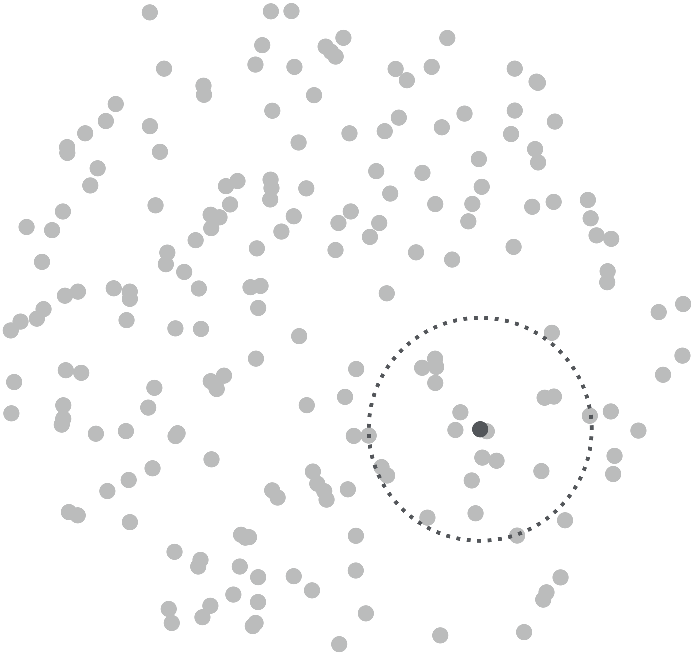
Acta Universitatis Agriculturae Sueciae

Doctoral Thesis No. 2023:80
Faculty of Forest Sciences

Contributions to the Theory of Environmental Sampling

Wilmer Prentius



Contributions to the Theory of Environmental Sampling

Wilmer Prentius
Faculty of Forest Sciences,
Department of Forest Resource Management,
Umeå

Doctoral Thesis
Swedish University of Agricultural Sciences
Umeå 2023

Acta Universitatis agriculturae Sueciae
2023:80

Illustration: The author.

ISSN 1652-6880
ISBN (print version) 978-91-8046-212-9
ISBN (electronic version) 978-91-8046-213-6

Print: Arkitektkopia, Bromma 2023

Contributions to the Theory of Environmental Sampling

Abstract

Environmental monitoring plays a crucial role in guiding climate change and conservation policy decisions. To obtain reliable insights from environmental populations, it is essential to adopt probability sampling. Furthermore, the availability of auxiliary variables can greatly enhance the quality by reducing estimator variability.

Auxiliary information can be used in different ways in a sampling design. Some designs aim to satisfy the balancing equation, i.e. selecting samples where the sample means of the auxiliary variables equal the population means. Other designs are constructed in an attempt to obtain samples well-spread, or spatially balanced, in auxiliary space, creating the sample as a miniature of the population. Paper III provides an improvement of an existing design, making it possible to increase the average spread of the sample. In Paper IV, a novel metric is introduced to assess a design's capability to yield spatially balanced samples.

Papers II and V introduce sampling designs for different types of populations one might encounter in nature. The variant of adaptive cluster sampling developed in Paper II facilitates the study of rare and clustered populations, utilizing circular plot shapes popular among practitioners. Paper V addresses the sampling of linear objects like storm-felled trees, employing aerial photographs from drones in the data collection processes.

When data are gathered from multiple surveys, various methods exist to consolidate results. A common approach involves constructing a linear combination weighted by variances. Paper I introduces a novel estimator that employs a linear combination, particularly valuable when a correlation is suspected between the estimator and the variance estimator – a frequently encountered scenario in studies involving environmental populations.

In conclusion, this thesis contributes to the field of environmental monitoring by emphasizing the critical role of probability sampling, utilization of auxiliary variables, and introducing innovative sampling designs tailored to the intricacies of environmental populations.

Keywords: area frame sampling, auxiliary variables, design-based inference, sampling design, spatially balanced sampling

Author's address: Wilmer Prentius, SLU, Department of Forest Resource Management, Skogsmarksgränd, SE-901 83 Umeå, Sweden
E-mail: wilmer.prentius@slu.se

Utveckling av samplingmetoder för miljöövervakning

Sammanfattning

Miljöövervakning har en viktig roll för att vägleda beslut inom klimatförändrings- och naturvårdsområdet. För att erhålla pålitlig information från populationer i naturen är användningen av sannolikhetsbaserade urval avgörande. Vidare kan nyttjandet av hjälpvariabler öka informationskvaliteten genom att minska estimatorernas variabilitet.

Hjälpinformation kan användas på olika sätt i en urvalsdesign. Vissa designer syftar till att uppfylla balansekvationen, det vill säga att välja urval där stickprovsmedelvärdena för hjälpvariablerna motsvarar populationsmedelvärdena. Andra designer är konstruerade för att producera urval som är väl spridda, eller rumsligt balanserade, i hjälpvariablerna, vilket kan skapa stickprov som efterliknar miniatyrer av populationen. I Studie III förbättras en befintlig design genom att öka den genomsnittliga spridningen av enheter i urvalet. Studie IV introducerar en ny metrik för att bedöma en designs förmåga att producera rumsligt balanserade urval.

Studie II och V introducerar urvalsdesigner för olika typer av populationer som är vanligt förekommande i naturen. Den variant av adaptivt klusterurval som utvecklats i Studie II underlättar undersökningar av sällsynta och klustrade populationer. Designen använder cirkulära provytor, vilka är populära bland fältarbetare. Studie V behandlar urval av linjära objekt, som stormfällda träd, genom användandet av flygfotografier från drönare i datainsamlingen.

När data samlas in från flera undersökningar finns olika metoder för att konstruera ett gemensamt resultat. En vanlig metod är att använda en linjär kombination av estimatorerna, viktade med varianser. I Studie I introduceras en ny estimator, baserad på en linjär kombination, som är särskilt inriktad på populationer där det misstänks finnas en korrelation mellan estimatorn och variansestimatorn – något som är vanligt förekommande bland populationer i naturen.

Sammanfattningsvis bidrar denna avhandling till utvecklingen av miljöövervakningen genom att betona den avgörande roll som sannolikhetsbaserade urval spelar, hur hjälpvariabler kan användas och analyseras, samt genom att introducera innovativa urvalsdesigner anpassade till de förutsättningar som finns hos populationer i naturen.

"I may not have gone where I intended to go, but I think I have ended up where I needed to be."
— Dirk Gently.

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 Prentius, W., Zhao, X. & Grafström, A. (2021). Combining environmental area frame surveys of a finite population. *Journal of Agricultural, Biological and Environmental Statistics*, 26, 250–266.

II Prentius, W. & Grafström, A. (2022). Two-phase adaptive cluster sampling with circular field plots. *Environmetrics*, 33(5), e2729.

III Prentius, W. (2023). Locally correlated Poisson sampling. (submitted)

IV Prentius, W. & Grafström, A. (2023). How to find the best sampling design: A new measure of spatial balance. (submitted)

V Grafström, A., Prentius, W., Mensah, A. A., Roberge, C., Wallerman, J. & Petersson, H. (2023). Spatially balanced line intersect sampling with curve-shaped tracts: Estimating the total volume of storm-felled trees. (submitted)

Papers I–II are published as open access.

The contribution of Wilmer Prentius 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 Monte-Carlo simulations. Was responsible for the review process.

II Planned the study together with the coauthor. Wrote most of the paper. Performed Monte-Carlo simulations. Was responsible for the review process.

III Planned the study, wrote the paper, performed the Monte-Carlo simulations, responsible for the review process.

IV Planned the study together with the coauthors. Wrote large parts of the paper. Performed Monte-Carlo simulations. Was responsible for the review process.

V Planned the study together with the coauthors. Participated in writing the manuscript.

Abbreviations

2PACS	Two-Phase Adaptive Cluster Sampling
ACS	Adaptive Cluster Sampling
CPS	Correlated Poisson Sampling
GRTS	General Random-Tessellation Stratified
HH	Hansen-Hurwitz
HT	Horvitz-Thompson
LCPS	Locally Correlated Poisson Sampling
LPM	Local Pivotal Method
MC	Multiple Count
NFI	National Forest Inventory
pdf	probability density function
SC	Single Count
SCPS	Spatially Correlated Poisson Sampling
SRS	Simple Random Sampling without replacement
SRSWR	Simple Random Sampling With Replacement
SYG	Sen-Yates-Grundy

1 Introduction

Environmental monitoring is becoming increasingly important, as policy makers seek information to guide decisions on climate change and conservation. In the European Union, the European Commission has decided on the Biodiversity strategy for 2030, outlining several environmental goals and measures, relying on better and more precise data and information (The European Commission, 2020). Simultaneously, the Forest strategy for 2030 directly outlines forest monitoring, reporting and data collection as strategic targets for the European Union (The European Commission, 2021).

In order to collect objective information about a population, surveys are often used. When a population is small, it might be possible to measure every unit in the population. However, for large populations, a survey often consists of utilizing a sample, as complete censuses are deemed to costly to perform. Different samples will yield different information. If one has a good knowledge of the population, a sample may be selected purposely to consist of the most average population units with respect to the quantity of interest (Kiær, 1976). The information that can be gained from a sample taken this way might be very good, i.e. close to the truth, however it would be hard to quantify the goodness of the information without looking into the process that selected the sample. If the pre-acquired knowledge of the population was poor, the information that can be gained from the sample will most likely also be poor.

Probability sampling is an objective way of selecting a sample. When using probability sampling, each possible sample is assigned a probability to be selected, in a way that guarantees a possibility for all units in the population to be selected. An algorithm then decides on which unit to select, while respecting the probabilities. The randomization of a probability sample ensures objectivity, and eliminates any selection bias (Särndal et al., 2003, 9). The probability distribution of the possible samples is called a sampling design.

An estimator is a function of the data, which is used to tell us something about a population parameter. If the estimator is unbiased, the probability weighted average of the estimator for all possible samples – the estimator's expected value – is equal to the desired quantity in the population. As an example, the sample mean is a commonly used unbiased estimator for the population mean, when the sample is taken through simple random sampling. The sampling design and an accompanying estimator is called a sampling strategy (Särndal, 1978).

The estimator described above adheres to the design-based inference paradigm. In design-based inference, randomness is only considered as a consequence of the sampling design. The other paradigm is model-based inference, in which we assume the gathered data as a realization of some abstract model or super-population distribution. Commonly, model-based inference is used to infer the properties of this model, as the expected value of the model distribution, or relationships between different variables in the model-based world, rather than infer the properties of any observed or unobserved realization of this model. Model-based inference mainly considers the randomness as a consequence of drawing values from the random distribution which constitutes the model. See the papers of Särndal (1978) and Gregoire (1998) for more on the differences between the inferential paradigms. This thesis only considers design-based inference, and as such any randomness invoked is only a consequence of the sampling design.

The variation of the estimator in a sampling strategy can be measured through the variance as the probability weighted squared deviation around the estimators expected value. Generally, it is desirable to use a strategy which gives an unbiased, or at least approximately unbiased estimator. Of the set of unbiased or approximately unbiased strategies, it is sensible to opt for the strategy with the lowest variance.

