographical study area. Then a two-dimensional random start Halton sequence with uniformly spread points in the rectangle is generated. The first n points that are observed in the study area will constitute the sample of size n. An alternative design to BAS which is called Halton iterative partitioning (HIP) was introduced by Robertson, McDonald, Price, and Brown (2018) to overcome some drawbacks of BAS for finite populations.

Another feature of environmental population is that there are often some auxiliary information (other than geographical coordinates) available. Nowadays, there is a wealth of remotely sensed information available from satellite, aerial photography, or laser scanning that can be used to efficiently distribute the sample units. Therefore, we shall take the full advantage of the available auxiliary information and the properties of the environmental populations when we derive the sampling strategies. The local pivotal method (LPM) and SCPS proposed by Grafström (2012) and Grafström, Lundström, and Schelin (2012) are two spatial sampling designs that employ auxiliary variables (often including geographical coordinates plus several other attribute variables) to spread the samples based on distances. By adding more variables that are related to the target variables when spreading the samples, we may get more representative samples and then improve the precision of the estimators. Since the samples are spread in all of the auxiliary variables, they will also provide a good basis for model-based inference.

Grafström (2012) and Benedetti, Piersimoni, and Postiglione (2015, ch. 7) showed that SCPS was more efficient than GRTS when the auxiliary variables were only the geographical coordinates. Robertson et al. (2013) argued that the statistical performance of BAS was comparable with LPM and SCPS when we only spread the samples in the geographical space. Compared with GRTS and BAS (HIP), the biggest advantage of the LPM and SCPS designs is that they can be used with any type and any number of auxiliary variables. Different from BAS (HIP), the selections are based on distances. By applying LPM and SCPS, we can spread the samples in the auxiliary variables even if they do not constitute dimensions of the population (Grafström & Matei, 2018a).

For environmental monitoring programs that cover large areas, many target variables usually vary rapidly over the landscape with respect to the low sampling intensity (Dobbie, Henderson, & Stevens, 2008). This means that geographically spread samples over the landscape may not optimally capture the distribution of the target variables (Grafström, Zhao, Nylander, & Petersson, 2017). A sample which is well spread in auxiliary variables implies that the sample has the

potential to be well spread also for the target variables that are related to the auxiliary variables. Moreover, a sampling design that uses auxiliary variables to spread the sample is particularly useful for multipurpose monitoring programs (Grafström & Schelin, 2014). To avoid misunderstanding, we would like to clarify that a well-spread sample in this article means that the sample is well spread in the auxiliary variables.

Various sampling strategies have been developed based on sample coordination (e.g., Ernst, 1999; Keyfitz, 1951; Kish & Scott, 1971; Patterson, 1950). There are two kinds of coordination: positive coordination and negative coordination. They intend to maximize or minimize the overlap of two or more samples, respectively. We will focus only on positive coordination in order to produce good estimators of change.

Consider sampling from a dynamic population, the coordinated samples selected at different time occasions are dependent. The degree of coordination is measured by the expected size of the overlap between samples. In essence, the positive coordination method based on PRNs consists of assigning a uniformly distributed random number on the interval [0,1] to each unit in the frame. The numbers assigned remain with the units over time, and such a number is called a PRN. These PRNs are used to decide the sampling outcomes (inclusions or exclusions) at each time point.

As the population changes over time, there is a need to update the sample to account for the changes in the auxiliary variables. If we use independent well-spread samples at different time occasions, we can have good estimators of states. However, since the variance of the estimator of change equals the summation of the variances for the two state estimators minus two times their covariance, the estimator of change may not be the best as the covariance between the state estimators will be zero. A permanent sample has a major drawback, as new auxiliary information cannot be entered into the design. A sample that matches the distribution for "today" will be unlikely to match the distribution for "tomorrow." This is because the developments of the sample and the population may be different over time. Thus, if we select a sample that is good at the time of selection, the quality of such a sample is likely to deteriorate over time, which impacts both state and change estimators. Therefore, we need to have an adaptive design that update the sample to maintain the representativeness of the changing population. When a new sample is selected, we want to have overlap (common units) between the new and the previous samples, so that we also improve estimators of change.

Compared with four reference strategies, (i) a strategy that uses permanent geographical-spread samples; (ii) a strategy that employs permanent well-spread samples; (iii) a strategy with independent well-spread samples; (iv) a strategy that applies a split-panel design), the superiority of the new strategy is illustrated by a forest inventory application. By using Monte-Carlo simulations, we show that the new strategy can outperform the reference strategies for both state and change estimators.

This article is mainly intended for those with a solid background in statistics and those who work with survey designs related to environmental monitoring. The rest of the article is organized as follows. In Section 2, a continuous sampling framework and the double sampling approach are introduced, sample coordination for a finite population is explained, and spatial balance is defined. Details about the new strategy are presented in Section 3. In Section 4, we use a simulation study to compare the new strategy against the four reference strategies. Conclusions and comments are given in Section 5.

2 | SAMPLING FRAMEWORK

Let F denote the continuous population that we sample from, and F is assumed to be a bounded open subset of the Euclidean plane $\mathbb{R}^2$, with surface area $\ell(F)$. The set F is considered to be fixed over time. The response of a target variable for a point $\mathbf{x} \in F$ at time t is denoted as $y_t(\mathbf{x})$. The population total of the response for the target variable at time t can hence be expressed as $Y(t) = \int_F y_t(\mathbf{x})\, d\mathbf{x}$. The sampling design of size $n(t)$ on F at time t is specified by a joint distribution of $n(t)$ random variables $\mathbf{x}_1, \ldots, \mathbf{x}_{n(t)}$. The sampling intensity at time t is given by $\pi_t(\mathbf{x}) = \sum_{i=1}^{n_t} f_{ti}(\mathbf{x})$, where $f_{ti}(\mathbf{x})$ is the marginal probability density function of $\mathbf{x}_i$ at time t. We have $\int_F \pi_t(\mathbf{x})\, d\mathbf{x} = n(t)$.

2.1 | Double sampling approach

To use auxiliary information in the sampling design, a double sampling approach can be employed for the continuous population. In the first-phase sample, a large number N of locations are selected independently using a constant sampling intensity, $\pi(\mathbf{x}) = N/\ell(F)$. We let U be the indexes of the geographical locations, $U = \{1, \ldots, i, \ldots, N\}$. Then U serve as a dynamic (but permanent with respect to locations and indexes) frame over time. Auxiliary information is then

derived for each unit from U at different time occasions. According to the Glivenko–Cantelli theorem and its multivariate generalizations, the empirical distribution of the auxiliary variables in the first-phase sample converges uniformly almost surely to the population distribution as the size of the sample increases (see Dehardt, 1971; Wolfowitz, 1954). Because of the large sample size, the empirical distribution of any variable in the first-phase sample U will closely match the population distribution. The realization of the first-phase sample of size N can be treated as a permanent frame over all repeated surveys. Then, we select the second-phase samples $S(t)$ at different time occasions, $S(t) \subset U$. Denote the first-phase sample of N random locations as $S = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$. Conditioned on S, the inclusion probability of unit $i \in U$ to be included in $S(t)$ is denoted as $\pi_i(t)$. We define the target variable as $y_i(t) = y_t(\mathbf{x}_i)/\pi(\mathbf{x}_i)$, for $i \in U$. Then, the total of the target variable in the first-phase sample at time t can be written as $Y_U(t) = \sum_{i \in U} y_i(t)$. Moreover, let $I_i(t)$ denote the inclusion indicator for unit i at time t, so that $I_i(t) = 1$ if $i \in S(t)$ and $I_i(t) = 0$ otherwise.

To preserve the relationship between the auxiliary variables and the target variables, we define the response of the auxiliary variables as $\mathbf{z}_i(t) = \left(z_{1i}(t), \dots, z_{h_t i}(t)\right)^T \in \mathbb{R}^{h_t}$, and $\mathbf{z}_i(t) = \mathbf{z}_t(\mathbf{x}_i)/\pi(\mathbf{x}_i)$, where $\mathbf{z}_t(\mathbf{x}_i)$ is the auxiliary response for the point $\mathbf{x}_i$ at time t. The target responses are observed for the locations in the second-phase sample in order to estimate the state $Y(t)$. The unbiased Horvitz–Thompson (HT) estimator is then defined as

$$\widehat{Y}(t) = \sum_{i \in S(t)} \frac{y_i(t)}{\pi_i(t)} = \sum_{i \in S(t)} \frac{y_t(\mathbf{x}_i)}{\pi_t(\mathbf{x}_i)}, \tag{1}$$

where $\pi_t(\mathbf{x}_i) = \pi(\mathbf{x}_i) \cdot \pi_i(t)$. The estimator $\widehat{Y}(t)$ is conditionally unbiased for $Y_U(t)$ and unconditionally unbiased for $Y(t)$.

As N is supposed to be much larger than the sample sizes, we allow ourself to do estimation and variance estimation conditioned on the first-phase sample. In that case, the variance of the HT-estimator (1) can be expressed as

$$\operatorname{var}(\widehat{Y}(t)) = \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\pi_{ij}(t) - \pi_i(t)\pi_j(t)\right) \frac{y_i(t)}{\pi_i(t)} \frac{y_j(t)}{\pi_j(t)}, \tag{2}$$

where $\pi_{ij}(t) = \Pr\left(i \in S(t), j \in S(t)\right)$ is the second-order inclusion probability for a pair of points (i, j) at time t. An estimator of (2) is

$$\widehat{\operatorname{var}}(\widehat{Y}(t)) = \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\pi_{ij}(t) - \pi_i(t)\pi_j(t)\right) \frac{y_i(t)}{\pi_i(t)} \frac{y_j(t)}{\pi_j(t)} \frac{I_i(t) I_j(t)}{\pi_{ij}(t)}. \tag{3}$$

The estimator (3) is unbiased for (2) provided that all second-order inclusion probabilities are strictly positive.

We define the change of the states between two time points as $\Delta_Y(1,2) = Y(2) - Y(1)$ and its estimator as $\widehat{\Delta}_Y(1,2) = \widehat{Y}(2) - \widehat{Y}(1)$. Since $\widehat{Y}(t)$ is unbiased at any time t, the estimator of change is unbiased as well. When estimating the variance of change from one time occasion to another time occasion, say Occasion 1 and 2, we are also interested in estimating the covariance $\operatorname{cov}(\widehat{Y}(1), \widehat{Y}(2))$. The variance of the change estimator is

$$\operatorname{var}\left(\widehat{\Delta}_Y(1,2)\right) = \operatorname{var}\left(\widehat{Y}(2)\right) + \operatorname{var}\left(\widehat{Y}(1)\right) - 2\operatorname{cov}\left(\widehat{Y}(1), \widehat{Y}(2)\right). \tag{4}$$

The covariance term in (4) is given by

$$\operatorname{cov}(\widehat{Y}(1), \widehat{Y}(2)) = \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\pi_{ij}(1,2) - \pi_i(1)\pi_j(2)\right) \frac{y_i(1)}{\pi_i(1)} \frac{y_j(2)}{\pi_j(2)}, \tag{5}$$

where $\pi_{ij}(1,2) = \Pr\left(i \in S(1), j \in S(2)\right)$. It is possible to construct the HT-estimator of (5) based on the two samples, that is,

$$\widehat{\operatorname{cov}}(\widehat{Y}(1), \widehat{Y}(2)) = \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\pi_{ij}(1,2) - \pi_i(1)\pi_j(2)\right) \frac{y_i(1)}{\pi_i(1)} \frac{y_j(2)}{\pi_j(2)} \frac{I_i(1) I_j(2)}{\pi_{ij}(1,2)}. \tag{6}$$

The estimator (6) is unbiased for (5) provided that the $\pi_{ij}(1,2)$ are strictly positive for all i, j.