Use of auxiliary variables, i.e. variables for which the measurements are known for all population units prior to the sampling effort, has been of great interest in efforts to reduce the variance of the estimator in a strategy. Stratified or ordered systematic sampling are examples of early uses of auxiliary variables in the design, so is sampling with probabilities proportional to size. Historically, the focus has mainly been on reducing the variance of the estimator, through the use of ratio, difference and regression estimators.

As available computer power has increased, together with an increase in available auxiliary information, recent years has seen a shift in focus to the development of more efficient designs. An example is balanced sampling methods, which serves to respect the means of the auxiliary variables in the sample. This group of methods have historically been associated with purposive sampling (Brewer, 1994). With recent developments, it is possible to select at least approximately balanced samples using probability sampling. If there is a linear relationship between the auxiliary variables and the variable of interest, balanced samples can be very efficient for certain estimators (Cochran, 1977). Other approaches have instead focused on selecting samples well-spread in the auxiliary variables, in the hope of capturing the general distribution of the population. In addition, well-spread designs are also approximately balanced (Grafström and Lundström, 2013).

In Section 2, the general estimators will be introduced, together with some methods for incorporating auxiliary information in the design. Section 2.3 will introduce some measures for spatial balance, i.e. tools for determining how well a design captures the general distribution of the population. As many environmental applications lack sampling frames, indirect sampling designs are needed, such as sampling from an area frame, for which Section 3 is devoted. Finally, some final remarks are given in Section 4.

2 Sampling from a finite population

Sampling is often performed when wanting to estimate some function of a variable of interest y, observable for all units in the population, while deeming a complete census infeasible due to financial or practical reasons. A common function is the population total $Y = \sum_{i \in U} y_i$, where U denotes the set of labels $1, \ldots, N$ corresponding to the N units in the population. Although other variations may be of interest, such as the population mean Y/N, we will solely consider the population total in this thesis. In order to draw a sample, we need a sampling design, directing how the selection process should be.

A common sampling design for sampling from finite populations is the Simple Random Sampling without replacement (SRS). For SRS, a sample of size n is drawn without replacement, where each unit in the population has an equal probability of being included into the sample. We call this probability the inclusion probability π_i of a unit i. For SRS, $\pi_i = n/N$.

A general estimator of Y for sampling designs performed without replacement, i.e. Single Count (SC) designs, is the Horvitz-Thompson (HT)-estimator

$$\hat{Y} = \sum_{i \in U} \frac{y_i}{\pi_i} \mathrm{I}(S_i > 0)\,, \tag{1}$$

where $\mathrm{I}(\cdot)$ denotes the indicator function, and S_i is the number of inclusions of unit i (Horvitz and Thompson, 1952). This estimator is unbiased for all designs where $\pi_i > 0$ for all units $i \in U$.

For designs producing fixed sized samples, such as the SRS, the variance of (1) is

$$\mathrm{V}\big(\hat{Y}\big) = -\frac{1}{2} \sum_{i \in U} \sum_{j \in U} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 (\pi_{ij} - \pi_i \pi_j), \tag{2}$$

where π_{ij} denotes the second order inclusion probability, or the probability of two units i, j being included together in a sample. The second order inclusion probability for SRS is $\pi_{ij} = n(n-1)/(N(N-1))$ when $i \neq j$. The form of variance (2) is called the Sen-Yates-Grundy (SYG)-form variance (Sen, 1953; Yates and Grundy, 1953), and can be estimated through

$$\hat{\mathrm{V}}(\hat{Y}) = -\frac{1}{2}\sum_{i\in U}\sum_{j\in U}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\mathrm{I}(S_i>0\cap S_j>0)\,. \tag{3}$$

Not all designs can guarantee fixed sized samples. For such designs, the general variance of the HT-estimator is

$$\mathrm{V}(\hat{Y}) = \sum_{i\in U}\sum_{j\in U}\frac{y_iy_j}{\pi_i\pi_j}(\pi_{ij}-\pi_i\pi_j),$$

which can be estimated through

$$\hat{\mathrm{V}}(\hat{Y}) = \sum_{i\in U}\sum_{j\in U}\frac{y_iy_j}{\pi_i\pi_j}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\mathrm{I}(S_i>0\cap S_j>0)\,. \tag{4}$$

Both (3) and (4) are unbiased if $\pi_{ij}>0$ for all pair of units i, j.

Some designs allows the same unit to be selected multiple times, i.e. Multiple Count (MC) designs. An example of an MC design is the Simple Random Sampling With Replacement (SRSWR), where n units are randomly drawn, and each unit in the population has an equal expectation of number of inclusions in the sample. This expectation is called the expected number of inclusions μ_i of a unit i. For SRSWR, $\mu_i = n/N$, as each unit has a draw probability of $1/N$, and n independent draws are made. A general estimator for MC designs is the Hansen-Hurwitz (HH)-estimator

$$\hat{Y} = \sum_{i\in U}\frac{y_i}{\mu_i}S_i, \tag{5}$$

which is unbiased for designs where $\mu_i>0$ for all units i (Hansen and Hurwitz, 1943).

When the sample size is fixed, such as for the SRSWR design, the variance in SYG-form of (5) is

$$\mathrm{V}(\hat{Y}) = -\frac{1}{2}\sum_{i\in U}\sum_{j\in U}\left(\frac{y_i}{\mu_i}-\frac{y_j}{\mu_j}\right)^2(\mu_{ij}-\mu_i\mu_j),$$

which has an estimator

$$\hat{\mathrm{V}}(\hat{Y}) = -\frac{1}{2}\sum_{i\in U}\sum_{j\in U}\left(\frac{y_i}{\mu_i}-\frac{y_j}{\mu_j}\right)^2\frac{\mu_{ij}-\mu_i\mu_j}{\mu_{ij}}S_iS_j. \tag{6}$$

The general variance of (5) is

$$\mathrm{V}\big(\hat{Y}\big) = \sum_{i\in U}\sum_{j\in U} \frac{y_i y_j}{\mu_i \mu_j}(\mu_{ij} - \mu_i \mu_j),$$

and can be estimated through

$$\hat{\mathrm{V}}\big(\hat{Y}\big) = \sum_{i\in U}\sum_{j\in U} \frac{y_i y_j}{\mu_i \mu_j} \frac{\mu_{ij} - \mu_i \mu_j}{\mu_{ij}} S_i S_j, \tag{7}$$

where μ_{ij} denotes the second order expected number of inclusions for a pair of units i, j. The estimators (6) and (7) are unbiased for all designs where $\mu_{ij} > 0$ for all pair of units. For SRSWR, $\mu_{ij} = n(n-1)/N^2$ for all pairs $i \neq j$.

2.1 Sampling designs for finite populations

For a more formal definition of a sampling design, we first consider the random vector $\mathbf{S} \in \mathbb{N}_0^N$ containing the numbers of inclusions of the population units, with elements $S_i = k$ if the ith unit in U is included in the sample k times. The sampling design can then be defined by the probability distribution p on the inclusion vector $\mathbf{S}$. Let $\mathscr{R} = \{\mathbf{s} \in \mathbb{N}_0^N : p(\mathbf{s}) > 0\}$ be the support of p, i.e. the set containing the possible outcomes of the designs, where $\mathbf{s}$ denotes the realization of the random vector $\mathbf{S}$. If $\mathscr{R} \subseteq \{0,1\}^N$, the sampling design is without replacement, as elements in $\mathbf{S}$ is confined to having values $\{0,1\}$. On the other hand, the design is with replacement if $\mathscr{R} \setminus \{0,1\}^N \neq \emptyset$, i.e. it is possible to select some unit multiple times. A design produces fixed sized samples of size n if $\mathscr{R} \subseteq \mathscr{R}_n$, where $\mathscr{R}_n = \{\mathbf{s} \in \mathbb{N}_0^N : \sum_{i\in U} s_i = n\}$. (Tillé, 2006, 7).

Through the design, the inclusion probabilities of the units in the population can be described as

$$\pi_i = \sum_{\mathbf{s}\in\mathscr{R}} p(\mathbf{s})\,\mathrm{I}(s_i > 0) = \mathrm{P}(S_i > 0)\,,$$

whereas the expected number of inclusions is defined by

$$\mu_i = \sum_{\mathbf{s}\in\mathscr{R}} p(\mathbf{s}) s_i = \mathrm{E}(S_i)\,.$$

We call these terms the (first order) design properties. Similarly, the second order design properties are defined as

$$\pi_{ij} = \sum_{\mathbf{s}\in\mathscr{R}} p(\mathbf{s})\,\mathrm{I}(s_i > 0 \cap s_j > 0) = \mathrm{P}(S_i > 0 \cap S_j > 0)\,,$$

and

$$\mu_{ij} = \sum_{\mathbf{s} \in \mathscr{R}} p(\mathbf{s}) s_i s_j = \mathrm{E}(S_i S_j) \, .$$

The estimators (1) and (5) are unbiased as long as the support of the design allows all units to be selected, i.e.

$$\forall i \in U \; \exists \, \mathbf{s} \in \mathscr{R} : s_i > 0,$$

whereas the variance estimators are unbiased if the support allows all pairs of units to be selected together

$$\forall i, j \in U \; \exists \, \mathbf{s} \in \mathscr{R} : s_i s_j > 0.$$

2.2 Using auxiliary variables

The use of auxiliary variables in sampling is almost as old as sampling itself (Kiær, 1976). As an example, consider a survey of the average use of public transportation, where prior knowledge is available on the subjects in the population on whether or not they live in a rural or urban area. If we expect that the availability of public transformation differs between rural and urban areas, it seems natural to want to ensure that the sample includes subjects from both types of areas. Using an SRS could, however, result in a sample containing only rural or only urban subjects. Thus, we might be inclined to use a stratified sampling design, i.e. a design constructed to respect the proportion of rural and urban subjects. In this case, it is reasonable to suspect that using the prior information – the auxiliary variable – would lead to a decrease in variance of the estimator.

For another example, the prior knowledge instead contains information about the subjects ages. If we expect that school-aged subjects and elderly subjects uses the public transportation more than others, we would be inclined to use this knowledge in the sampling process. However, compared to the previous example, we no longer have categorical data, but continuous, why we cannot employ the stratified design. Instead, we can use an ordered systematic sample, where every subject is ordered by age, selecting every kth subject, $k = N/n$, with a randomized start. By doing so, we eliminate the risk of selecting a sample containing only school-aged subjects, and ensures a greater similarity between the distribution of ages in the population and the distribution of ages in the sample.

When the Swedish National Forest Inventory (NFI) started in 1923, the initial design selected trees to survey through placing belts in a regular pattern over Sweden, measuring all trees along the 10m wide corridors. Later, in 1953, the strategy shifted to randomly placing circular plots in a square formation over the landscape, as information about the forest state acquired during the previous years hinted towards similarity between nearby units (Fridman et al., 2014). The observed tendency has later been popularized by Tobler (1970), as Tobler's first law of geography: "everything is related to everything else, but near things are more related than distant things." This tells us that we should aim to separate the units in the sample, increasing the distance between them, as close units are more likely to convey the same information compared to units farther apart.

This idea is analogous to the designs in the examples. The stratified and ordered systematic designs ensures that no part of the population is over-represented in the sample – i.e. the risk of clustering is reduced. However, while efficient, these designs have an obvious drawback in that they are limited to a single or few auxiliary variables, and other methods are needed if we want to utilize multiple auxiliary variables in the sampling design, especially if wanting to use unequal probability sampling.

Although the specific scenario of the Swedish NFI is a case of sampling finite populations using area frames, which is the topic of Section 3, there are many other examples of direct sampling from finite populations in environmental studies. In many of these, longitude and latitude is an obvious set of auxiliary variables.

An early method which attempted to spread samples in longitude and latitude is the General Random-Tessellation Stratified (GRTS) design (Stevens and Olsen, 2004). The GRTS works similarly to Peano curves, or space-filling curves, by recursively tessellating the quadrants of the two-dimensional auxiliary space into smaller and smaller squares, until at most one unit exists in each square, while preserving the order through the depth of the tessellation. The main difference to Peano curves is that the ordering of the quadrant in each step of the tessellation process is random. When the tessellation process is complete, a one-dimensional ordering has been constructed, from which it is possible to draw a sample using an ordered systematic design, with unequal probabilities if needed.

While the GRTS retains some kind of ordering of the population, information is obviously lost when going from two dimensions to one dimension. Another approach was introduced by Grafström (2012), called Spatially Correlated Poisson Sampling (SCPS), an adaptation of Correlated Poisson Sampling (CPS) by Bondesson and Thorburn (2008). In SCPS, a decision is taken for a single unit at a time, to either include or exclude it from the sample, by random. After the decision is made, any undecided units gets their probabilities updated in a way that ensures that the unconditional inclusion probabilities remains the same, i.e. the conditional inclusion probabilities are martingales (Bondesson and Thorburn, 2008). The main difference between CPS and SCPS is that probabilities are updated according to a priority order, updating units close to the included/excluded unit before units far away. In short, the process can be described as:

a) Initialize the conditional inclusion probabilities $\pi_i^{(t)}$ at step $t = 0$ to the prescribed inclusion probabilities, i.e. $\pi_i^{(0)} = \pi_i$. Any unit that has integer conditional probability mass, i.e. 0 or 1, at any step, is considered (definitely) excluded from or included in the sample.

b) Select a unit $i(t)$ from the undecided units $U^{(t)}$, i.e. units for which $0 < \pi_i^{(t)} < 1$. This unit can be selected at random or by traversing through the frame. Draw a random number $U^{(t)} \sim U(0,1)$. If $u^{(t)} \leq \pi_{i(t)}^{(t)}$, set $\pi_i^{(t+1)} = 1$, otherwise set $\pi_i^{(t+1)} = 0$.

c) Order the remaining units by some distance measure in auxiliary space, into an ordered list.

d) Sequentially, through the ordered list, update the conditional inclusion probabilities to

$$\pi_{j(t+1)} = \begin{cases} \pi_{j(t)} - (1 - \pi_{i(t)}) w_{i(t),j(t)}, & \text{if } i(t) \text{ was included,} \\ \pi_{j(t)} + \pi_{i(t)} w_{i(t),j(t)}, & \text{if } i(t) \text{ was excluded,} \end{cases}$$

where w denotes the largest weights

$$0 \leq w_{i(t),j(t)} \leq \min\left(\frac{\pi_{j(t)}}{1 - \pi_{i(t)}}, \frac{1 - \pi_{j(t)}}{\pi_{i(t)}} \right),$$

while ensuring that the sum of the used weights is not larger than 1. Figure 1 shows an example of a unit selected in b), and two units which would get their probabilities updated as an result of the outcome of the selected unit.

e) Repeat the process from b) as step $t+1$, until all units are either considered included or excluded. If only one unit remains, the process is finished after b).

The inclusion of c) in the process described above creates a negative correlation between the numbers of inclusions S_i, and S_j, dependent on the distance between i, j, which effectively spreads the sample in auxiliary space. A similar idea is behind Local Pivotal Method (LPM), introduced by Grafström et al. (2012). In LPM, or more specifically LPM 2, one competing unit

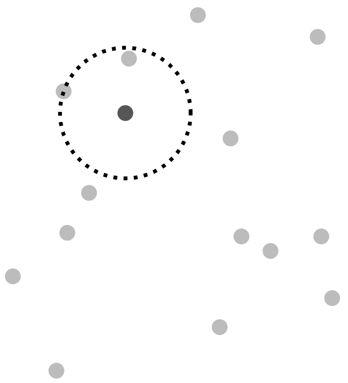


Figure 1: A unit (dark dot) is selected at some step in the SCPS process. The two closest units, i.e. units within the circle, will have their probabilities updated as a result of the outcome of the selected unit.

is chosen at random, together with its nearest neighbour, amongst the units without integer conditional inclusion probabilities. An example of the set of pairs possible to choose randomly from is shown in Figure 2. By a random decision, the conditional inclusion probabilities of the competing units are updated, moving as much probability mass as is possible in the direction of one of the units, ensuring that at least one of the competing units get integer conditional inclusion probability.

In the other variant of LPM, LPM 1, competing units are selected only from the set of pairs of mutual nearest neighbours. Figure 2 highlights these pairs by solid lines.

As the average distance between competing units in LPM 1 is lower, compared to LPM 2, the negative correlation created between the numbers of inclusions S_i and S_j is generally higher for close units in LPM 1, resulting in a greater spread between the units (Grafström et al., 2012). In Paper III, this idea is applied on SCPS by modifying the selection part in b), resulting in the method Locally Correlated Poisson Sampling (LCPS). For SCPS, the selected unit $i(t)$ in b) is randomly drawn, or sequentially traversed through U. By instead choosing $i(t)$ by minimizing the distance of the furthest away unit with a positive weight, see Figure 3, we can increase the negative correlation introduced between numbers of inclusions S_i and S_j that are nearby each other.

Another set of designs are designs which aims to satisfy the balancing equation

$$\sum_{i\in U} x_i = \sum_{i\in U} \frac{x_i}{\pi_i} \mathrm{I}(S_i > 0) , \tag{8}$$

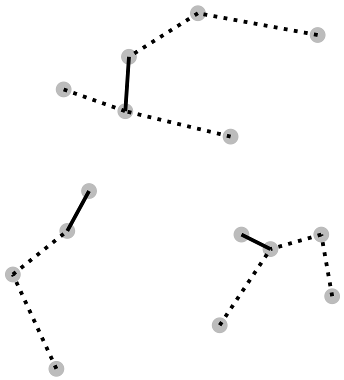


Figure 2: Example of possible pairs of units, which can be selected at random at some step in the LPM process. In LPM 2, any of the pairs can be selected, whereas in LPM 1, only the pairs with a solid line between them can be selected.

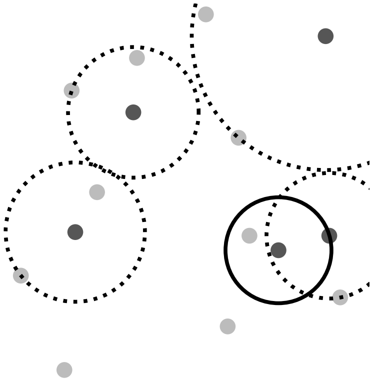


Figure 3: Example of five candidate units (dark dots) to select at some step in the LCPS process. The decision to include or exclude any of these units would affect the conditional inclusion probabilities for units with positive weights. The affected units are shown to be inside some distance metric around the candidate units (circles). The LCPS chooses to decide the outcome of the candidate unit which affects units within the smallest possible distance, shown by the unit with a solid line circle.

for the available auxiliary variables x. This is equivalent to selecting a sample where the estimate of the total (or mean) of the auxiliary variables are exactly the total (or mean) of the population (Yates, 1946), and has been shown to have desirable properties, especially if there exists a linear relationship between the auxiliary variables and the variable of interest (Royall and Herson, 1973). It has been shown that designs that produces well-spread samples are approximately balanced (Grafström and Lundström, 2013).

The first general approach which managed to approximately respect the balancing equation for multiple variables is the Cube method by Deville and Tillé (2004). Similarly to LPM and SCPS, the Cube method works sequentially, deciding the outcomes for at least one unit at each step of the process. Being a SC design, Section 2.1 stated that the support of the number of inclusions is $\mathscr{R} = \{0, 1\}^N$, which can be considered as the vertices of the hypercube $[0, 1]^N$, for which the vector of inclusion probabilities $\boldsymbol{\pi}$ is a member. The Cube method operates by finding the hyperplane within this hypercube for which the balancing equation holds, and at each step of the process moving in the direction of at least one vertex of the hypercube in a manner that respects the inclusion probabilities.

Today, there exists a multitude of sampling designs that successfully incorporates auxiliary variables. Some more recent includes the Balanced Acceptance Sampling, which selects well-spread samples using quasi-random numbers, i.e. evenly spread uniform random numbers (Robertson et al., 2013); a sampling design with probability function proportional to the distances, which iteratively reshapes the sample in order to achieve a greater spread, measured through an index dependent on the distance matrix (Benedetti and Piersimoni, 2017); the Local Cube method, a variant of the Cube method, aiming to achieve both well-spread and balanced samples (Grafström and Tillé, 2013); the Weakly Associated Vector (WAVE) method, another variant of the Cube method, however traversing through a hyperplane defined by a stratification matrix instead of the balancing equation (Jauslin and Tillé, 2020).

The principle behind the mentioned methods, excluding the design with probability function proportional to the distances, is that they all operate by adjusting the second order inclusion probabilities. Some in order to create a negative correlation between nearby numbers of inclusions S_i and S_j, others in order to achieve balance. As such, variance estimators such as (3) and (4) can often not be used, either due to it being infeasible to calculate the second order inclusion probabilities, or because some might be 0. For these designs, alternative variance estimators are needed. A strategy is to use the variance estimator of assuming a SRS desing, another to use some local variance estimator (see Grafström and Schelin, 2014; Stevens Jr and Olsen, 2003; Zhao and Grafström, 2023).

2.3 Measuring the Spatial Balance

The idea behind designs that are well-spread in auxiliary space is the assumption that the values which are similar in auxiliary space may also be similar among the variables of interest. By spreading the observations, it should therefore be possible to get a sample with a better representation of the population. As exemplified through Tobler's first law of geography, this assumption is well accepted in the study of environmental populations. If a well-spread design is used and no relationship exists between the variables of interest and the auxiliary variables, one would still be left with a design that cannot perform any worse than if the auxiliary variables had been disregarded.

The quantification of how well a design manages to represent the population is the measure of spatial balance of the design B. The basis for this quantification is the similarity between the distribution of the auxiliary variables of the samples produced by the design, and the population distribution. However, for an unequal probability design, such a basis would need to disregard the inclusion probabilities, as the unequal probability design can be seen as an intentional transformation of the auxiliary space.