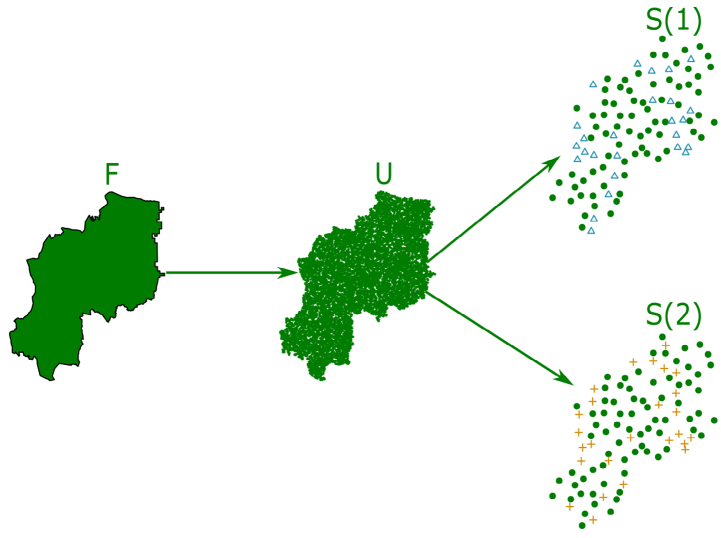


FIGURE 1 The general framework of the strategy for two time occasions. F is the continuous population frame. U stands for the first-phase sample with $N = 10,000$. $S(1)$ and $S(2)$ represent the positively coordinated second-phase samples, $n(1) = n(2) = 100$. The overlapped sample points are marked with solid circles, and the points that are on different locations are marked with triangles and crosses for $S(1)$ and $S(2)$, respectively

2.2 | Sample coordination for a finite population

The general framework of sample coordination was summarized by Grafström and Matei (2015). Here, we only consider the situation where we select samples from the same population U at different time occasions. Consider only two time occasions, that is, Occasion 1 and 2, the overall sampling design p is defined on $U \times U$, with marginal designs p_1 and p_2. A random sample of $n(1)$ locations selected at time Point 1 is denoted by $S(1)$, and a random sample of size $n(2)$ selected at time Point 2 is denoted by $S(2)$. The overall sampling design is said to be coordinated if the joint probability of selection of two samples is not equal to the product of the probabilities of selecting each separate sample, that is, if $p\,(s(1), s(2)) \neq p_1\,(s(1))\,p_2\,(s(2))$ (see Cotton & Hesse, 1992; Mach, Reiss, & Şchiopu-Kratina, 2006).

Our aim of coordination is to maximize the overlap between several samples drawn successively from U. Therefore, the selection of a new sample will depend on the samples previously drawn. In order to obtain a larger or a smaller overlap of the samples, a dependence between the samples must be introduced. This dependence will determine the expected number of common units in the selected samples. Let $\mathcal{O}$ denote the random variable "size of the overlap," $\mathcal{O} = \sum_{i \in U} I_i(1) I_i(2)$. The coordination degree between two samples is measured by the expected size of the overlap

$$E(\mathcal{O}) = \sum_{i \in U} E\,(I_i(1) I_i(2)) = \sum_{i \in U} \pi_i(1, 2), \tag{7}$$

where $\pi_i(1, 2) = \Pr\,(i \in S(1) \cap S(2))$ is the probability for unit i to be included in both $S(1)$ and $S(2)$. According to Mach et al. (2006), the expected size of the overlap may also have impact on the precision of the change estimators between two occasions. Figure 1 is an illustration of how we construct the finite framework from F and how we select positively coordinated samples from U for two time occasions. In this example, percentage of overlap of the two samples is 74%.

2.3 | Spatial balance

Spatial balance is a measure to check the spread of a spatial sample. It is often used when the auxiliary space is multidimensional. The measure is based on Voronoi polytopes (Stevens & Olsen, 2004). For a sample of size $n(t)$, we need to construct $n(t)$ polytopes. For each $i \in s(t)$, the polytope $\rho_i(t)$ includes all units in the population closer to i than to any other

sample unit $j \in s(t), j \neq i$. The distance used when we construct the polytopes is the Euclidean distance on standardized variables. Spatial balance of a sample at time occasion t can be measured by

$$B(t) = \frac{1}{n(t)} \sum_{i \in s(t)} (v_i(t) - 1)^2, \tag{8}$$

where $v_i(t) = \sum_{j \in \rho_i(t)} \pi_j(t)$ is the total probability mass within the polytope at time occasion t. Spatial balance can be interpreted as a measure of the variance of the total probability mass within the polytopes. A small value of $B(t)$ indicates the sample is well spread at time t. If the sample is spread perfectly, the total probability mass within $\rho_i(t)$ equals to 1. To measure how well a design succeeds in selecting spatially balanced samples, simulation to find the expected value of $B(t)$ under the design is needed.

3 | SAMPLE COORDINATION FOR SPATIALLY BALANCED SAMPLES

By using well-spread samples at each point in time, we are likely to reduce the variances of the state estimators. Increasing the expected size of the overlap between the samples, by a coordinated sample selection, may lead to a higher positive covariance between the state estimators. As Duncan and Kalton (1987) said, the reason for the increased covariance is that many sample units are the same in the two samples and their values tend to be similar at the two time occasions. Thus, by also introducing coordination, we will most likely achieve a smaller variance of the change estimator.

Under our framework of coordination for two (or more) samples, the first time occasion sample $S(1)$ and the second time occasion sample $S(2)$ have fixed sample sizes $n(1)$ and $n(2)$, respectively. The sample overlap $S(1,2) = S(1) \cap S(2)$ contains $n(1,2)$ units and $n(1,2)$ is not fixed. We would like to achieve a high degree of overlap without loosing the spatial balance compared with independent selection of samples.

3.1 | Spatially correlated Poisson sampling

With the aid of the auxiliary variables derived from the first-phase sample at time occasion t, we select a second-phase sample $S(t)$ of size $n(t)$ from the large first-phase sample. We would like to select a second-phase sample whose distribution of the auxiliary variables matches the distribution of the first-phase sample, and thus also match the population distribution at time t. Assuming the dependence between sampling units decrease as the distance between them increase, to minimize the sampling variance, we should select the sampling units so that we maximize the distance between them (Benedetti et al., 2015, ch. 7). In other words, the sample should be well spread over the auxiliary variables.

SCPS is a list-sequential sampling method of selecting well-spread samples. The method was first derived as a spatial application of correlated Poisson sampling, proposed by Bondesson and Thorburn (2008). It is a fixed size πps design that achieves a good spread of the selected samples by using the auxiliary variables. The main idea of SCPS is motivated by a generalization of Tobler's first law of geography. According to that law, geographically nearby locations tend to have more similar properties than locations farther apart. As the distance measure applied in SCPS's algorithm is the standardized Euclidean distance in the space of the auxiliary variables, we can call it "law of auxiliary variables" instead. If the auxiliary variables have high explanatory power for the target variables, then two units with a small distance in the auxiliary space will tend to have more similar values on the target variables than two units farther apart. Generally, in SCPS, the selection of nearby units is avoided to the furthest extent possible, which creates well-spread samples. This is guaranteed by creating a strong negative correlation between the inclusion indicators of nearby units.

3.2 | Algorithm of SCPS

It is assumed that we have a list U of the units to be sampled. The sampling outcome is first decided for the first unit in the list and then for the second, and so forth. After each sampling decision, the inclusion probabilities for the remaining

units in the list are updated. Denote the prescribed inclusion probability of each unit i at time t as $\pi_i(t), i = 1, 2, \dots, N$, with $\sum_{i=1}^{N} \pi_i(t) = n(t)$. Then, we have a starting vector of inclusion probabilities $(\pi_1(t), \dots, \pi_N(t))$. In the end of the algorithm, we will get a vector of inclusion indicators by gradually updating the vector of inclusion probabilities in maximum of N steps. The updating can be illustrated as

$$\begin{array}{lllllll} \boldsymbol{\pi}^{(0)}(t): & \pi_1(t) & \pi_2(t) & \pi_3(t) & \pi_4(t) & \dots & \pi_N(t) \\ \boldsymbol{\pi}^{(1)}(t): & I_1(t) & \pi_2^{(1)}(t) & \pi_3^{(1)}(t) & \pi_4^{(1)}(t) & \dots & \pi_N^{(1)}(t) \\ \boldsymbol{\pi}^{(2)}(t): & I_1(t) & I_2(t) & \pi_3^{(2)}(t) & \pi_4^{(2)}(t) & \dots & \pi_N^{(2)}(t) \\ \boldsymbol{\pi}^{(3)}(t): & I_1(t) & I_2(t) & I_3(t) & \pi_4^{(3)}(t) & \dots & \pi_N^{(3)}(t) \\ \quad \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\pi}^{(N)}(t): & I_1(t) & I_2(t) & I_3(t) & I_4(t) & \dots & I_N(t) \end{array}. \tag{9}$$

The first unit is included with probability $\pi_1^{(0)}(t) = \pi_1(t)$ at time occasion t. If the first unit was included, we set $I_1(t) = 1$, otherwise $I_1(t) = 0$. The sampling outcome at each step is decided by comparing the inclusion probability of the step unit with a random number associated with the unit. Denote the random number associated with the unit $i \in U$ at time t as $r_i(t)$, with $r_1(t), r_2(t), \dots, r_N(t)$ i.i.d. $U(0, 1)$. When the values for $I_1(t), \dots, I_{j-1}(t)$ have been decided for the first $j-1$ units in the list at time occasion t, the step unit j is included in the sample $S(t)$ at time t, that is, $I_j(t) = 1$, if $r_j(t) < \pi_j^{(j-1)}(t)$, and $I_j(t) = 0$ otherwise. The inclusion probabilities for the rest of the units in the list are updated according to

$$\pi_i^{(j)}(t) = \pi_i^{(j-1)}(t) - \left(I_j(t) - \pi_j^{(j-1)}(t)\right) w_i^{(j)}(t), \tag{10}$$

where $i = j+1, \dots, N$ and $i \geq 2$, $\pi_i^{(0)}(t) = \pi_i(t)$, $w_i^{(j)}(t)$ is the weight received by unit i from the step unit j at time t, $\sum_{i=j+1}^{N} w_i^{(j)}(t) = 1$. The weight $w_i^{(j)}(t)$ depends on the sampling outcomes of the first $j-1$ units. To make sure that $0 \leq \pi_i^{(j)}(t) \leq 1$ holds, the weights need to satisfy the following restrictions

$$-\min\left(\frac{1-\pi_i^{(j-1)}(t)}{1-\pi_j^{(j-1)}(t)}, \frac{\pi_i^{(j-1)}(t)}{\pi_j^{(j-1)}(t)}\right) \leq w_i^{(j)}(t) \leq \min\left(\frac{\pi_i^{(j-1)}(t)}{1-\pi_j^{(j-1)}(t)}, \frac{1-\pi_i^{(j-1)}(t)}{\pi_j^{(j-1)}(t)}\right). \tag{11}$$

In the SCPS, the decision (include or not) is always made for the step unit j, that is, the outcome of $I_j(t)$ is decided in step j. From Equation (11), we can see that it is possible for the weights to be negative. However, to achieve spatial balance, the weights need to be positive. When updating the inclusion probability for unit i, the weight it receives depends on the distance between i and the step unit j. If the inclusion indicator for the step unit is 1, then based on the inclusion probability in step $j-1$, the unit j needs to "steal" more probability from its nearby units. The closer the unit to the step unit, the more probability mass will be "stolen" by the step unit until the updated inclusion probability of the step unit becomes 1, that is, the nearest unit to j will receive as much weight as possible from j, then as much as possible weight will received by the second nearest unit from j, and so forth. By contrast, if the step unit has an inclusion indicator equal to 0, it will give away all its probability mass to its neighbors in a similar way. The above strategy is called maximal weight strategy. Thus, with the restriction of the maximum weight of each unit and the sum of the weights equal to 1, we will have a sample of a fixed size. The algorithm as well as an example of updating the inclusion probabilities of the SCPS can be found in Grafström (2012).

3.3 | Positive coordination under SCPS

Coordination of well spread samples using the SCPS was introduced for finite populations by Grafström and Matei (2018b). A coordination method based on PRN is used in the algorithm, where the random number associated with each unit in the algorithm of SCPS remains over time, that is, $r_i(1) = r_i(2) = r_i$ for two time occasions. Positive coordination is achieved by using the same comparison rule for all time occasions at each step, that is, $I_j(t) = 1$, if $r_j < \pi_j^{(j-1)}(t)$, and $I_j(t) = 0$ otherwise. The implementation of positive coordination using SCPS can be found in the R package "BalancedSampling" (Grafström & Lisic, 2019). The algorithm for two time occasions is illustrated in Example 1.