Grafström and Schelin (2014) makes a distinction between representative samples and spatially balanced samples. A representative sample is said to be a sample where

$$\sum_{i \in U^*} \mathrm{I}(S_i > 0) = \frac{n}{N} N^*, \tag{9}$$

for every coherent subset $U^* \subset U$ with size N^*. A subset is considered coherent if it is possible to construct a convex region in auxiliary space containing only the units in the subset. Thus, a sample is representative if it resembles a miniature of the population, i.e. the empirical distribution of the sample resembles the empirical distribution of the population. A sample is said to be spatially balanced if

$$\sum_{i \in U^*} \mathrm{I}(S_i > 0) = \sum_{i \in U^*} \pi_i, \tag{10}$$

for every coherent subset $U^* \subset U$. This can be interpreted as the design selecting the correct amount of sample units within each subset U^*. It is easy to see that if $\pi_i = n/N$ for all units i, there is an agreement between (9) and (10).

We make a distinction between the measure of spatial balance for a sample, and the measure of the spatial balance for a design, where the latter is the

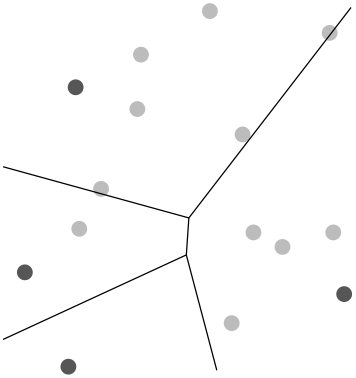


Figure 4: The Voronoi tessellation of the auxiliary space, with population units as dots and four sample units marked by dark colour.

expected spatial balance for all possible samples in the design. A measure of the spatial balance of a sample is the Voronoi spatial balance measure, introduced by Stevens and Olsen (2004). Let $U_1, \ldots, U_n$ be the partition of units in the population around the sample points, such that U_j includes all units closer to the jth sample unit than any other sample point, i.e. a partition of the units created by the Voronoi tessellation of the auxiliary space, see Figure 4. A spatially balanced sample should, on average, have a probability mass

$$v_j = \sum_{k \in U_j} \pi_k,$$

of 1 inside each Voronoi polytope (Grafström and Schelin, 2014). The Voronoi spatial balance measure can then be defined as

$$B_{\mathrm{VO}}(\mathbf{s}) = \frac{1}{n} \sum_{j=1}^{n} (v_j - 1)^2.$$

A drawback with the Voronoi spatial balance measure is that the measure has no clearly defined boundaries, as it is highly dependent on the spatial pattern of the population. In order to evaluate a design, one needs to compare the design with that of a baseline, normally SRS. The spatial balance measure based on Moran's I is a modification of the Moran's I measure of spatial correlation (Moran, 1950), normalized to be defined on $[-1, 1]$, indicating perfect spatial balance to maximum clustering, with a clear benchmark value at 0 (Tillé et al., 2018). The measure is dependent on specifying a weight matrix, with two variants given in Tillé et al. (2018) and Jauslin and Tillé (2020).

In Paper IV, a new definition of the spatial balance (10) is proposed. Let G be the design-weighted empirical distribution of the population for some auxiliary variable vector $\mathbf{x}$, as

$$G_{\mathbf{x}}(U^*) = \frac{1}{X} \sum_{i \in U^*} x_i,$$

where X denotes the total of the auxiliary variable $\mathbf{x}$. Furthermore, let $\widehat{G}$ denote the design-weighted empirical distribution of the sample, as

$$\widehat{G}_{\mathbf{x}}(U^*, \mathbf{s}) = \frac{1}{X} \sum_{i \in U^*} \frac{x_i}{\pi_i} \mathrm{I}(s_i > 0).$$

A sample has perfect spatial balance if

$$\widehat{G}_{\mathbf{x}}(U^*, \mathbf{s}) = G_{\mathbf{x}}(U^*), \tag{11}$$

for all coherent subsets $U^* \subset U$ and auxiliary variables $\mathbf{x}$. This can be seen as a generalization of (10), as when $\mathbf{x} = \boldsymbol{\pi}$, (10) and (11) are equivalent. Notably, $\widehat{G}_{\mathbf{x}}(U, \mathbf{s}) = G_{\mathbf{x}}(U)$ is proportional to the balancing equation (8).

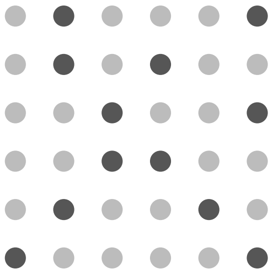
Paper IV also proposes a new measure of spatial balance, based on a Voronoi partitioning of the population, similar to the Voronoi spatial balance measure. The proposed measure promotes samples where the balancing equation is fulfilled locally within each Voronoi polytope. Let d_x be the disparity of the design-weighted empirical distributions, i.e.

$$d_{\mathbf{x}}(U_j, \mathbf{s}) = X \left(\widehat{G}_{\mathbf{x}}(U_j, \mathbf{s}) - G_{\mathbf{x}}(U_j) \right),$$

and let $\mathbf{d}(U_j, \mathbf{s})$ be the vectorization over all auxiliary variables. The proposed balance measure is defined as

$$B_{\mathrm{LB}}(\mathbf{s}) = \sqrt{\frac{1}{N} \sum_{i=1}^{n} \mathbf{d}(U_j, \mathbf{s})^{\mathsf{T}} \mathbf{Q}^{-1} \mathbf{d}(U_j, \mathbf{s})},$$

where $\mathbf{Q} = \sum_{i \in U} \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}}$. As the proposed measure takes into account the local balancing of other auxiliary variables, the measure can be used to more effectively discriminate between individual samples, something that is harder to with the Voronoi spatial balance measure. Furthermore, simulations shows that the proposed measure yields more consistent results over different sample sizes, compared to the other measures, which makes it more suitable to use if comparing samples of different sizes. An example of the different measures, denoted VO and MI for the measures based on Voronoi polytopes and Moran's I respectively, and LB for the proposed measure, is shown in Figure 5.



VO	MI	LB
0.03	-0.71	0.34
0.11	-0.86	0.34
0.11	-1.00	0.48
0.00	–	0.41
0.00	-1.00	0.00
0.00	-0.50	0.41

Figure 5: Rows of six samples, from a population of six units choosing two, and their corresponding measures of spatial balance. The measure based on Moran's I is undefined for the 4th sample.

3 Sampling finite populations using area frames

In the previous section, we considered sampling designs using list frames, i.e. frames where we are able not only to conceptualize the population U, but can actually label the frame according to the labels in U. For many environmental studies, such as the Swedish NFI previously used in an example, a list frame like that cannot be constructed without an effort almost as tedious as conducting a complete census.

An approach to the problem of sampling a finite population without a list frame is to use an indirect sampling scheme, such as cluster sampling, where we cluster the population through some auxiliary information, creating a sampling frame. After the sampling frame is constructed, a sample is drawn according to some finite population design, and the results are aggregated from the observational units to the sampling units, and a finite population estimator can be used. An example of a clustered sampling procedure could be a random selection of school classes, where all children within each selected class would be surveyed. In environmental studies, the area frame (i.e. the map) is readily available to be used in the clustering process, and a sampling frame can be created by tessellating the area frame, creating a finite frame of geographical units from which a sample can be drawn, see Figure 6.

Another approach is to use plot sampling, where sample units are selected by placing plots, through some random procedure, on the area frame, surveying any units which falls within the boundaries of these plots (Gregoire and Valentine, 2007, 207). The random procedure placing the plots on the area frame is defined by drawing a random point as the plot centre using some probability density function (pdf) f on the area frame. Another way to look at the selection process is to imagine each unit having an inclusion zone, as the horizontal and vertical reflection of the plot shape, see Figure 7. A unit is included if the plot centre is located within the unit's inclusion zone.

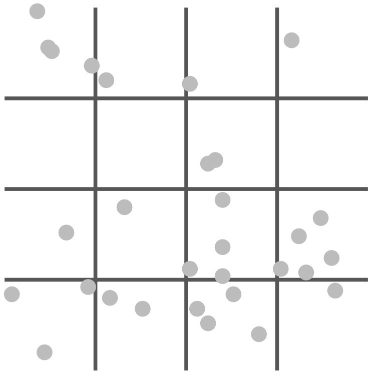


Figure 6: 16 sampling units created by tessellating an area frame by a regular grid. The value of the variables of interest are determined by aggregating the observational units (dots) within each cluster.

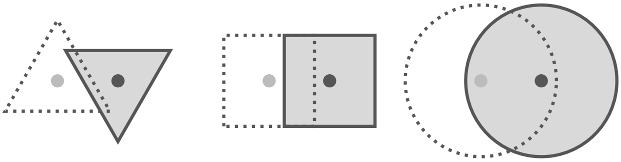


Figure 7: Three examples of plots (solid lines) around plot centres (dark dots), and inclusion zones (dashed lines) around units (light dots). If a plot centre falls within the inclusion zone of a unit, the unit will be selected, as it will be inside the plot border.

We can describe plot sampling as follows. A population U exists within an area frame $\mathscr{F} \subset \mathbb{R}^2$. Around each unit $i \in U$, there is an inclusion zone A_i, such that if a sample point $\mathbf{z}_k$ drawn from a distribution defined by the pdf f lands within A_i, unit i is included in the sample. A derivation of the general design properties is done in Paper I.

It is sometimes desirable to ensure equal inclusion probabilities for all population units, i.e. by choosing f to be the uniform distribution over $\mathscr{F}$. If any unit has an inclusion zone not fully within $\mathscr{F}$, that unit would have a smaller inclusion probability compared to units with inclusion zones fully within $\mathscr{F}$. A simple way to handle this issue is to extend the area frame so that it is guaranteed to include all inclusion zones fully, i.e. by extending the area frame by the radius of the plot.

3.1 Continuous population approach

When the population is continuous, another approach is needed in order to estimate a total. For continuous populations, instead of having countable population units, the population exists as a Lebesgue integrable function y on the frame $\mathscr{F} \subset \mathbb{R}^k$, as an per-area density. Here, the total of the function y, $Y = \int_{\mathscr{F}} y(\mathbf{x})d\mathbf{x}$, or some variation thereof, is of interest. In environmental sampling, it is usually the case that $\mathscr{F}$ is a subspace of $\mathbb{R}^2$, i.e. a map. A sample is drawn by selecting n sample points $\mathbf{Z}_i$ on $\mathscr{F}$, each according to some (marginal) pdf f_i. The joint pdf then constitutes the sampling design on $\mathscr{F}$ (Cordy, 1993).

From the marginal pdf's, we can derive the sampling intensity function

$$\pi(\mathbf{z}) = \sum_{i=1}^{n} f_i(\mathbf{z}),$$

and the second order sampling intensity function

$$\pi(\mathbf{z},\mathbf{z}') = \sum_{i=1}^{n} \sum_{j \neq i} f_{ij}(\mathbf{z},\mathbf{z}'),$$

where f_{ij} denotes the joint marginal pdf of sample points i, j.

An estimator for the total is the continuous HT estimator

$$\hat{Y} = \sum_{i=1}^{n} \frac{y(\mathbf{Z}_i)}{\pi(\mathbf{Z}_i)}. \tag{12}$$

The estimator is unbiased if $\pi(\mathbf{z}) > 0$ for all $\mathbf{z} \in \mathscr{F}$, and y is bounded or non-negative. Assuming that y is bounded, the variance of (12) is

$$V(\hat{Y}) = \int_{\mathscr{F}} \frac{y(\mathbf{z})^2}{\pi(\mathbf{z})} d\mathbf{z} + \int_{\mathscr{F}} \int_{\mathscr{F}} \frac{y(\mathbf{z})y(\mathbf{z}')}{\pi(\mathbf{z})\pi(\mathbf{z}')} (\pi(\mathbf{z},\mathbf{z}') - \pi(\mathbf{z})\pi(\mathbf{z}')) d\mathbf{z} d\mathbf{z}'.$$

An estimator of the variance is

$$\hat{V}(\hat{Y}) = \sum_{i=1}^{n} \frac{y(\mathbf{Z}_i)^2}{\pi(\mathbf{Z}_i)^2} + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} \frac{y(\mathbf{Z}_i)y(\mathbf{Z}_j)}{\pi(\mathbf{Z}_i)\pi(\mathbf{Z}_j)} \frac{\pi(\mathbf{Z}_i,\mathbf{Z}_j)-\pi(\mathbf{Z}_i)\pi(\mathbf{Z}_j)}{\pi(\mathbf{Z}_i,\mathbf{Z}_j)},$$

and is unbiased if $\pi(\mathbf{z},\mathbf{z}') > 0$ for all $\mathbf{z},\mathbf{z}' \in \mathscr{F}$ (Cordy, 1993).

When discrete populations are sampled from an area frame through plot sampling, it can be useful to map the values y_i of the discrete population U to the continuous function

$$y(\mathbf{z}) = \sum_{i \in U} \frac{y_i}{|A_i|} \mathrm{I}(\mathbf{z} \in A_i),$$

assuming fixed shaped plots and inclusion zones fully within $\mathscr{F}$ (Mandallaz, 2007, 56). The mapping above ensures that the function total $\int_{\mathscr{F}} y(\mathbf{z})d\mathbf{z}$ equals the discrete population total Y (Grafström et al., 2017).

3.2 Adaptive cluster sampling

A common feature of environmental populations is clustering or fragmentation. An example is dead wood, which occurs in higher rates in old-growth forests, compared to managed forests (Talvitie et al., 2006). When populations show these kind of patterns, plot sampling might be relatively costly, as the sample size will have to increase in order to achieve acceptable levels of confidence when many plots will experience absence (Thompson, 1990).

If clustering is expected, it might be beneficial to use Adaptive Cluster Sampling (ACS), which makes use of the observed values y in the sampling process, increasing the sampling effort near plots where occurrence has been found.

Similarly to the cluster sampling procedure previously described, ACS begins by tessellating the area frame into a grid of plots, see Figure 8. From the plots created by the tessellation, a sample is drawn using a finite sampling design. When a plot is visited, and occurrence is observed, neighbouring plots are also visited. If occurrence is observed in any neighbouring plots, unvisited neighbouring plots are visited as well, and so forth, until the cluster has been fully surveyed (Thompson, 1990). Thus, if any plot in a cluster is selected during the sampling process, the whole cluster will be surveyed. As such, the design properties cannot be calculated when taking the sample, but can be calculated after the survey. The ACS can be effective, if the cost of checking occurrence is low relative to the cost of measuring.

Let $\mathbf{S}^{(1)}$ denote the number of inclusions of plots resulting from the initial sample, taken without replacement, with first and second order inclusion

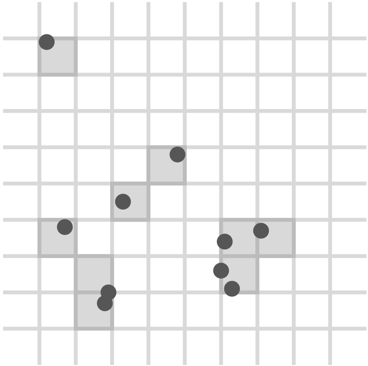


Figure 8: Tessellation of an area frame into 100 square plots. Out of the 100 square plots, the population of interest (dots) only exists within 9 dark squares.

probabilities π_i and π_{ij} respectively. Let $C(i)$ denote a cluster formed around plot i, or i itself if empty. Furthermore, let $\mathbf{S}$ denote the number of inclusions of plots resulting from the survey, i.e. the sum of plots either surveyed in $\mathbf{S}^{(1)}$ or as a result of the expansion around any such plot, as

$$S_i = \sum_{k \in C(i)} S_k^{(1)}. \tag{13}$$

The expected number of inclusions for a plot becomes

$$\mu_i = \sum_{k \in C(i)} \pi_k, \tag{14}$$

with second order expected number of inclusions

$$\mu_{ij} = \sum_{k \in C(i)} \sum_{l \in C(j)} \pi_{kl}, \tag{15}$$

and the HH-estimator (5) can be used for the total.

Plot sampling commonly uses circular plot shapes. One reason may be that circular plots are relatively easy to use in field work, as one needs to locate a centre point, and then only have to account for the distance to this centre point. In Paper II, we introduce a development of the ACS, named Two-Phase Adaptive Cluster Sampling (2PACS), where the plots in the initial sample consists of circles, instead of squares. As circles cannot fully tessellate an area, a two-phase sampling design is needed.

In the first phase, a regular lattice of circular plots is randomly placed upon the surface with centre points $\mathbf{z}$, see Figure 9. Let U_1 denote the set of

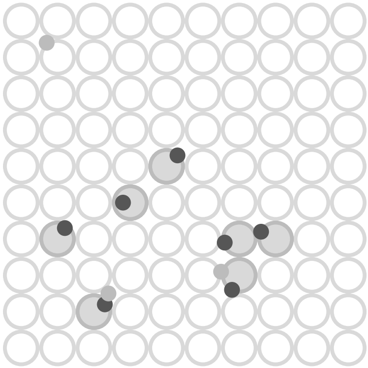


Figure 9: A lattice of 100 circular plots on an area frame. Out of the 100 plots, the population of interest (dots) only exists within 7 dark circles.

these plots. This first phase constitutes a sample according to the continuous population approach, with plot values

$$y_i = \frac{y(\mathbf{z}_i)}{\pi(\mathbf{z}_i)},$$

where the sampling intensity $\pi(\mathbf{z}) = \lambda^{-2}$ is defined by the separation of the plot centres λ. The first phase continuous HT-estimator, according to (12), becomes

$$\hat{Y}_1 = \sum_{i \in U_1} y_i.$$

The clusters $C(i)$ formed around each plot i is conditional to this first phase sample U_1, and the second phase sample is taken by using a without replacement sampling design from U_1. This leads to the number of inclusions $\mathbf{S}$, defined as in (13), being conditioned on the first phase sample, and hence also the second phase design properties (14) and (15), resulting in a conditional HH-estimator

$$\hat{Y}_2 = \sum_{i \in U_1} \frac{y_i}{\mu_i} S_i.$$

While the law of total expectation shows this estimator as unbiased

$$\mathrm{E}\big(\hat{Y}_2\big) = \mathrm{E}_1\big(\mathrm{E}_2\big(\hat{Y}_2 | U_1\big)\big) = \mathrm{E}_1\big(\hat{Y}\big) = Y, \tag{16}$$

where $\mathrm{E}_1, \mathrm{E}_2$ denotes the expectation under the first and second phase respectively, an alternative variance estimator is needed and presented in Paper II.

If the population has a form which tends to create large clusters, it may be that the resulting surveying effort is considerable. In Paper II, we suggest a restriction on the definition of the clusters $C(i)$ so that it never expands farther away than a set distance from the plot i, reducing the sample size for worst-case scenarios. Other strategies for restricting the final sample size includes increasing the cut-off rate, i.e. expanding until $y_i > c$, and sequentially placing the initial sample plots (Brown, 2003; Brown and Manly, 1998).

3.3 Line intersect sampling

While many environmental populations can naturally be represented as point objects, such as standing trees, or to be more precise, the seedling points of trees, others are better represented by other shapes, for example road networks and rivers. When the population can be represented by lines, and have values that can be measured as a function along these lines, line intersect sampling can be useful. In line intersect sampling, the sampling units consists of line segments, being placed randomly over the area frame. The line segments are traversed, and any crossings of the line segments and the population are recorded. In the simple case, consider a single line object i of length l_i placed inside an area frame $\mathscr{F}$, and a single sampling unit of length L placed uniformly with uniform rotation on $\mathscr{F}$. The probability of a crossing occurring, i.e. the line object being included in the sample, is

$$\mathrm{P}(S_i = 1) = \frac{2}{\pi}\frac{l_i L}{|\mathscr{F}|},$$

as stated by Matérn (1964). If n sampling units are placed independently on the area frame, the expected number of inclusions for the line object becomes

$$\mu_i = n\frac{2}{\pi}\frac{l_i L}{|\mathscr{F}|},$$

which can be shown to hold for line objects of any form, and/or linear sampling units of any form.

In Paper V, we apply line intersect sampling together with spatially balanced sampling in a two-phased approach in order to estimate the total volume of felled trees using aerial photographs. In a first phase sample, a large number of linear sampling units are placed in a systematic pattern, with random start and rotations, on the area frame, and auxiliary information is collected for each unit. From this first phase sample, a second phase sample is taken by a sampling design utilizing the collected auxiliary information. The second phase sample is surveyed, registering any crossings between the linear sampling units and any storm-felled trees.

Let y_i denote the volume of a storm-felled tree $i \in U$. The systematic first phase sample is taken by placing n_1 linear sampling units in a grid, spaced out

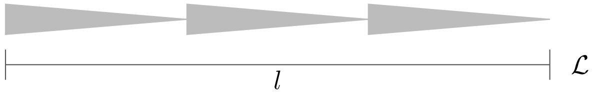


Figure 10: Representation of storm-felled trees as a function along a line.

by F_1, F_2 such that $n_1 F_1 F_2 = |\mathscr{F}|$, as U_1. The expected number of crossings of U by U_1 in this first phase sample is

$$\sum_{i \in U} \mu_i = \frac{n_1}{n_1} \frac{2}{\pi} \frac{L}{F_1 F_2} \sum_{i \in U} l_i = \frac{2}{\pi} \frac{lL}{F_1 F_2}, \tag{17}$$

where l is the total length of the storm-felled trees. An alternative representation of the total volume Y of the storm-felled trees is

$$Y = \int_{\mathscr{L}} y(z) dz,$$

where $y(z)$ denotes the cross-sectional area of the storm-felled trees, represented as a continuous line segment, see Figure 10.

Through the line segment representation, the sampling intensity of the crossings along the line segment can be deduced from (17) as

$$\pi(z) = \frac{2}{\pi} \frac{L}{F_1 F_2},$$

since the intensity is uniform along all storm-felled trees. Denote as $Z(i)$ the set of crossings on $\mathscr{L}$ between a first phase sampling unit $i \in U_1$ and the storm-felled trees U. A first phase estimator can then be given by

$$\hat{Y}_1 = \sum_{i \in U_1} \sum_{z \in Z(i)} \frac{y(z)}{\pi(z)}.$$

The second phase sample $\mathbf{S}$ is drawn from the first phase sample U_1 with inclusion probabilities π_i, resulting in a conditional HT-estimator

$$\hat{Y}_2 = \sum_{i \in U_1} \frac{\mathrm{I}(S_i > 0)}{\pi_i} \sum_{z \in Z(i)} \frac{y(z)}{\pi(z)}.$$

As with (16), this estimator is unbiased.