Unit	$z_i(1)$	$z_i(2)$	$\pi_i(t)$	r_i
1	5	5	3/4	0.9821
2	3	1	1/2	0.6782
3	6	7	1/2	0.8060
4	8	2	1/4	0.6342

Note: $z_i(t)$ is the value of the auxiliary variable of unit i at time t. $\pi_i(t)$ represents the prescribed inclusion probability of unit i at time t, it does not change from Time 1 to Time 2 in this example. r_i is the permanent random number associated with unit i.

TABLE 1 Auxiliary variables, prescribed inclusion probabilities, as well as the random numbers associated with the units for two time occasions in Example 1

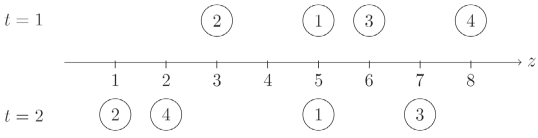


FIGURE 2 The distances between units of the auxiliary variable for two time occasions in Example 1

Example 1. Suppose we have a population of size $N = 4$, we want to select two positively coordinated samples of size $n = 2$ using SCPS at two time occasions. Table 1 presents the auxiliary value, prescribed inclusion probability as well as the random number associated with each unit at two time occasions. Besides the random number, the prescribed inclusion probability for each unit is also same at each time. According to Table 1, the distances between different units at both time occasions can be calculated, and they are illustrated by Figure 2. At each time occasion, the distance between units is calculated by comparing the value of the auxiliary variable for each unit. The smaller the differences for the values, the closer the units will be. The visiting order is chosen to be 1, 2, 3, 4, the decision is made for each unit according to the same order at both time occasions. The updating of inclusion probabilities for the units at both time occasions for positive coordination of SCPS is illustrated by Figure 3. The maximum weight and the updated inclusion probability for each unit at each step is calculated by using (10) and (11) at both time occasions. According to Figure 3, the overlap for the selected samples is 50% in the example.

When it comes to estimation under positively coordinated samples selected with SCPS, we may use the unbiased HT-estimator (1). However, for designs that produce well-spread samples, many of the second-order inclusion probabilities may be zero. Hence, it is not possible to have a design-based unbiased variance estimator. Based on squared local deviations, Grafström and Schelin (2014) derived an approximate variance estimator for the HT-estimator, under spatially balanced sampling. It can be expressed as

$$\widehat{\text{var}}(\hat{Y}(t)) = \frac{1}{2}\sum_{i\in S(t)}\left(\frac{y_i(t)}{\pi_i(t)} - \frac{y_{i'}(t)}{\pi_{i'}(t)}\right)^2, \tag{12}$$

where i' is the nearest neighbor to i in $S(t)$ at time t. The distance measure we apply to find i' is the same as we used when selecting the sample.

Estimation of the covariance between successive state estimators is difficult. The estimator (6) is only unbiased for (5) provided the $\pi_{ij}(1,2)$ are strictly positive for all i,j. However, that requirement does not hold in general for positively coordinated and spatially balanced samples. The reason is that it is likely that inclusion of a unit i at Time 1 often imply inclusion of unit i at Time 2 and hence also exclusion of neighboring units at Time 2. In such a case, the $\pi_{ij}(1,2)$ cannot be guaranteed to be strictly positive for all i,j. Finding a suitable estimator for the covariance (5) under positively coordinated and spatially balanced samples remains a challenging problem for the future.

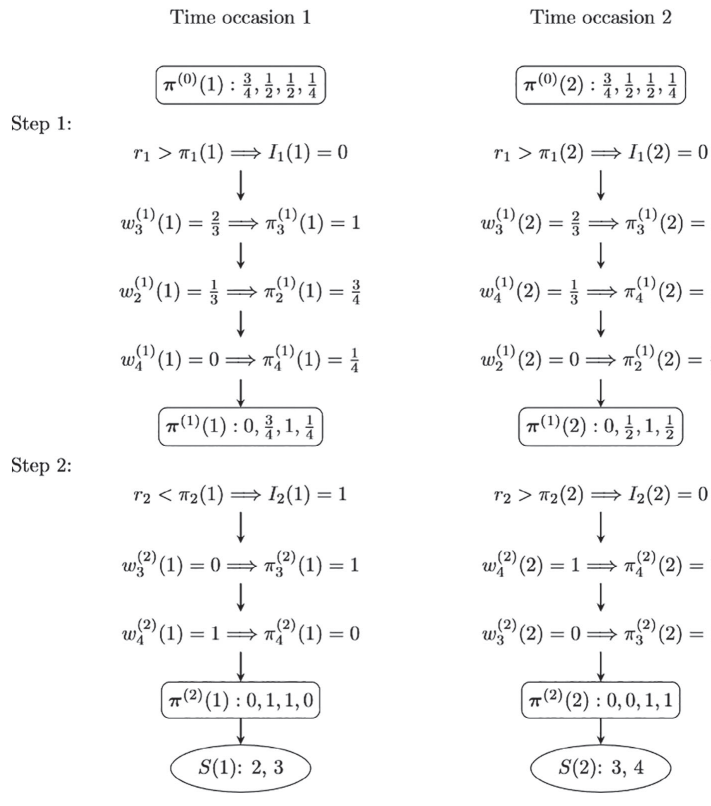


FIGURE 3 Sampling procedure for positively coordinated samples selected using spatially correlated Poisson sampling in Example 1. At each time and step, the order for updating the inclusion probabilities for the remaining units is decided by the distance between the step unit and the remaining units. The distances are measured in auxiliary variables

4 | A FOREST INVENTORY APPLICATION

In forest inventories, we usually sample circular plots or clusters of circular plots. The field inventories are based on the samples selected. Before each survey, the field staff need to find the center of the plots (clusters), then survey each plot with a fixed radius (survey each cluster with a fixed configuration). Since we lack a list frame for the individual objects, the continuous sampling framework can be applied in forest inventories (see e.g. Mandallaz, 2007, ch. 4; Grafström, Schnell, Saarela, Hubbell, & Condit, 2017). Suppose we have a finite population $U(t)$ of objects (e.g., trees) at time t, $U(t) = \{1, \dots, N(t)\}$. The number of objects can be different at different time occasions because there will be births and deaths. As the finite population cannot be partitioned into circular plots, it is impractical to sample the objects directly. Thus, we need to construct a continuous framework from the original finite framework, then do the sample selections based on the continuous framework.

In this section, the new strategy is compared with four reference strategies by a Swedish National Forest Inventory (NFI) example using simulation. The new sampling strategy as well as the four reference strategies are listed here.

- Strategy 1: The new strategy, which employs well-spread and positively coordinated samples selected by SCPS over time.
- Strategy 2: The first reference strategy. Use a permanent geographical-spread sample over time.
- Strategy 3: The second reference strategy. A sampling strategy that uses a permanent sample selected by SCPS, which is well spread at the first time occasion.
- Strategy 4: The third reference strategy, which uses independent well-spread samples selected by SCPS over time without sample coordination.

- Strategy 5: The last reference strategy. Use split-panel designs to split the sample into two parts: a panel with a permanent geographical-spread sample and a panel with well-spread samples. This strategy is similar to the current strategy of the Swedish NFI.

A region in the middle of Sweden is selected as our study region. In this region, each cluster consists of 12 circular plots of 7-m radius. Plots in a cluster are placed along a square formation with a side-length of 1,500 m and with 500 m between plots. Denote the response of target at time t as $\xi_i(t)$, then the population total at time occasion t can be expressed as $Y(t) = \sum_{i\in U(t)}\xi_i(t)$. When sampling clusters with a given configuration and a fixed orientation over time, the inclusion zone for an object i on location $\mathbf{x}_i$ is the collection of potential sample points, which lead to inclusions of the object. Mathematically, it can be denoted by $K_i \subset F, K_i = K(\mathbf{x}_i) = \{\mathbf{x} \in F : \mathbf{x}_i \in C(\mathbf{x})\}$, and $C(\mathbf{x})$ is the cluster centered on $\mathbf{x}$. Any cluster $C(\mathbf{x})$ with its cluster center $\mathbf{x}$ within K_i includes the object in one of its plots. Details about inclusion zone and the cluster configuration of the study region can be found in Grafström, Zhao, et al. (2017). Different formulations for the density function of the target variable have been discussed by Grafström, Schnell, et al. (2017). For constructing the continuous framework of the NFI application, the density function at time t can be expressed as

$$y_t(\mathbf{x}) = \sum_{i\in U_t} \frac{I_{ti}(\mathbf{x})\xi_i(t)}{\ell\,(K_i)}, \tag{13}$$

where $I_{ti}(\mathbf{x}) = 1$ if $\mathbf{x} \in K_i$ at time t, and 0 otherwise, $\ell\,(K_i)$ is the area of the inclusion zone for object i. By using the expression (13), the continuous population total is identical to the corresponding finite population total. This is because

$$Y(t) = \int_F y_t(\mathbf{x})\,d\mathbf{x} = \int_F \sum_{i\in U(t)} \frac{I_{ti}(\mathbf{x})\xi_i(t)}{\ell\,(K_i)}\,d\mathbf{x} = \sum_{i\in U(t)} \frac{\xi_i(t)}{\ell\,(K_i)} \int_F I_{ti}(\mathbf{x})\,d\mathbf{x} = \sum_{i\in U(t)} \xi_i(t). \tag{14}$$

Once we have constructed the continuous framework, we can easily employ the general framework in Section 2 for our example.

First, a number of 100,000 clusters were independently selected as an initial first-phase sample, then a subset of size 10,000 clusters were selected using SCPS as the first-phase sample, since the algorithm is quite computationally intensive. (This is only needed to do the simulation, we use the initial first-phase sample as the first-phase sample in reality because we only select the second-phase sample once.) Then, from the first-phase sample, a sample of size 100 was selected as the second-phase sample for both time occasions, respectively, using different strategies.

The auxiliary variables we used were the geographical coordinates, the mean tree height, the mean basal area, and the mean elevation. Denote the five auxiliary variables at the plot level as $\mathbf{q}_k(t) = \left(q_{x_k}(t), q_{y_k}(t), q_{h_k}(t), q_{b_k}(t), q_{e_k}(t)\right)^T \in \mathbb{R}^5$. According to Grafström, Zhao, et al. (2017), we only had auxiliary information for one time occasion. Stand-level growth models were applied to generate the 5-year's growth in the plot level for the mean tree height and the mean basal area. The growth models were based on data from permanent samples of the Swedish NFI established during 1983–1987 and reinventoried three to four times between 1988 and 2010 (Fridman, Holm, Nilsson, Ringvall, & Ståhl, 2014). We also applied a clear cutting with a rate of 5% for the second time occasion. First, 20% of the plots who had the highest mean tree height values were selected as the potential plots. Then, 1/4 of the plots among them were randomly selected for clear cutting. Based on the growth and the clear cutting, we got the mean tree height and the mean basal area at the next time occasion with a 5-years' time difference. The geographical coordinates as well as the elevation remained the same at time Occasion 2.

The 5-years' growth models are expressed for the mean tree height and the mean basal area in the plot level in Equations (15) and (16), respectively.

$$\Delta q_{h_k} = a_0 + a_1 q_{h_k}(1) + a_2 log\left(q_{h_k}(1)\right) + a_3 q_{b_k}(1) + \varepsilon_{h_k}, \tag{15}$$

$$\Delta q_{b_k} = b_0 + b_1 q_{b_k}(1) + b_2 log\left(q_{b_k}(1)\right) + b_3 q_{h_k}(1) + \varepsilon_{b_k}, \tag{16}$$

where $\varepsilon_{h_k} \sim N(0, \sigma_{h_k})$ and $\varepsilon_{b_k} \sim N(0, \sigma_{b_k})$. Before employing the auxiliary information to the sampling design, we aggregated the plot level auxiliary information $\mathbf{q}_k(t)$ to cluster-level $\mathbf{z}_i(t)$. Whether a plot contributes to the aggregated cluster value or not depends on if the plot center is inside of the region or not.