3.4 Combining surveys

Sometimes there exists a need for combining multiple, independent surveys. If domain estimates are needed of a subset of the population, a primary survey of the whole population might lack sufficient information about the subset to achieve acceptable levels of confidence, why complementary domain level surveys are needed. In other cases, one might want to incorporate existing data from several surveys into a larger survey. In either case, it would be unwise to discard already collected data.

A general strategy is to calculate the design properties for the combined design. Let D denote a set of independent designs, and let $S_i^{(d)}, \pi_i^{(d)}, \pi_{ij}^{(d)}, \mu_i^{(d)}, \mu_{ij}^{(d)}$ be the design properties for units $i, j \in U$ for a single design $d \in D$. Naturally, the number of inclusions S_i of the combined design is just the sum of the inclusions

$$S_i = \sum_{d \in D} S_i^{(d)}. \tag{18}$$

From this, we can deduce that the probability of inclusion for a unit i in the combined design becomes the complement to the probability of the unit not being included in any sample

$$\pi_i = 1 - \prod_{d \in D} \left(1 - \pi_i^{(d)}\right).$$

Similarly, for the second order inclusion probability, it can be shown that

$$\pi_{ij} = \pi_i + \pi_j - \prod_{d \in D} \left(1 - \pi_i^{(d)} - \pi_i^{(d)} - \pi_{ij}^{(d)}\right).$$

For a combined MC-design, the expected number of inclusions is

$$\mu_i = \sum_{d \in D} \mu_i^{(d)},$$

with the second order expected number of inclusions as

$$\mu_{ij} = \mu_i^{(d)} \mu_j^{(d)} + \sum_{d \in D} \left(\mu_{ij}^{(d)} - \mu_i^{(d)} \mu_j^{(d)}\right).$$

Using these combined design properties, it is possible to use the estimators (1) or (5) (Grafström et al., 2019).

An estimator for combining multiple surveys over the same frame is the linear combination of the form

$$\hat{Y} = \sum_{d \in D} \alpha_d \hat{Y}^{(d)}, \tag{19}$$

where α_d is the weight of design d, with $\sum \alpha_d = 1$. The values of α_d yielding the lowest variance of the linear combination is found by

$$\alpha_d = \mathrm{V}\left(\hat{Y}^{(d)}\right)^{-1} / \sum_{k \in D} \mathrm{V}\left(\hat{Y}^{(k)}\right)^{-1},$$

where $\mathrm{V}(\cdot)$ denotes the variance operator. As the variances are often not known, a strategy could be to use the estimators of the variances instead

$$\widehat{\alpha}_d = \hat{\mathrm{V}}\left(\hat{Y}^{(d)}\right)^{-1} / \sum_{k \in D} \hat{\mathrm{V}}\left(\hat{Y}^{(k)}\right)^{-1}.$$

The estimator (19) using $\widehat{\alpha}_d$ is unbiased if the design variance estimators are independent from the design estimators (Rubin and Weisberg, 1974).

For environmental surveys, it is rather unlikely that the estimators are independent from the variance estimators, as positive correlation is expected between the two (Grafström et al., 2019). As a simple example, if a survey only manages to sample bushes, the variance will probably be rather low, compared to a survey which only manages to sample in old growth forest, where the plants generally are taller with higher variation between them. A combination of these two surveys would promote the sample with only bushes, as the variance is lower.

In Paper I, we propose a pooled variance estimator to be used as weights, of the forms

$$\hat{\mathrm{V}}_p\left(\hat{Y}^{(d)}\right) = \sum_{i \in U} \sum_{j \in U} \frac{y_i y_j}{\pi_i^{(d)} \pi_j^{(d)}} \frac{\pi_{ij}^{(d)} - \pi_i^{(d)} \pi_j^{(d)}}{\pi_{ij}} \mathrm{I}(S_i > 0 \cap S_j > 0),$$

for HT-estimators, and

$$\hat{\mathrm{V}}_p\left(\hat{Y}^{(d)}\right) = \sum_{i \in U} \sum_{j \in U} \frac{y_i y_j}{\mu_i^{(d)} \mu_j^{(d)}} \frac{\mu_{ij}^{(d)} - \mu_i^{(d)} \mu_j^{(d)}}{\mu_{ij}} S_i S_j,$$

for HH-estimators, using all available information in estimating the variances. The results shows that this can help reduce the bias for the linear combination, however with an additional effort required in order to match sample units of different surveys.

4 Final remarks

Environmental studies have a long history of utilizing auxiliary variables, through the locations of the population units on the map. Through digitization and registers, more and more populations gain complete coverage of some information prior to the surveying effort. The remote sensing efforts provide new possibilities of having full coverage of new landscape attributes. The line transect design in Paper V provides a practical example of how to incorporate auxiliary information in the sampling phase, as well as utilizing remote surveying equipment.

The design in Paper II utilizes the auxiliary variables in order to provide a well-spread sample. When running a design utilizing auxiliary variables on a computer, two main constraints exist in time complexity and space complexity. For taking a single sample, space complexity is probably going to be the restricting factor, which would imply that sampling algorithms which do not rely on the computation of large matrices are favourable compared to those that do. Nevertheless, the development of faster algorithms remains important for the broad adaptation of spatially balanced or balanced sampling designs.

As more focus is put upon conservation and monitoring of important and protected habitats, there is an increased need for sampling designs able to efficiently capture these kind of populations (Adler et al., 2020). The design in Paper III can be useful if the population structure is clustered, and the linear combination of Paper I can be used if there exists data from multiple surveys. Nevertheless, further research is needed in order to provide accurate information as the level of detail required through legislation increases.

When auxiliary variables are available, there is no disadvantage in using them in the sampling design, as the worst case scenario implies effectiveness on par with SRS. The conclusion therefore seems to be that sampling designs should always make use of any auxiliary information, regardless of the aim of the study. Paper IV discusses the various ways to measure spatial balance.

If well-spread designs are able to provide miniatures of the population, such designs seem desirable even for model-based studies, as a way to reliably get a good coverage, of the population, and therefore minimize the risk of extrapolation.

References

Adler, S., Christensen, P., Gardfjell, H., Grafström, A., Hagner, A., Hedenås, H., and Ranlund, A. (2020). Ny design för riktade naturtypsinventeringar inom NILS och THUF. Technical Report 513, Sveriges lantbruksuniversitet, Instutitionen för skoglig resurshushållning, Umeå.

Benedetti, R. and Piersimoni, F. (2017). A spatially balanced design with probability function proportional to the within sample distance. *Biometrical Journal*, 59(5):1067–1084.

Bondesson, L. and Thorburn, D. (2008). A list sequential sampling method suitable for real-time sampling. *Scandinavian Journal of Statistics*, 35(3):466–483.

Brewer, K. (1994). Survey sampling inference: Some past perspectives and present prospects. *Pakistan Journal of Statistics*, 10(1):213–233.

Brown, J. A. (2003). Designing an efficient adaptive cluster sample. *Environmental and Ecological Statistics*, 10:95–105.

Brown, J. A. and Manly, B. (1998). Restricted adaptive cluster sampling. *Environmental and Ecological Statistics*, 5:49–63.

Cochran, W. G. (1977). *Sampling Techniques*. Wiley, New York, 3 edition.

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.

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

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 Fennica*, 48(3).

Grafström, A. (2012). Spatially correlated Poisson sampling. *Journal of Statistical Planning and Inference*, 142(1):139–147.

Grafström, A., Ekström, M., Jonsson, B. G., Esseen, P.-A., and Ståhl, G. (2019). On combining independent probability samples. *Survey Methodology*, 45(2):349–364.

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

Grafström, A., Lundström, N. L., and Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, 68(2):514–520.

Grafström, A. and Schelin, L. (2014). How to select representative samples. *Scandinavian Journal of Statistics*, 41(2):277–290.

Grafström, A., Schnell, S., Saarela, S., Hubbell, S., and Condit, R. (2017).

The continuous population approach to forest inventories and use of information in the design. *Environmetrics*, 28(8):e2480.

Grafström, A. and Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics*, 24(2):120–131.

Gregoire, T. G. (1998). Design-based and model-based inference in survey sampling: appreciating the difference. *Canadian Journal of Forest Research*, 28(10):1429–1447.

Gregoire, T. G. and Valentine, H. T. (2007). *Sampling Strategies for Natural Resources and the Environment*. Chapman & Hall/CRC Press, Boca Raton.

Hansen, M. H. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14(4):333–362.

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

Jauslin, R. and Tillé, Y. (2020). Spatial spread sampling using weakly associated vectors. *Journal of Agricultural, Biological and Environmental Statistics*, 25(3):431–451.

Kiær, A. N. (1976). *Den repræsentative undersøgelsesmethode : ny utgave med engelsk oversettelse ved Statistisk sentralbyrås 100-års jubileum 1976*. Statistisk sentralbyrå, Oslo.

Mandallaz, D. (2007). *Sampling Techniques for Forest Inventories*. Chapman & Hall/CRC Press, Boca Raton.

Matérn, B. (1964). A method of estimating the total length of roads by means of a line survey. *Studia Forestalia Suecica*, 18:68–70.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.

Robertson, B., Brown, J., McDonald, T., and Jaksons, P. (2013). Bas: Balanced acceptance sampling of natural resources. *Biometrics*, 69(3):776–784.

Royall, R. M. and Herson, J. (1973). Robust estimation in finite populations i. *Journal of the American Statistical Association*, 68(344):880–889.

Rubin, D. B. and Weisberg, S. (1974). The variance of a linear combination of independent estimators using estimated weights. *ETS Research Bulletin Series*, 1974(2):i–5.

Särndal, C.-E. (1978). Design-based and model-based inference in survey sampling. *Scandinavian Journal of Statistics*, 5(1):27–43.

Särndal, C.-E., Swensson, B., and Wretman, J. H. (2003). *Model Assisted Survey Sampling*. Springer, New York.

Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5(1194):127.

Stevens, D. L. J. and Olsen, A. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American statistical Association*, 99(465):262–278.

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

Talvitie, M., Leino, O., and Holopainen, M. (2006). Inventory of sparse forest populations using adaptive cluster sampling. *Silva Fennica*, 40(1):101.

The European Commission (2020). EU Biodiversity Strategy for 2030 (COM/2020/380).

The European Commission (2021). New EU Forest Strategy for 2030 (COM/2021/572).

Thompson, S. K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, 85(412):1050–1059.

Tillé, Y. (2006). *Sampling Algorithms*. Springer, New York.

Tillé, Y., Dickson, M. M., Espa, G., and Giuliani, D. (2018). Measuring the spatial balance of a sample: A new measure based on Moran's I index. *Spatial Statistics*, 23:182–192.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, 46(sup1):234–240.

Yates, F. (1946). A review of recent statistical developments in sampling and sampling surveys. *Journal of the Royal Statistical Society*, 109(1):12–43.

Yates, F. and 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. and Grafström, A. (2023). Estimation of change with partially overlapping and spatially balanced samples. *Environmetrics*, page e2825.

Popular science summary

In order to make informed decisions, one needs access to reliable data. For the decisions regarding climate change and conservation issues, information of the current state and changes in natural resources and habitats is needed.

Information about the unknown can be gained in multiple ways. If it is possible, one can survey everything deemed relevant. Oftentimes, the cost associated which a full scale survey is significant, and a sample of the whole is needed. In order to ensure objectivity, random samples are preferred over subjective samples, as it reduces the risk of the researcher influencing the outcome.

Sometimes information is known about the whole population prior to the sampling effort. In such cases, it makes sense to use this information when selecting the random sample. This can decrease the risk of getting a highly unrepresentative sample. For example, if the locations of the population units are known in advance, we rarely want to have a sample that is all clustered together on the map, as nearby things tend to be more similar compared to things far apart. Thus, a strategy can be to use a sample selection algorithm that spreads the sample on the map.

Furthermore, for different types of populations we need different types of tools in order to select a sample. When surveying the trees in a forest, one can randomly place plots in the landscape, and measure any trees inside the plots, as trees can be thought of as points on a map, if seen from above. If the population is rare, not enough plots may catch the phenomenon we are interested in, and some other sampling technique is needed in order to ensure that enough data can be gathered. Other combinations of populations and variables of interest may not even be suitable to survey using plot sampling, and yet different designs are needed.

This thesis explores different ways of selecting samples, focusing on the utilization of prior information. The different papers give new tools for selecting well-spread samples (Paper III), how to evaluate if a sample is well-spread

(Paper IV), how to sample from rare and clustered populations (Paper II), how to sample storm-felled trees using aerial drone photos (Paper V), and how to combine multiple surveys in order to increase the accuracy of our estimators (Paper I).

Populärvetenskaplig sammanfattning

För att kunna fatta välinformerade beslut krävs tillgång till pålitlig data. För beslut som rör klimatförändringar och naturvård krävs information om det nuvarande tillståndet samt om förändringar i naturresurser och arters utbredningsområden.

Ett sätt som kan användas för att få information är genom en totalundersökning, där man fullständigt undersöker en hel population. Ofta är dock kostnaden för en sådan totalundersökning betydande, vilket gör att en stickprovsundersökning är nödvändig. För att säkerställa objektiviteten i en stickprovsundersökning behöver man göra ett slumpmässigt urval, då den som genomför undersökningen annars riskerar att påverka resultatet.

Ibland finns det information om enheterna i populationen innan urvalet tas. I sådana fall är det möjligt att använda denna information när man genomför det slumpmässiga urvalet, för att minska risken för att ett skevt urval väljs. Som exempel, om vi vet positionerna för enheterna i populationen i förväg så vill vi sällan ha ett stickprov där alla enheter är väldigt nära varandra. Detta eftersom närliggande objekt tenderar att vara mer lika varandra jämfört med objekt som är långt ifrån varandra. En strategi kan därför vara att använda en urvalsalgoritm som sprider ut urvalet på kartan.

Vidare behövs olika typer av samplingverktyg för olika typer av populationer. När man undersöker träd i en skog kan man slumpmässigt placera ut provytor i landskapet, och sedan mäta eventuella träd som hamnat inom provytorna, då träd kan betraktas som punkter om de ses uppifrån. För populationer där enheterna är sällsynt förekommande så riskerar dock inte tillräckligt många provytor att träffa av det fenomen vi är intresserade av, och andra tekniker behövs för att säkerställa att vi får tillräckligt med data för att kunna dra säkra slutsatser. För andra typer av populationer och variabler kan det vara så att provytor inte lämpar sig alls, varför ytterligare andra metoder behövs.

Denna avhandling behandlar olika sätt att välja urval med fokus på nyttjande av hjälpinformation. De olika studierna presenterar nya verktyg för att ta spridda urval (Studie III), hur man utvärderar om ett stickprov har en bra spridning (Studie IV), hur man väljer urval från sällsynta och klustrade populationer (Studie II), hur man väljer urval av stormfällda träd med hjälp av flygfoton från drönare (Studie V) och hur man kombinerar resultaten flera undersökningar för att öka noggrannheten för en skattning (Studie I).

Acknowledgements

I would like to start to thank my main supervisor Anton Grafström. Thank you for taking me on, guiding, and believing in me. Thank you for being challenging but open, giving and having time. I feel privileged to have had the opportunity to work with and learn from such a skilled, enthusiastic, and friendly statistician and scientist. I would also like to thank my co-supervisors Cornelia Roberge and Magnus Ekström for the support and guidance I received along the way, and for your patience in answering my sometimes unstructured questions. A special thanks to Xin, who welcomed me into the group and showed the way forward.

Thank you Hans Petersson, for the support I've received from the department. Thank you to Hilda and Léna for the sense of community. Also a big thanks to all my former and present colleagues.

A fool, a fool, I met a fool in the forest! To my fellow PhD students at the department: Thanks for the laughs and the company. Because of you, the forest has felt less scary.

Thank you to my childhood friends, for giving me so many valuable insights in life and statistics: Ivar that independence is something declared, not gained. John that regardless of the problem presented, no challenge should go unaccepted. Klintan that there is no interest too obscure to be disregarded.

Leo, Pontus, Fredde, Erik, Eric and Peter, thanks for making studying both bearable and unbearable. Thank you for the pancakes: big pharma Robin, business economist GAE, bernie Bore, Carl Lars Karlsson Larsson, paint magnate Hanna, archmaster Arvid, comedic Christian, long-legged Raul.

To my family: my sisters Carolina and Fanny, my cousins Ella and Harry, and my uncle Curt. Thank you for the support I've received throughout my life. Without you, I would not have made it to where I am today. The gratitude I feel towards you all cannot be understated, and will always stay with me.

Finally, to my girlfriend Sofia, thank you for your patience, support and

determination, especially during the more stressful times of writing this thesis.

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

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

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

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

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

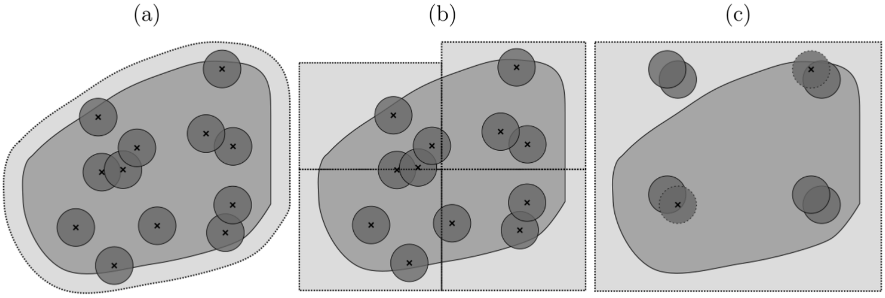


Figure 1. Examples of **a** i.i.d., **b** stratified, and **c** systematic frames and inclusion zones. The outer areas represent the sample frames (F), the inner areas represents the areas of interest (F_U), and the circles represents the inclusion zones (A) for units. In both **a** and **b**, the sample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

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

where S_i denotes the number of inclusions of unit i, $\pi_i = \Pr(S_i > 0)$ denotes the inclusion probability for unit i, i.e., the probability for unit i to be included in the sample, and $\mathrm{I}(\cdot)$ 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

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

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 $E_i = \mathrm{E}\left[S_i\right]$ 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

$$\begin{aligned}\pi_{ij}^{(\mathcal{D})} =& \pi_i^{(\mathcal{D})} + \pi_j^{(\mathcal{D})} \\ &- \left(1 - \prod_{P_d\in\mathcal{D}}\left(1 - \pi_i^{(P_d)} - \pi_j^{(P_d)} + \pi_{ij}^{(P_d)}\right)\right), \qquad (17)\\ E_i^{(\mathcal{D})} =& \sum_{P_d\in\mathcal{D}} E_i^{(P_d)}, \qquad (18)\end{aligned}$$

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

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

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

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

with sufficient counterpart

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

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

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

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

4.1. Combining Estimators by Linear Combinations

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

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

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

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

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

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

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

where

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

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

5. SIMULATION

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

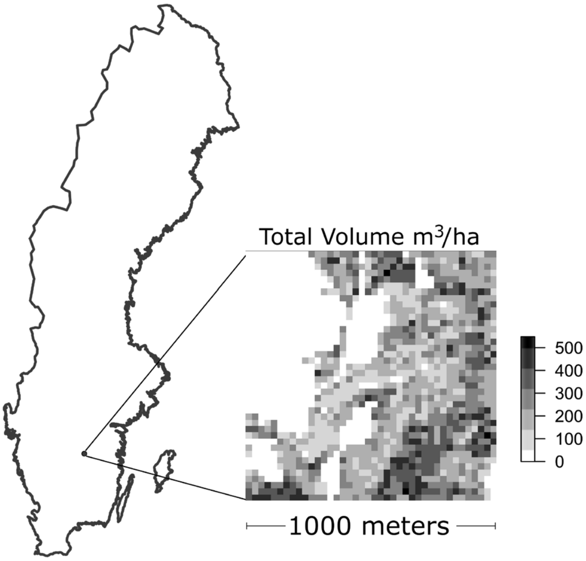


Figure 2. Location and the total biomass volume (m^3/ha) for the area used as a boilerplate for simulating the population. Darker colors indicate higher volumes (Color figure online).

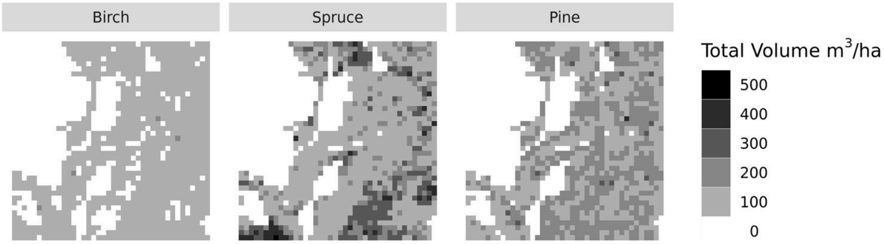


Figure 3. Total biomass volume (m^3/ha) per species for the simulated population. Darker colors indicate higher volumes (Color figure online).

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

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

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

Table 1. Sample designs used in the simulation study

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

n Sample size; *Radius* Radius of inclusion zones

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6. DISCUSSION

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

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

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

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

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

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

Another scenario where both presented methods might be useful are when combining designs like those used in the simulation here, where it is not possible to get an unbiased variance estimator for one or more of the individual designs. The pooled combination is unbiased if the combined second-order sample properties are positive for all units in the population, whereas the naive combination needs positive second-order sample properties for all units and all designs. Furthermore, the combined samples approach has none of these restrictions and is also more relaxed in terms of first-order sample properties.