TABLE 2 Example National Forest Inventory

St	$B(1)$	$B(2)$	Overlap	$V\left(\hat{\bar{Z}}_h(1)\right)$	$V\left(\hat{\bar{Z}}_h(2)\right)$	$V\left(\hat{\Delta}_{\bar{Z}_h(1,2)}\right)$	$V\left(\hat{\bar{Z}}_b(1)\right)$	$V\left(\hat{\bar{Z}}_b(2)\right)$	$V\left(\hat{\Delta}_{\bar{Z}_b(1,2)}\right)$
1	0.129	0.127	62	0.809	0.744	0.978	0.017	0.017	0.022
2	0.239	0.238	100	10.311	10.165	2.130	0.233	0.242	0.048
3	0.129	0.171	100	0.809	2.475	1.969	0.017	0.058	0.043
4	0.129	0.128	1	0.809	0.776	1.608	0.017	0.018	0.035
P_{28}	0.197	0.167	20	3.157	1.115	4.169	0.071	0.026	0.094
P_{55}	0.224	0.209	50	4.694	2.122	5.986	0.106	0.050	0.138
P_{73}	0.236	0.228	70	6.486	4.111	7.370	0.148	0.098	0.171
P_{82}	0.241	0.236	80	7.869	6.125	7.411	0.177	0.144	0.170

Note: The first-phase sample size is $N = 10,000$, the second-phase sample size is $n(1) = n(2) = 100$, the repetition time is 10,000. For tree height, the sample mean for the two occasions is 100.27 and 101.50 dm, respectively. The percentage of change for the auxiliary is 95.17% and correlation of the two time occasions for the auxiliary is .9186. For basal area, the sample mean is 13.76 and 14.64 m^2/ha, respectively, for the two time occasions. The percentage of change is 95.17% and correlation coefficient of the two time occasions for the auxiliary variable is .9205. $B(i)$ is the mean of spatial balance for the sample selected at time occasion i. Overlap is the mean of the percentage of overlap for the sample units selected at two time occasions. $V\left(\hat{\bar{Z}}_i(t)\right)$ represents the empirical variance of the estimator for the mean of the auxiliary variable. $V\left(\hat{\Delta}_{\bar{Z}_i(1,2)}\right)$ stands for the variance of the estimator of change. P_{ij} is the split panel design, the value of i corresponding to the percentage of the permanent sample and value of j corresponding to the percentage of temporary sample. For example, P_{28} means the permanent panel is 20% and the temporary panel is 80%.

We applied the growth models since we would like to generate more realistic auxiliary variables at the second time occasion. Separate research can be done within this topic. Since we do not have any target variable in the simulation, results are only presented for auxiliary variables. As a multipurpose inventory, there are a wide range of target variables in the Swedish NFI, some examples can be the proportions of different land types, volume and number of trees per hectare, mean age, damages, amount of different berries, and so forth. Among them, most of the variables are related to the selected auxiliary variables.

The simulation results of the example is shown in Table 2 for the new strategy and the four reference strategies. According to Dobbie et al. (2008), the optimal panel design depends on the balance between needing to detect trend and report on the states, four different partition schemes are applied for Strategy 5 to test if there is an optimal way to split the two panels. The mean of spatial balance for samples at two time occasions, the percentage of overlap, the empirical variances of the estimator of the two states as well as change are listed for mean tree height and mean basal area.

We can clearly see that both state and change estimators obtained by employing the new strategy (Strategy 1) are better than what are possible by using reference strategies. There is no any optimal combination of the two panels for the current Swedish NFI strategy. The higher the proportions of the permanent panels, the worse the estimators will be. The main results can be summarized from two aspects: effect of using spatially balanced sampling designs and effect of the amount of overlap.

Strategies 1 and 4 that apply spatially balanced sampling designs produce the best estimators for states. Theoretically, the values of the estimators tend to be the same for the two strategies when the number of repetitions in the simulation is large enough. Strategy 2, which only spreads the sample in the geographical space, leads to the worst estimators for states. This confirms the importance of spreading the samples also in auxiliary variables other than the geographical space. Comparing Strategies 1 and 4 against Strategy 2, reduction of variances of the state estimators for both auxiliary variables is more than 92%, and the decrease in width of confidence intervals for both auxiliary variables is more than 70%. For Strategy 3, the quality of the state estimators reduced at the second time occasion, since the permanent sample is not as well spread anymore at that time point. For the strategy that applies split panel designs, the value of the variances for state estimators reduced as we increased the proportions of the well-spread panel.

Comparing Strategy 1 against Strategy 4, the reduction of variances for the estimators of change is more than 37%, and the decrease in width of confidence intervals is more than 20% for both auxiliary variables. This is because we get a much higher overlap (62%) between the samples selected at both time occasions using Strategy 1, which leads to higher covariances between the two state estimators compared with Strategy 4. For mean tree height, the estimate for covariance is 0.288 and −0.012 for Strategy 1 and Strategy 4, respectively. For mean basal area, the values are 0.006 and 8.21×10^{-5}, respectively. Based on Equation (4), when fixing the state variances, the strategy that has the larger covariance will

produce a smaller variance for the estimator of change. At first glance, results produced by Strategy 5 do not seem to be reasonable in terms of the estimators of change. It appears to be a contradictory statement that the variances of estimators of change increase as the percentages of overlap increase. However, by careful observations, we can find that the reason for the increased variances of change estimators is not because we do not have higher values of the covariances. For tree height, the covariance between two state estimators increased from 1.115 to 3.291 when we increased the overlap from 20% to 80%. For basal area, the covariance increased from 0.001 to 0.075. The reason of the increased variances of the change estimators are the increased values of variances for state estimators. According to Equation (4), when the increase in covariance cannot compensate the increase in the two variances of state estimators, the variance of estimator of change will still increase, even if we have a very high percentage of overlap.

5 | CONCLUSION AND DISCUSSION

We proposed a new sampling strategy with its main focus on monitoring totals of environmental variables. In practice, it is not restricted to monitor only the totals, it can also be applied for parameters such as quantiles (Grafström & Schelin, 2014). Positive coordination is studied by using SCPS within a continuous population framework. Based on an application, with settings similar to the Swedish NFI, we illustrated that the proposed new strategy performed better than all reference strategies. When matching the sample distribution of the auxiliary variables to the population distribution and at the same time use positively coordinated samples, we improve the precision for both the state and the change estimators.

If we use a sample, which is well spread only at the first time occasion as a permanent sample, then there is a big risk that the sample evolves differently from the population. The sample may become less balanced over time and, as a result, the state estimators also become less efficient over time. Although the sample overlap is 100% for a permanent sample, the estimator of change will also gradually become worse over time as the sample can differ more and more from the population in terms of the distributions of the auxiliary variables (and hence the target variables).

For the current Swedish NFI, two different types of clusters are used: permanent clusters and temporary clusters. In fact, the Swedish NFI design that combines temporary and permanent clusters has almost become an international standard toward which NFIs in other countries' aim (Fridman et al., 2014). The permanent clusters primarily aim to increase the accuracy of change estimation, and they are resurveyed regularly, whereas temporary ones are mainly intended to capture the current state of the forest and are only surveyed once (Tomppo, Gschwantner, Lawrence, & McRoberts, 2010, ch. 35). In the NFI example, we use positively coordinated and spatially balanced samples to target change and the current states of two time occasions simultaneously. Based on the simulation, we can see that the new strategy successfully improves the precision of the estimator for both state and change for the auxiliary variables. Thus, it has potential to also improve the precision of the estimators for the target variables that are related to the auxiliary variables. For those target variables that are not related to the selected auxiliary variables, by spreading the samples in the auxiliary variables, we get similar results as independent sample selections. Therefore, there is a potential to change the current design of the Swedish NFI. Instead of using the complex and less efficient combination of the permanent and the temporary samples, employing positively coordinated and well-spread samples can achieve both goals within a single sampling strategy.

A planner has plenty of options in choosing a sampling strategy. The main properties considered often include precision, unbiasedness, cost-efficiency, and simplicity to apply. As Scott (1984) mentioned, when estimating both state and change, a combination of remeasured (matched) plots, plots not remeasured (unmatched), and replacement (new) plots is generally the most cost-effective alternative. With the suggested strategy, at each point in time, the sample is a sample of the SCPS design with the prescribed inclusion probabilities. Thus, we make no compromise on the level of spatial balance of the different samples. Yet, we achieve a quite high degree of positive coordination. As the sample size can be varied over time, the strategy is also flexible for budget changes over time. With auxiliary variables available, they can be seen as proxies for the target variables. The sampling strategy that is superior for estimating the state and change for the auxiliary variables is likely to be superior for estimating the state and change of target variables related to those auxiliary variables.

Many spatially balance sampling designs can improve the state estimators compared with traditional designs (Benedetti et al., 2015, ch. 7). We focus only on SCPS among others, since it is efficient and easy to apply when it comes to positive sample coordination. The main reason of not including the GRTS or BAS designs as reference strategies in the application is because we have additional auxiliary variables beside the geographical coordinates, which are

not dimensions of the population. If we add an auxiliary variable such as elevation to the geographical coordinates, then we get a surface with zero three-dimensional volume. By enclosing the surface in a three-dimensional rectangular box and generate random points in the box, there is a zero probability to get points that lie on the surface. Thus, for example, BAS fails to use the additional information and can only spread the sample in the geographical coordinates.

The sample coordination method presented here is a probabilistic way for SCPS to define panels. This means by using this method we cannot fix the panels beforehand as we could do by using the traditional panel designs. The overlap between the successive samples at two time occasions depends on the change in the auxiliary variables between the two time occasions. Further studies on how to select well-spread samples with a prescribed percentage of overlap is of a great interest to us.

ACKNOWLEDGEMENTS

The authors are grateful to Kenneth Nyström for providing the growth models for the Swedish NFI example. They also thank two referees and an associate editor for valuable comments that improved the article.

REFERENCES

Artiola, J. F., Pepper, I. L., & Brusseau, M. L. (2004). *Environmental monitoring and characterization*. London, UK: Elsevier.

Awange, J. L. (2012). *Environmental Monitoring Using GNSS: Global navigation satellite systems*. Berlin/Heidelberg. Germany: Springer.

Benedetti, R., Piersimoni, F., & Postiglione, P. (2015). *Sampling spatial units for agricultural surveys*. Berlin/Heidelberg, Germany: Springer.

Bondesson, L., & Thorburn, D. (2008). A list sequential sampling method suitable for real-time sampling. *Scandinavian Journal of Statistics*, *35*, 466–483. https://doi.org/10.1111/j.1467-9469.2008.00596.x

Brewer, K. R. W., Early, L. J., & Joyce, S. F. (1972). Selecting several samples from a single population. *Australian Journal of Statistics*, *14*(3), 231–239. https://doi.org/10.1111/j.1467-842X.1972.tb00899.x

Cotton, F. & Hesse, C. (1992). Tirages coordonnés d'échantillons. Technical report E9206, Direction des Statistiques Économiques. Paris, France: INSEE.

Dehardt, J. (1971). Generalizations of the Glivenko-Cantelli theorem. *The Annals of Mathematical Statistics*, *46*(2), 2050–2055. https://doi.org/10.1214/aoms/1177693073

Dobbie, M. J., Henderson, B. L., & Stevens, D. L. (2008). Sparse sampling: Spatial design for monitoring stream networks. *Statistics Surveys*, *2*(2008), 113–153. https://doi.org/10.1214/07-SS032

Duncan, G. J., & Kalton, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, *55*(1), 97–117. https://doi.org/10.2307/1403273

Ernst, L. R. (1999). The maximization and minimization of sample overlap: A half century of results. *Bulletin of the International Statistical Institute*, *57*, 293–296.

Fridman, J., Holm, S., Nilsson, M., Ringvall, A. H., & Ståhl, G. (2014). Adapting national forest inventories to changing requirements – the case of the Swedish national forest inventory at the turn of the 20th century. *Silva Fennica*, *48*(3), 1–29. https://doi.org/10.14214/sf.1095

Grafström, A. (2012). Spatially correlated Poisson sampling. *Journal of Statistical Planning and Inference*, *142*(1), 139–147. https://doi.org/10.1016/j.jspi.2011.07.003

Grafström, A. & Lisic, J. (2019). Balanced sampling: Balanced sampling and spatially balanced sampling. *R package version 1.5.5*.

Grafström, A., Lundström, N. L. P., & Schelin, L. (2012). Spatially balanced bampling through the pivotal method. *Biometrics*, *68*(2), 514–520. https://doi.org/10.1111/j.1541-0420.2011.01699.x

Grafström, A., & Matei, A. (2015). Coordination of conditional poisson samples. *Journal of Official Statistics*, *31*(4), 649–672. https://doi.org/10.1515/jos-2015-0039

Grafström, A., & Matei, A. (2018a). Spatially balanced sampling of continuous populations. *Scandinavian Journal of Statistics*, *45*(3), 792–805. https://doi.org/10.1111/sjos.12322

Grafström, A. & Matei, A. (2018b). Coordination of spatially balanced samples. *Survey Methodology, 44*(2), 215-238 (Statistics Canada, Catalogue No. 12-001-X).

Grafström, A., & Schelin, L. (2014). How to select representative samples. *Scandinavian Journal of Statistics*, *41*(2), 277–290. https://doi.org/10.1111/sjos.12016

Grafström, A., Schnell, S., Saarela, S., Hubbell, S. P., & Condit, R. (2017). The continuous population approach to forest inventories and use of information in the design. *Environmetrics*, *28*(8), 1–12. https://doi.org/10.1002/env.2480

Grafström, A., Zhao, X., Nylander, M., & Petersson, H. (2017). A new sampling strategy for forest inventories applied to the temporary clusters of the Swedish national forest inventory. *Canadian Journal of Forest Research*, *47*(9), 1161–1167. https://doi.org/10.1139/cjfr-2017-0095

Keyfitz, N. (1951). Sampling with probabilities proportional to size: Adjustment for changes in the probabilities. *Journal of the American Statistical Association*, *46*(253), 105–109. https://doi.org/10.1080/01621459.1951.10500773

Kish, L. (1983). *Data collection for details over space and time*. In T. Wright (Ed.), *Statistical methods and the improvement of data quality* (pp. 73–84). Cambridge, MA: Academic Press.

Kish, L. (1986). Timing of surveys for public policy. *Australian Journal of Statistics*, *28*(1), 1–12. https://doi.org/10.1111/j.1467-842X.1986.tb00579.x

Kish, L., & Scott, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, *66*(335), 461–470. https://doi.org/10.2307/2283509

Mach, L., Reiss, P. T., & Şchiopu-Kratina, I. (2006). Optimizing the expected overlap of survey samples via the northwest corner rule. *Journal of the American Statistical Association*, *101*(476), 1671–1679. https://doi.org/10.1198/016214506000000320

Mandallaz, D. (2007). *Sampling techniques for forest inventories*. Boca Raton, FL: CRC Press.

Marker, D. A., & Stevens, D. L., Jr. (2009). *Sampling and inference in environmental surveys*. In D. Pfeffermann & C. R. Rao (Eds.), *Handbook of statistics 29A: Sample surveys: Design, methods and applications* (pp. 487–512). Cambridge, MA: Elsevier.

McDonald, T. L. (2003). Review of environmental monitoring methods: Survey designs. *Environmental Monitoring and Assessment*, *85*(3), 277–292. https://doi.org/10.1023/A:1023954311636

Patterson, H. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Series B (Methodological)*, *12*(2), 241–255. https://doi.org/10.1111/j.2517-6161.1950.tb00058.x

Qualité, L., & Tillé, Y. (2008). Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added. *Survey Methodology*, *34*(2), 173–181.

Robertson, B., Brown, J. A., McDonald, T., & Jaksons, P. (2013). BAS: Balanced acceptance sampling of natural resources. *Biometrics*, *69*(3), 776–784. https://doi.org/10.1111/biom.12059

Robertson, B., McDonald, T., Price, C., & Brown, J. (2018). Halton iterative partitioning: Spatially balanced sampling via partitioning. *Environmental and Ecological Statistics*, *25*(3), 305–323. https://doi.org/10.1007/s10651-018-0406-6

Scott, C. T. (1984). A new look at sampling with partial replacement. *Forest Science*, *30*(1), 157–166.

Scott, C. T. (1998). Sampling method for estimating change in forest resources. *Ecological Applications*, *8*(2), 228–233. https://doi.org/10.2307/2641062

Sen, A. R. (1973). Theory and application of sampling on repeated occasions with several auxiliary variables. *Biometrics*, *29*(2), 381–385. https://doi.org/10.2307/2529401

Stevens, D. L., & Olsen, A. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, *99*(465), 262–278. https://doi.org/10.1198/016214504000000250

Tomppo, E., Gschwantner, T., Lawrence, M., & McRoberts, R. E. (2010). *National forest inventories pathways for common reporting*. Dordrecht, Netherlands: Springer.

Wolfowitz, J. (1954). Generalization of the theorem of Glivenko-Cantelli. *The Annals of Mathematical Statistics*, *25*(1), 131–138.

How to cite this article: Zhao X, Grafström A. A sample coordination method to monitor totals of environmental variables. *Environmetrics*. 2020;31:e2625. https://doi.org/10.1002/env.2625

III

Estimation of change with partially overlapping and spatially balanced samples

Xin Zhao [*1] and Anton Grafström[1]

[1] *Department of Forest Resource Management, Swedish University of Agriculture Sciences*

Abstract

Spatially balanced samples are samples that are well-spread in some available auxiliary variables. Selecting such samples has been proven to be very efficient in estimation of the current state (total or mean) of target variables related to the auxiliary variables. As time goes, or when new auxiliary variables become available, such samples need to be updated to stay well-spread and produce good estimates of the current state. In such an update, we want to keep some overlap between successive samples to improve the estimation of change. With this approach, we end up with partially overlapping and spatially balanced samples. To estimate the variance of an estimator of change, we need to be able to estimate the covariance between successive estimators of the current state. We introduce an approximate estimator of such covariance based on local means. By examples, we show that it can be applied in the estimation of the variance of an estimator of change based on partially overlapping and spatially balanced samples.

Key words: Spatially correlated Poisson sampling, well-spread samples, overlapping samples, repeated surveys.

*Email: xin.zhao@slu.se

1 Introduction

In repeated surveys, a common focus is to monitor the change between population totals over time. The estimation of the variance of an estimator of change is essential to judge whether the observed change is statistically significant. It is well known that, when estimating the variance of an estimator of change, we need to estimate the variance of the two state estimators as well as the covariance between them. To reduce the variance of the estimator of change, we can either make the variance of the two state estimators smaller or attempt to create a high positive covariance between the two state estimators or try both of them. The question of whether we should use independent samples, a permanent sample or partially overlapping samples over time arises.

For independent samples, we do not need to consider the covariance. Then, the variance of change depends only on the variance of the estimators at each time occasion. This simplifies the estimation problem. However, it will not be the best strategy to use independent samples when estimating changes over time. This is because the variance of the change estimator becomes about twice the variance of the estimator of the state when using independent samples. When the time between surveys is short and the values of the target variables have not changed much, a permanent sample might be employed to reduce the variance of an estimator of change. However, as the population changes over time, the permanent sample will not be as representative as it used to be at the following time occasion. If the sample changes in a different way than the population, which is out of our control, then there is a risk of a much larger variance of the state estimator at the following time occasion. Thus, even if the covariance between the two state estimators becomes large by having fully overlapping samples, it is not guaranteed that the variance of change will be reduced. There is a need for updating the sample at the next time occasion to account for changes while retaining as many units as possible from the old sample.

For environmental surveys, the units to be observed often have some spatial features which are represented by a set of auxiliary variables. In general, nearby units in the space spanned by the auxiliary variables tend to have more similar values than units that are farther apart. Especially, this is true if the auxiliary variables have some explanatory power for the target variables in the survey. It is well known that we should incorporate the geographical locations of the populations when selecting samples, see, e.g., Stevens & Olsen 2003; Grafström 2012. Spatially balanced samples are more efficient than, for instance, samples selected by simple random sampling, e.g. Stevens & Olsen (2004). Recently, Zhao & Grafström (2020) illustrated further that it is efficient to use spatially balanced and partially overlapping samples for monitoring the change of environmental variables. By employing spatially balanced samples, we can reduce the variance of the state estimators. When using also positively coordinated samples, we can reduce the variance of the change estimator. In Zhao & Grafström (2020), the advantages of this strategy were verified by comparison with several other strategies. However, the problem of estimating the variance of the estimator of change under the proposed strategy was left unsolved.

A large number of variance estimators (approximations) have been proposed under different sampling designs (e.g., Horvitz & Thompson 1952; Yates & Grundy 1953; Hartley &

Rao 1962; Hájek 1964; Berger 2004). For repeated surveys, researchers have also paid a lot of attention to the estimation of covariance. Tam (1985) was one of the earliest studies that considered covariance estimations from overlapping samples. Qualité (2009, ch.5) derived covariance estimators based on two overlapping samples by considering sampling designs that are essentially applicable to obtain rotating panels, i.e., panels where only a part of the sample at a previous time occasion is maintained, and the rest of the units in the sample are replaced by new units at a next time occasion.

As a result of better spread of the samples when employing spatially balanced sampling designs, it may not be optimal to apply conservative variance or covariance estimators, like the ones for simple random sampling. Grafström & Schelin (2014) introduced a local mean variance estimator under spatially balanced sampling designs. Instead of using the global mean, a local mean is adopted in the expression. In the variance estimator, only the nearest neighbours of a sample unit and the unit itself will be included in the computation of the local mean. The authors also considered the case that there will be equal distances between units. Therefore, the local neighbourhood size of each sample unit will not be fixed in the variance estimator. We modify this variance estimator by introducing a fixed size of the local neighbourhoods of all sample units. Starting from the settings in Qualité (2009, ch.5), we also derive a local mean covariance estimator and obtain a variance estimator for the estimator of change. By simulation, we illustrate that the proposed local mean estimators are stable and less biased compared to the estimators that do not employ local means. Therefore, the local mean variance and covariance estimators can be applied when estimating the variance of the estimator of change with partially overlapping and spatially balanced samples.

The rest of the paper is structured as follows. We begin with notations for estimating change with general designs in Section 2. In Section 3, we introduce an efficient sampling strategy for monitoring the change of environmental variables. In Section 4, starting from a local mean variance estimator, we derive a local mean covariance estimator for partially overlapping and spatially balanced samples. In Section 5 , two examples are considered to evaluate the estimators. Finally, Section 6 is dedicated to discussion and comments.

2 Estimation of change with general designs

Suppose we have a shared list frame $U = \{1, ..., i, ..., N\}$ over time. From U, a sample S_t can be selected at time t with a sample size n_t. Denote the target variable for unit i at time t as y_{it}. The total can be expressed as $Y_t = \sum_{i \in U} y_{it}$. Let $\pi_{it} = \Pr(i \in S_t)$ be the prescribed inclusion probability of unit i at time t. The Horvitz-Thompson (HT) estimator of Y_t can be expressed as

$$\widehat{Y}_t = \sum_{i \in S_t} \frac{y_{it}}{\pi_{it}}. \tag{1}$$

Our goal is to estimate the change of the population total between two occasions $\Delta = Y_2 - Y_1$ by using $\widehat{\Delta} = \widehat{Y}_2 - \widehat{Y}_1$. To know the precision of the estimation, we also need to

estimate the variance of the estimator of change. This variance is given by

$$V(\widehat{\Delta}) = V(\widehat{Y}_1) + V(\widehat{Y}_2) - 2C(\widehat{Y}_1, \widehat{Y}_2). \tag{2}$$

This means we need to estimate the variance of the separate state estimators and the covariance between the two estimators. The variance of the state estimator (1) can be expressed as

$$V(\widehat{Y}_t) = \sum_{i \in U} \sum_{j \in U} (\pi_{itj} - \pi_{it}\pi_{jt}) \frac{y_{it}}{\pi_{it}} \frac{y_{jt}}{\pi_{jt}}, \tag{3}$$

where $\pi_{ijt} = \Pr(i \in S_t, j \in S_t)$ is the second-order inclusion probability for a pair of points (i, j) at time t. An estimator of $V(\widehat{Y}_t)$ is

$$\widehat{V}(\widehat{Y}_t) = \sum_{i \in S_t} \sum_{j \in S_t} \frac{(\pi_{ijt} - \pi_{it}\pi_{jt})}{\pi_{ijt}} \frac{y_{it}}{\pi_{it}} \frac{y_{jt}}{\pi_{jt}}. \tag{4}$$

Estimator (4) is unbiased for (3) if all second-order inclusion probabilities π_{ijt} are strictly positive.

The covariance between two HT-estimators of two population totals can be expressed as

$$C(\widehat{Y}_1, \widehat{Y}_2) = \sum_{i \in U} \sum_{j \in U} \left(\pi_{ij}^{12} - \pi_{i1}\pi_{j2}\right) \frac{y_{i1}}{\pi_{i1}} \frac{y_{j2}}{\pi_{j2}}, \tag{5}$$

where $\pi_{ij}^{12} = \Pr(i \in S_1, j \in S_2)$. It is also possible to construct the HT-estimator of (5) based on the two samples, i.e.

$$\widehat{C}(\widehat{Y}_1, \widehat{Y}_2) = \sum_{i \in S_1} \sum_{j \in S_2} \frac{\left(\pi_{ij}^{12} - \pi_{i1}\pi_{j2}\right)}{\pi_{ij}^{12}} \frac{y_{i1}}{\pi_{i1}} \frac{y_{j2}}{\pi_{j2}}. \tag{6}$$

Similar to the variance estimator (4), the estimator (6) is unbiased for (5) provided the π_{ij}^{12} are strictly positive for all i, j. By employing (4) and (6) we can easily obtain the estimator of the variance for the estimator of change, provided that we have known positive second order probabilities.

3 An efficient sampling strategy to monitor the change of environmental variables

In environmental surveys, the spatial pattern of units is important, because the units themselves are defined using spatial criteria. To achieve good estimates of population characteristics, the spatial pattern of the sample should be similar to the spatial pattern of the population. Often, we do not know the spatial pattern of the target variable before the sample is selected. Instead, we have full access to some auxiliary variables that are related to the target variables. Stevens & Olsen (2004) introduced the generalized random tessellation

stratified (GRTS) design and coined the phrase "spatially balanced sampling". They also proposed a statistic that measures the spatial balance of a sample using Voronoi polygons. The local pivotal method (LPM) and spatially correlated Poisson sampling (SCPS) proposed by Grafström *et al.* (2012) and Grafström (2012) are two spatially balanced sampling designs that employ auxiliary variables (often including geographical coordinates plus several other attribute variables) to spread the samples based on distances. Grafström & Lundström (2013) illustrated that when the target variables are smooth functions of auxiliary variables, it is sufficient to spread the samples in the auxiliary variables. Because we get well spread and balanced samples by spreading in such auxiliary variables, we thus improve the precision of the estimators.

In spatially balanced sampling designs, auxiliary variables which are related to the target variables are often applied to spread the samples. There is a general assumption when using these designs, that nearby units are more similar than units that are farther apart. Intuitively, more information could then be obtained if the random sample avoids the selection of nearby units. The distance is measured over the auxiliary variables. In other words, the samples should be well spread over the auxiliary variables. It has been confirmed that, by using these designs, we gain in efficiency of design–based estimators of the totals of target variables (see e.g., Benedetti *et al.*, 2017). Regarding the monitoring of change, we need to be cautious about the determination of whether a sampling strategy is an efficient strategy or not. As we can see from (2), the variance of the estimator of change equals the sum of the separate variances of the state estimators minus two times the covariance between them. Therefore, to get a smaller variance of the estimator of change we can either reduce the variance of each state estimator or produce a high covariance between them, given the same variance of each estimator, or aim for both of them.

Zhao & Grafström (2020) proposed an efficient sampling strategy for monitoring the change of environmental variables. In this strategy, the concept of spatially balanced samples and positive sample coordination are combined. The spatially balanced samples are selected by the SCPS. When applying the SCPS, a set of auxiliary variables that are related to the target variables should be used to spread the sample. We choose the same set of auxiliary variables (with different values) at different time occasions. The positive sample coordination is achieved by assigning the same random number to each sample unit in the algorithm of SCPS. In this way, we will get partially overlapping and spatially balanced samples. Figure 1 illustrates two such samples selected by SCPS.

By using this strategy, we can reduce the variance of the state estimators and often achieve a large covariance between the state estimators at the same time. In Zhao & Grafström (2020), the empirical impact of using positively coordinated and spatially balanced samples was studied. In the next section, we will focus on the estimation problem and will provide a reasonable variance estimator for the estimator of the change under the sampling strategy.

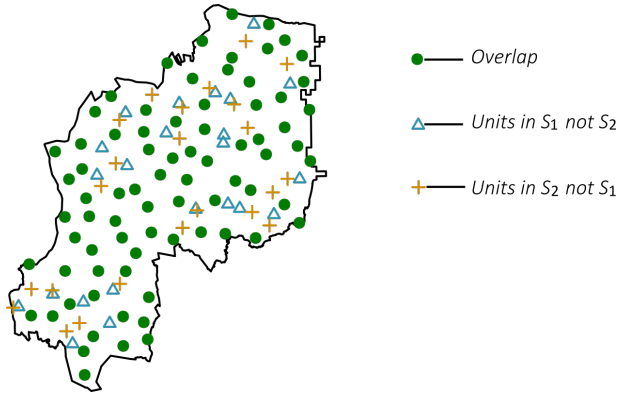


Figure 1: Illustration of two samples selected by the strategy

4 Estimation of change when samples are overlapping and well spread

Under a spatially balanced sampling design, it is often difficult to obtain π_{ijt} and π_{ij}^{12}. Moreover, many second-order inclusion probabilities may be zero. It will likely not be possible to use design-based unbiased variance estimators such as (4) and (6) under spatially balanced sampling designs. Even if it is possible, it will generally not be recommended as such variance estimators can become highly unstable when some second-order inclusion probabilities are very small.

4.1 Variance estimators for spatially balanced samples

Matérn (1947) introduced a variance estimator for systematic sampling from a regular grid of sample locations. In Matérn's variance estimator, the sample locations are split into several nonoverlapping groups of neighbours. A local variance is first constructed for each group, then an average over groups is calculated as the variance estimator. Motivated by this estimator, Grafström & Schelin (2014) also proposed a local mean variance estimator which was shown to perform well under spatially balanced sampling. In their variance estimator, the local neighbourhood for each sample unit i depends only on i and its nearest neighbours. The size of the local neighbourhood depends on the number of nearest neighbours of each unit in the sample. Their variance estimator can be applied in situations where units have many nearest neighbours. For well-spread environmental samples, it is rare for a sample unit to have equidistant neighbours. In environmental sampling, we often spread the samples geographically. Moreover, the geographical coordinates are different for each unit. Therefore, we usually get unique distances between sample units. Hence, there is no need to consider the case of equal distances in the local neighbourhood in the variance estimator for such samples. In principle, by only including the unit i and its nearest neighbour in the local

neighbourhood, we often have two units in the local neighbourhood when estimating the variance with well-spread samples.

In Stevens & Olsen (2003), the authors recommended using four sample units in the local neighbourhood. This is because they found that their local mean variance estimator became unstable when including fewer sample units in the local neighbourhood. We consider their suggestion and modify the local mean variance estimator in Grafström & Schelin (2014) by using a neighbourhood size proportional to the sample size. For $V\left(\widehat{Y}_t\right)$, the local mean variance estimator can be expressed as

$$\widehat{V}_{SB}(\hat{Y}_t) = \frac{n_{lt}}{n_{lt}-1}\sum_{i\in S_t}\left(\frac{y_{it}}{\pi_{it}} - \frac{1}{n_{lt}}\sum_{j\in S_{it}}\frac{y_{jt}}{\pi_{jt}}\right)^2, \tag{7}$$

where $S_{it} \subseteq S_t$ is the local neighbourhood of a sample unit i at time t. The neighbourhood S_{it} contains the unit i as well as its nearby units in the sample, the size n_{lt} is equal to $p_l n_t$ (rounded to the nearest integer). The proportion p_l can be chosen such that n_{lt} can be any integer between two and n_t. The same proportion p_l is suggested in estimation of the variance of $\widehat{Y}_t$ for all t. Then, for a fixed p_l, the number of neighbours included in the local neighbourhood depends only on the sample size.

Suppose all sample units are independently selected with the same set of drawing probabilities $p_i > 0$, $i = 1, 2, ..., N$, with $\sum_{i=1}^{N} p_i = 1$. For a sample S_t with sample size n_t, the expected number of inclusions of unit i is then $\pi_{it} = n_t p_i$. When enlarging the local neighbourhood to the full sample, i.e., if $S_{it} = S_t$, (7) becomes

$$\widehat{V}(\hat{Y}_t) = \frac{1}{n_t(n_t-1)}\sum_{i\in S_t}\left(\frac{y_{it}}{p_i} - \frac{1}{n_t}\sum_{i\in S_t}\frac{y_{it}}{p_i}\right)^2. \tag{8}$$

The estimator (8) corresponds to the unbiased variance estimator under the probability proportional to size (*pps*) sampling design. Furthermore, if we apply a constant inclusion probability $\pi_{it} = n_t/N$, we get

$$\widehat{V}(\hat{Y}_t) = \sum_{i\in S_t}\frac{N^2}{n_t(n_t-1)}\left(y_{it} - \frac{1}{n_t}\sum_{i\in S_t} y_{it}\right)^2 = \frac{N^2}{n_t}S_t^2, \tag{9}$$

where $S_t^2 = (n_t-1)^{-1}\sum_{i\in S_t}\left(y_{it} - n_t^{-1}\sum_{i\in S_t} y_{it}\right)^2$. Equation (9) is equivalent to the unbiased variance estimator under simple random sampling with replacement (SIR) design.

4.2 Covariance estimator for partially overlapping and spatially balanced samples

As illustrated in Section 4.1, the variance estimator (7) is a local mean version of the variance estimator for sampling with independent observations. In the case of overlapping samples, we

can introduce also a local mean version of an estimator of the covariance. As a starting point, we introduce the setting with independent observations. Let $p_i > 0$, $i = 1, 2, ..., N$, with $\sum_{i=1}^{N} p_i = 1$ be the drawing probabilities for units in U. First n_1 independent observations are drawn from U to S_1, and a subsample S_{12} of S_1 is retained as a part of S_2, with $n_{12} \geq 2$ observations. Next, an additional number of $n_2 - n_{12}$ independent observations are drawn from U to S_2 according to the drawing probabilities p_i, $i = 1, 2, .., N$. Now, the two samples S_1 and S_2 share n_{12} observations in the sample S_{12}. In this setting, we estimate the total $Y_t = \sum_{i \in U} y_{it}$ with $\widehat{Y}_t = \sum_{i \in S_t} y_{it} n_t^{-1} p_i^{-1}$ for $t = 1, 2$. Moreover, the covariance between $\widehat{Y}_1$ and $\widehat{Y}_2$ is

$$C\left(\widehat{Y}_1, \widehat{Y}_2\right) = n_{12} \sum_{i \in U} p_i \left(\frac{y_{i1}}{n_1 p_i} - \frac{Y_1}{n_1}\right)\left(\frac{y_{i2}}{n_2 p_i} - \frac{Y_2}{n_2}\right). \tag{10}$$

The covariance (10) can be estimated using S_{12} by the simple expansion (see e.g. Qualité, 2009, ch.5)

$$\widehat{C}\left(\widehat{Y}_1, \widehat{Y}_2\right) = \frac{n_{12}}{n_{12} - 1} \sum_{i \in S_{12}} \left(\frac{y_{i1}}{n_1 p_i} - \frac{\widehat{Y}_1'}{n_1}\right)\left(\frac{y_{i2}}{n_2 p_i} - \frac{\widehat{Y}_2'}{n_2}\right), \tag{11}$$

where $\widehat{Y}_t' = \sum_{i \in S_{12}} y_{it} n_{12}^{-1} p_i^{-1}$ is the estimator of Y_t based on the sample S_{12}. Even though $\widehat{Y}_t'$ is not the best estimator of Y_t as it only uses information of the shared observations in S_{12}, it is recommended. Using information outside of S_{12} can lead to undesired effects and is for that reason considered bad practice, see Qualité (2009, ch.5).

In the case of two overlapping and spatially balanced samples, we replace the expected number of inclusions $n_t p_i$ with the inclusion probabilities π_{it} and introduce local means. The estimator (11) then becomes

$$\widehat{C}_{SB}\left(\widehat{Y}_1, \widehat{Y}_2\right) = \frac{n_{l12}}{n_{l12} - 1} \sum_{i \in S_{12}} \left(\frac{y_{i1}}{\pi_{i1}} - \overline{y}_{i1}\right)\left(\frac{y_{i2}}{\pi_{i2}} - \overline{y}_{i2}\right), \tag{12}$$

where $\overline{y}_{i1} = n_{l12}^{-1} \sum_{j \in S_{i1}} y_{j1} \pi_{j1}^{-1}$, $\overline{y}_{i2} = n_{l12}^{-1} \sum_{j \in S_{i2}} y_{j2} \pi_{j2}^{-1}$ and S_{it} is the local neighbourhood for unit i in S_{12} at time t. The neighbourhood size n_{l12} is chosen as $p_l n_{12}$ (rounded to the nearest integer) and the same proportion p_l as in the local mean variance estimator is recommended. Since $n_{12} \leq n_t$, it is reasonable to decide the proportion by the size of the overlap when estimating the variance of the estimator of change, i.e, $p_l = n_{l12} n_{12}^{-1}$, where n_{l12} can be any integer between two and n_{12}. Then we make sure that $n_{lt} = \text{round}(p_l n_t) \geq n_{l12}$. If the size $n_{l12} = n_{12}$, we get back to the estimator (11). The estimator (12) of covariance under spatially balanced sampling is consistent with the estimator $\widehat{V}_{SB}\left(\widehat{Y}_t\right)$ of variance, i.e. $\widehat{C}_{SB}\left(\widehat{Y}_t, \widehat{Y}_t\right) = \widehat{V}_{SB}\left(\widehat{Y}_t\right)$. This is important for estimating the variance of an estimator of change. Combining (7) and (12), the expression of the variance estimator for the estimator of change with partially overlapping and spatially balanced samples follows.

Different from the sampling plans in Qualité (2009, ch.5), the size of overlap is random when using the strategy in Zhao & Grafström (2020). For the current algorithm in the

strategy, it is not possible to fix the size of the overlap and select a well-spread sample at a second time occasion. That is to say, we do not know how many sample units from S_1 that will also be selected into S_2 before we get the full sample on the second time occasion. The percentage of overlap between two samples depends mainly on the change over time of the auxiliary variables that we use to spread the samples. Similar to the variance estimator (7), the covariance estimator (12) is proposed as a general estimator for spatially balanced samples. In the next section, we study the performance of the proposed variance and covariance estimators, specifically under the strategy in Zhao & Grafström (2020).

5 Evaluation of the estimators

To evaluate the proposed estimators for positively coordinated and spatially balanced samples, two examples are considered. In the examples, we take different sizes of neighbourhoods into account to check how they will affect the estimators. Estimators which apply the full samples/overlap in the neighbourhoods are incorporated in the simulations as well. For each example, the empirical variance and covariance, the mean of the variance and covariance estimators are presented. We calculate the mean coverage rates for the 95% confidence intervals when using the variance estimators. The relative bias(RB) as well as the empirical relative root mean square error (RRMSE) for the estimators are also compared for different estimators. It is worth noting that, the samples selected at the two time occasions are well spread and positively coordinated in both examples.

Example 1. We use a surface to define the target variable at each time occasion. In the simulations, the samples are spread in geographical coordinates. We set the population size $N = 2500$, sample size at the first time occasion is $n_1 = 100$, a smaller sample size with $n_2 = 50$ is used at time 2. The number of repetitions is 10000, and equal inclusion probabilities $\pi_{it} = n_t/N$ are applied at each time. The surfaces are displayed in Figure 2. The simulation results are listed in Table 1 for estimators that use neighbourhoods of different sizes.

Example 2. In this example, the same data set as in Zhao & Grafström (2020) is applied to evaluate the estimators. It is an application of the Swedish national forest inventory. In Zhao & Grafström (2020), five different auxiliary variables (geographical coordinates, elevation, tree height, and basal area) were employed to spread the samples and the performance of the strategy was only evaluated for the auxiliary variables. Because tree height and basal area are strongly correlated, we here choose the basal area as our target variable and spread the samples using the rest of the variables to learn the performance of the proposed estimators for a possible target variable. In the simulation, we have a population size of $N = 10000$, sample size $n_1 = n_2 = 100$, equal inclusion probabilities $\pi_t = n_t/N = 0.01$ and the number of repetitions is 10000. The results are illustrated for the basal area in Table 2.

Simulation results of both examples are also illustrated in Figure 3 for all estimators. From the figure and the tables, we can see that all estimators are generally conservative.

Table 1: Simulation results of Example 1. The correlation coefficient between the target variables at the two time occasions is 0.9550. The total of the target is 72861.65 at time 1 and 104398.8 at time 2. The mean of percentage of overlap $2E\left(n_{12}\right)/\left(n_1+n_2\right)$ between samples at the two time occasions is 43.35%.

		y			
		$n_{l12}=2$	$n_{l12}=4$	$n_{l12}=6$	$n_{l12}=n_{12}$
Empirical	$V\left(\widehat{Y}_1\right)$		5603276		
	$V\left(\widehat{Y}_2\right)$		24663288		
	$C\left(\widehat{Y}_1,\widehat{Y}2\right)$		2348328		
	$V(\hat{\Delta})$		25563549		
Estimated	$\overline{\widehat{V}}_{SB}\left(\widehat{Y}_1\right)$	18307756 (0.999)	29398474 (1)	34905003 (1)	51394294 (1)
	$\overline{\widehat{V}}_{SB}\left(\widehat{Y}_2\right)$	59566510 (0.995)	80948404 (0.999)	95541122 (0.999)	150657134 (1)
	$\overline{\widehat{C}}_{SB}\left(\widehat{Y}_1,\widehat{Y}_2\right)$	19688631	22676798	25939517	38593385
	$\overline{\widehat{V}}_{SB}(\hat{\Delta})$	38497004 (0.960)	64993281 (0.996)	78567092 (0.999)	124864658 (1)
Relative bias	$RB_{\overline{\widehat{V}}_{SB}(\widehat{Y}_1)}$	2.267	4.247	5.229	8.172
	$RB_{\overline{\widehat{V}}_{SB}(\widehat{Y}_2)}$	1.415	2.282	2.874	5.109
	$RB_{\overline{\widehat{C}}_{SB}(\widehat{Y}_1,\widehat{Y}_2)}$	7.384	8.657	10.046	15.434
	$RB_{\overline{\widehat{V}}_{SB}(\hat{\Delta})}$	0.506	1.542	2.073	3.885
RRMSE	$RRMSE_{\overline{\widehat{V}}_{SB}(\widehat{Y}_1)}$	2.308	4.290	5.270	8.213
	$RRMSE_{\overline{\widehat{V}}_{SB}(\widehat{Y}_2)}$	1.501	2.350	2.935	5.171
	$RRMSE_{\overline{\widehat{C}}_{SB}(\widehat{Y}_1,\widehat{Y}_2)}$	7.964	9.140	10.535	16.042
	$RRMSE_{\overline{\widehat{V}}_{SB}(\hat{\Delta})}$	0.782	1.652	2.162	3.963

Table 2: Simulation results of Example 2. For basal area, the total is 137487.4 m^2/ha and 146575.3 m^2/ha, respectively for the two time occasions. Correlation coefficient between basal area at time 1 and 2 is 0.9225 and for tree height it is 0.9201. The correlation coefficient between basal area and tree height is 0.9494 and 0.9568 respectively at the two time occasions. The mean of the overlap is 64.05%.

		y_b			
		$n_{l12}=2$	$n_{l12}=4$	$n_{l12}=6$	$n_{l12}=n_{12}$
Empirical	$V\left(\widehat{Y}_1\right)$		4450984		
	$V\left(\widehat{Y}_2\right)$		4117761		
	$C\left(\widehat{Y}_1,\widehat{Y}2\right)$		1973211		
	$V(\widehat{\Delta})$		4622964		
Estimated	$\overline{\widehat{V}}_{SB}\left(\widehat{Y}_1\right)$	5371933 (0.965)	6443658 (0.980)	8161177 (0.991)	34256433 (1)
	$\overline{\widehat{V}}_{SB}\left(\widehat{Y}_2\right)$	5131544 (0.970)	6209523 (0.983)	7935950 (0.992)	34776755 (1)
	$\overline{\widehat{C}}_{SB}\left(\widehat{Y}_1,\widehat{Y}_2\right)$	2152611	3225581	4338479	23229975
	$\overline{\widehat{V}}_{SB}(\widehat{\Delta})$	6198256 (0.975)	6202019 (0.974)	7420169 (0.986)	22573238 (1)
Relative bias	$RB_{\overline{\widehat{V}}_{SB}\left(\widehat{Y}_1\right)}$	0.207	0.448	0.834	6.696
	$RB_{\overline{\widehat{V}}_{SB}\left(\widehat{Y}_2\right)}$	0.246	0.508	0.927	7.446
	$RB_{\overline{\widehat{C}}_{SB}\left(\widehat{Y}_1,\widehat{Y}_2\right)}$	0.091	0.635	1.199	10.773
	$RB_{\overline{\widehat{V}}_{SB}(\widehat{\Delta})}$	0.341	0.342	0.605	3.883
RRMSE	$RRMSE_{\overline{\widehat{V}}_{SB}\left(\widehat{Y}_1\right)}$	0.278	0.486	0.859	6.716
	$RRMSE_{\overline{\widehat{V}}_{SB}\left(\widehat{Y}_2\right)}$	0.312	0.544	0.952	7.465
	$RRMSE_{\overline{\widehat{C}}_{SB}\left(\widehat{Y}_1,\widehat{Y}_2\right)}$	0.334	0.718	1.255	10.858
	$RRMSE_{\overline{\widehat{V}}_{SB}(\widehat{\Delta})}$	0.437	0.432	0.676	3.985

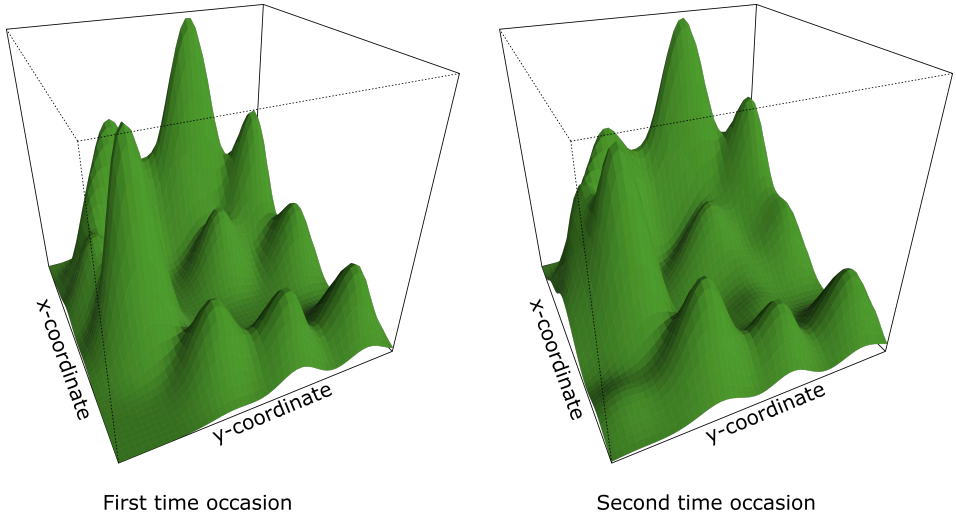


Figure 2: Target surfaces in Example 1

Comparing the estimators which apply the full samples/overlap in the neighbourhood, we can reduce the bias by using local neighbourhood estimators. The smaller the neighbourhood size, the less biased the estimator tends to be. The coverage rate of the confidence intervals also increases as the neighbourhood size grows. Note that, in each iteration we can only fix the neighbourhood size of the local mean covariance estimator. The size of the neighbourhood of the local mean variance estimator varies according to the size of the overlap. If the aim is to estimate the variance of the estimator of the total at each time occasion, we can fix the neighbourhood size directly.

6 Discussion

Thus far, we have evaluated the proposed estimators by using well-spread samples and equal inclusion probabilities (representative samples). The use of equal probabilities is, however, the most common case in multipurpose environmental surveys. As the strength of the relation between different target variables and the auxiliary variables that we use to spread the samples are not the same for different target variables, it is safer to spread the samples with equal inclusion probabilities.

More examples have been investigated to verify the performance of the estimators. The conclusions we get from other examples are in accordance with the two examples that are presented in the manuscript. We find that if we apply the same number of units to both the variance and covariance estimators, the overestimation by the local mean covariance estimator will become bigger than the overestimation by the local variance estimators. In such a case, it may produce a negative bias for the variance of the estimator of change. This

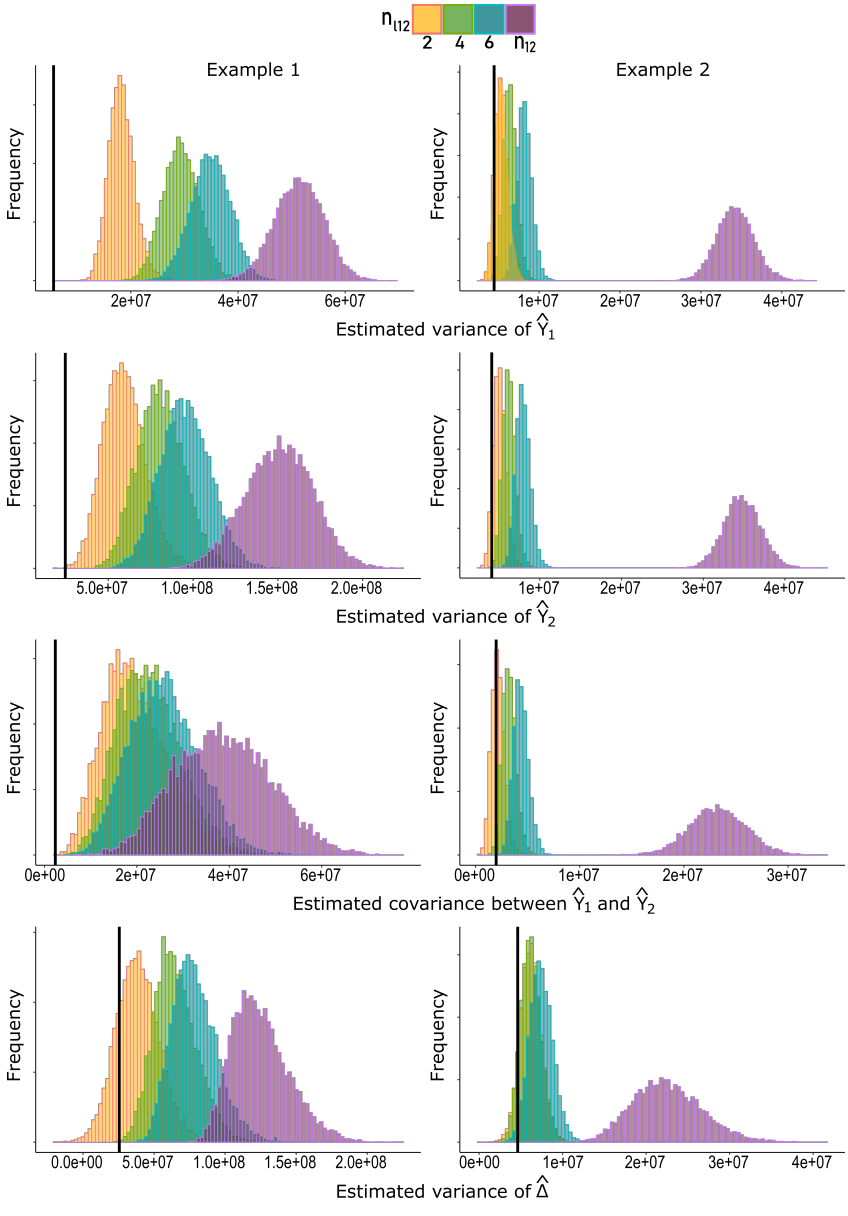


Figure 3: Comparing the estimators for different neighbourhood sizes. The bold black vertical lines represent the empirical variances/covariances.

is because the local covariance estimator is based only on the overlap, and the neighbours tend to have larger distances, thereby causing bigger differences in the overlap than in the full sample. Therefore, the more the neighbours in the neighbourhood, the bigger the difference between the value of unit i and its local mean will be, thus the larger positive bias it will produce. By the method we proposed, fewer neighbours are used in the local mean covariance estimator than the separate variance estimators. Therefore, we reduce the impact of the distance in the estimation of the variance of the estimator of change.

Besides the distance, the performances of the local mean covariance estimators are also affected by the rate of overlap. The bias tends to become bigger for a small percentage of the overlap. We need to notice that, for repeated surveys that are carried out with more tight time intervals, permanent samples are likely to be better. Especially when we only want to reduce the variance of the estimator of change in the short run. In that case, the best strategy is probably to use a permanently well-spread sample (the sample S_1 is well-spread in the first survey, thereafter the same sample will be applied in the second survey). At short intervals, if S_2 is only partially overlapping with S_1, it will lead to a smaller covariance compared to a permanent sample. Although we reduce the variance of the state at the second time occasion by updating the sample, the reduction of the variance may not compensate for the reduction of the covariance. In the long run, it will be preferable to apply the new strategy, because the quality of S_1 is likely to become worse over time. The reduction of the variance will then compensate for the reduction of the covariance compared to a permanent sample. Thus, the planner needs to be aware of these trade-offs when dealing with complex surveys.

References

Benedetti, R., Piersimoni, F. & Postiglione, P. (2017). Spatially balanced sampling: A review and a reappraisal. *International Statistical Review*, 85(3), 439–454.

Berger, Y. G. (2004). A Simple variance estimator for unequal probability sampling without replacement. *Journal of Applied Statistics*, 31(3), 305-315.

Grafström, A. (2012). Spatially correlated Poisson sampling. *Journal of Statistical Planning and Inference*, 142(1), 139-147. https://doi.org/10.1016/j.jspi.2011.07.003.

Grafström, A. Lundström, N.L.P., & Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, 68(2), 514-520. doi:10.1111/j.1541-0420.2011.01699.x.

Grafström, A. & Lundström, N.L.P. (2013). Why well spread probability samples are balanced. *Open Journal of Statistics*, 3(1), 36-41.

Grafström, A. & Schelin, L. (2014). How to select representative samples. *Scandinavian Journal of Statistics*, 41(2), 277-290. https://doi.org/10.1111/sjos.12016.

Grafström, A. & Matei, A. (2018). Coordination of spatially balanced samples. *Survey Methodology*, Statistics Canada, Catalogue No. 12-001-X, 44(2), 215-238. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2018002/article/54953-eng.htm.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35(4), 1491–1523. https://doi.org/10.1214/aoms/1177700375

Hartley, H.O., & Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *The Annals of Mathematical Statistics*, 33(2), 350 –374.

Horvitz, D.G. & Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663–685.

Matei, A & Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics*, 21(4), 543–570.

Matérn, B. (1947). *Metoder att uppskatta noggrannheten vid linje- och provytetaxering.(Methods of estimating the accuracy of line and sample plot surveys)*. Meddelanden från Statens Skogsforskningsinstitut 36(1). Statens Skogsforskningsinst, Stockholm, Sweden.

Stevens, D.L. & Olsen, A.R. (2003). Variance estimation for spatially balanced samples of environmental resources. *Environmetrics*, 14(6), 593-610.

Stevens, D.L. & Olsen, A.R. (2004). Spatially Balanced Sampling of Natural Resources. *Journal of the American Statistical Association*, 99(465), 262-278. doi:10.1198/016214504000000250.

Tam, S.M. (1984). On covariance in finite population sampling. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 34(4), 429–433.

Qualité, L. (2009). *Unequal probability sampling and repeated surveys* (Doctoral dissertation, Université de Neuchâtel).

Yates, F. & Grundy, P.M. (1953.) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, Series B (Methodological)*, 15(2), 253–261.

Zhao, X., & Grafström, A. (2020). A sample coordination method to monitor totals of environmental variables. *Environmetrics*, 31(6). https://doi.org/10.1002/env.2625.

IV

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

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

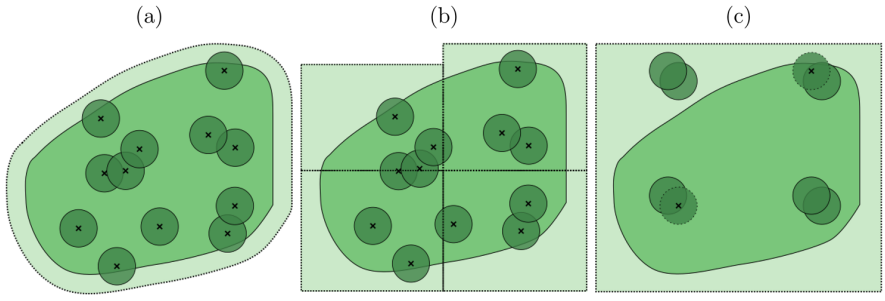


Figure 1. Examples of **a** i.i.d., **b** stratified, and **c** systematic frames and inclusion zones. The outer areas represent the sample frames (F), the inner areas represents the areas of interest (F_U), and the circles represents the inclusion zones (A) for units. In both **a** and **b**, the sample 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

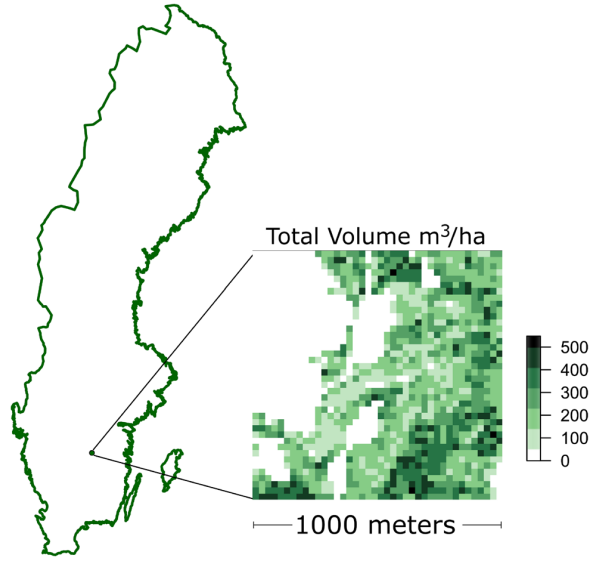


Figure 2. Location and the total biomass volume (m^3/ha) for the area used as a boilerplate for simulating the population. Darker colors indicate higher volumes (Color figure online).

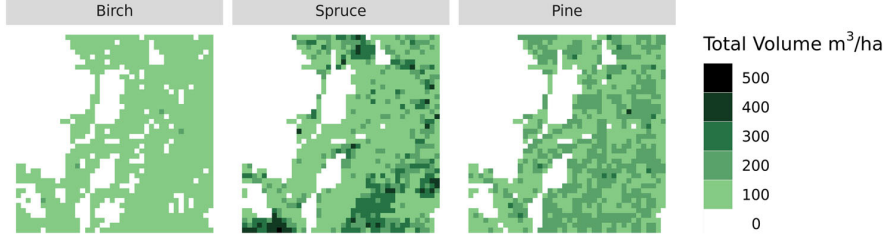


Figure 3. Total biomass volume (m^3/ha) per species for the simulated population. Darker colors indicate higher volumes (Color figure online).

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

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

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

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

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

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

and conversely, apparent from (33), we have

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

leading to

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

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