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

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

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

ACKNOWLEDGEMENTS

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

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

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

APPENDIX: UNIT DESIGN PROPERTIES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

and conversely, apparent from (33), we have

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

leading to

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

Received: 30 June 2021 | Revised: 9 February 2022 | Accepted: 9 April 2022

DOI: 10.1002/env.2729

RESEARCH ARTICLE

WILEY

Two-phase adaptive cluster sampling with circular field plots

Wilmer Prentius | **Anton Grafström**

Department of Forest Resource Management, Swedish University of Agricultural Sciences, Umeå, Sweden

Correspondence
Wilmer Prentius, Department of Forest Resource Management, Swedish University of Agricultural Sciences, SE-90183 Umeå, Sweden.
Email: wilmer.prentius@slu.se

Abstract

Adaptive cluster sampling (ACS) is extended to the case when the primary sampling units consist of circular field plots. When conducting field work for environmental monitoring, circular field plots are often preferred as they are easily set up by field workers. ACS was developed by tessellating the area frame into square plots. By using a two-phase sampling procedure, a first-phase sample of circular field plots can be established as the primary sampling units, from which ACS can be performed. However, the two-phase approach introduces some additional complexity in estimation. We derive estimators and conservative variance estimators for two-phase ACS using circular field plots. For some populations, ACS may produce a highly variable sample size. To deal with this issue, we provide a way to reduce the maximal possible sample size. By using simulated populations, we compare the efficiencies of two-phase methods with ordinary simple random sampling. The simulations show that the two-phase approach is a competitive alternative to regular ACS, and that adding a restriction to the maximal possible sample size makes ACS a viable alternative for a larger set of populations.

KEYWORDS

adaptive sampling, area frame sampling, design-based sampling, environmental monitoring, spatially clustered populations

1 | INTRODUCTION

For environmental surveys, we often have some well-defined area on which the population of interest is located. However, we rarely know the number of units in the population or their exact locations within the area. In such a situation, it is not feasible to construct a comprehensive list frame to select a sample. Instead, environmental surveys are often conducted using an area frame.

A popular way to select a sample from the population is then through plot sampling, where fieldworkers collect data from circular sample plots randomly placed on the area frame. Examples of such sampling efforts are national forest inventories, forest damage inventories, and landscape inventories. The use of circular field plots eases the field effort, as circles are easy to emulate using a rope or a stick of a fixed length around the center point, whereas other shapes are harder to recreate with accuracy.

If the population units are spread out and can be found in a large part of the area, then it is preferable to make sure that the field plots are also well spread in the area. By doing so, more of the variation is captured and less redundant information is collected. The reason for this comes from Tobler's first law of geography (Tobler, 1970): "everything is related to everything else, but near things are more related than distant things."

Another implication of Tobler's first law of geography would be that if gold is struck, we expect more to be nearby. Thus, for rare populations, that is, populations for which it is hard to find sufficiently large samples through conventional sampling methods, it may be advantageous to increase a sample with nearby units, even though the additional units are slightly less informative.

One dynamical sampling effort, where nearby plots are sampled if the primary plots recorded any of the features of interest, is the adaptive cluster sampling (ACS) (Thompson, 1990). In ACS, the primary sampling units are the squares created by superimposing a regular grid over the area frame. Whenever a sample unit matches some criteria, all nearby units are included in the sample, and if any of these match the criteria, its neighbors are added until the cluster has been fully included in the sample. In Section 2, the ACS method is reiterated.

This article presents a further development of the ACS, where the sampling units consist of a lattice of circular field plots over the area of interest, instead of rectangles, in order to reduce the in-field effort. As any lattice of circles cannot fully cover an area frame, positive inclusion probabilities are ensured by deciding the lattice through a first-phase sample. This approach will increase the variance of the estimator. However, the variance added will tend to be negligible and is likely to be fully compensated for by a reduced cost, as a wall-to-wall rectangular lattice of circles would correspond to about 80% of the frame. A similar method has been proposed before, but without determining the effects of the additional sampling phase or providing a rigorous theory (Talvitie et al., 2006).

The final sample size of ACS is random, which is an undesirable property in many environmental surveys or monitoring programs. There have been some attempts to reduce the variability of the final sample size for ACS. Among others, Brown and Manly (1998) proposed a method that sequentially updates the initial sample with new plots if the total sample size is below some threshold, however, without providing an unbiased estimator. Salehi and Seber (1997) showed that by stratifying the primary plots, the final sample size can be reduced by not allowing expansion of plots over the stratum boundaries. In this article, we propose to reduce the variability in the final sample size through restricting how far away from the initial sample plots the expansion is allowed, which we present together with the general theory of the proposed two-phase ACS in Section 3.

In Section 4, the methods are evaluated using simulated populations with different spatial structures. Finally, we discuss the results in Section 5.

2 | ADAPTIVE CLUSTER SAMPLING

A region F is tessellated into N square plots of size λ^2 as S_1. F contains some finite population of units, with some variable of interest y, assumed to be clustered together into one or more clusters, with a total of Y over F. Let y_i be the plot total of y over a plot i, making it possible to express the total as

$$Y = \sum_{i \in S_1} y_i.$$

Let $C(i)$ denote a cluster of plots including plot i. There exists an indicator of the fulfillment of some kind of conditions or criteria on plot i. If the criteria of plot i are not fulfilled, then $C(i)$ includes only plot i. Otherwise, $C(i)$ includes plot i together with all connected neighbors also fulfilling the criteria, where a neighbor is defined by some measure of distance. In Figure 1, the concept of a cluster is shown.

From S_1, a sample of plots is drawn without replacement as S_{2a}, with first- and second-order inclusion probabilities

$$\pi_i = \Pr(i \in S_{2a}), \qquad \pi_{ij} = \Pr(i \in S_{2a}, j \in S_{2a}).$$

Let M_i denote the number of inclusions of a plot i, either directly by itself being included into S_{2a}, or indirectly through the inclusion of any other plot in $C(i)$ into S_{2a}. Furthermore, let E_i be the first-order expected number of direct or indirect inclusions of a plot i, such that

$$M_i = \sum_{k \in C(i)} I(k \in S_{2a}), \qquad E_i := \sum_{k \in C(i)} \pi_k.$$

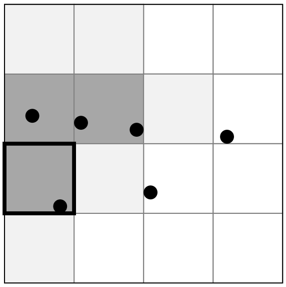


FIGURE 1 Cluster (green) defined around the plot with bold borders, with a criteria for inclusion $y_i > 0$. Gray plots are plots that will be checked during the sampling effort to determine if they belong to the cluster. Black dots represent some objects with positive values y.

From this, the Hansen-Hurwitz estimator (Hansen & Hurwitz, 1943) can be expressed as

$$\hat{Y} = \sum_{i \in S_2} \frac{y_i}{E_i} M_i, \tag{1}$$

where $S_2 = \cup_{i \in S_{2a}} C(i)$. The variance of (1) is

$$V(\hat{Y}) = \sum_{i \in S_2} \sum_{j \in S_2} \frac{y_i}{E_i} \frac{y_j}{E_j} (E_{ij} - E_i E_j),$$

which can be estimated through

$$\hat{V}(\hat{Y}) = \sum_{i \in S_2} \sum_{j \in S_2} \frac{y_i}{E_i} \frac{y_j}{E_j} \frac{E_{ij} - E_i E_j}{E_{ij}} M_i M_j,$$

where the second-order expected number of direct or indirect inclusions of plots i, j is

$$E_{ij} := E\left(M_i M_j\right) = \sum_{k \in C(i)} \sum_{l \in C(j)} \pi_{kl}.$$

3 | TWO-PHASE ADAPTIVE CLUSTER SAMPLING

Let S_1 be a sample of N circular plots i of radius r, centered on $\mathbf{x}_i$. The plots in S_1 are taken as a systematic sample from a region F, such that the plot centers are separated by $\lambda \geq 2r$. Note that N might be random, for certain shapes of F.

The region F contains some finite population of units U, with some variable of interest y', assumed to be clustered together in one or more clusters, with a total of Y over F. Any unit $j \in U$ is inside a circular plot i if $\mathbf{x}_i$ falls inside its inclusion zone K_j of size $|K_j|$. Let $y(\mathbf{x})$ denote the per-area plot mean of y' on a circular plot centered around $\mathbf{x}$, such that

$$y(\mathbf{x}) = \sum_{j \in U} \frac{I_{K_j}(\mathbf{x})}{|K_j|} y'_j,$$

where $I_{K_j}(\mathbf{x})$ is equal to 1 if $\mathbf{x} \in K_j$, 0 otherwise. If some buffer of at least r is added around F, then $|K_j| = r^2\pi$ for all units $j \in U$. As shown by Grafström et al. (2017), the population parameter Y can thus be expressed as

$$Y = \int_F y(\mathbf{x}) d\mathbf{x} = \sum_{j \in U} y'_j.$$

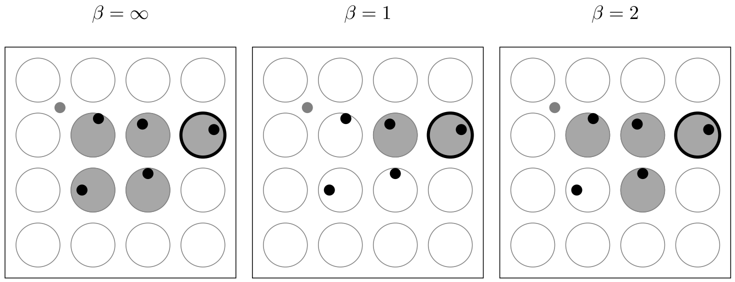


FIGURE 2 Clusters defined around the plot with bold borders with unrestricted (left) and restricted (middle, right) expansion, with a criterion for inclusion $y_i > 0$.

As the sampling intensity of any point $\mathbf{x}$ is $\pi(\mathbf{x}) = \lambda^{-2}$, a Horvitz–Thompson (HT) estimator (Cordy, 1993) of the total Y is

$$\hat{Y}_1 = \sum_{i \in S_1} \frac{y(\mathbf{x}_i)}{\pi(\mathbf{x}_i)} = \sum_{i \in S_1} y_i, \tag{2}$$

where $y_i = y(\mathbf{x}_i)/\pi(\mathbf{x}_i)$.

Given the sample S_1, let $\mathcal{U}_C$ denote the set of clusters C_i, where each cluster C_i is defined around and includes plot $i \in S_1$. There exists an indicator of the fulfillment of some kind of conditions or criteria on plot i.

If the criteria of inclusion of plot i are not fulfilled, then C_i contains only i. If the criteria of inclusion of plot i are fulfilled, then the cluster C_i contains all connected neighbors also fulfilling the criteria, where a neighbor is defined by some measure of distance. Furthermore, the expansion can be restricted, such that a cluster C_i never expands further away than β plots from the initial plot i, measured by some node distance. In this article, the node distance measure is exemplified by using the minimum number of plots traversed, travelling horizontally and/or vertically, in order to reach another plot. Figure 2 shows such unrestricted (i.e., $\beta = \infty$) and restricted clusters. It can be noted that if unrestricted expansion is allowed, $C_i = C_j$ if $j \in C_i$.

From S_1, a second sample of plots is drawn without replacement as S_{2a}. According to the design from which S_{2a} is selected, the first- and second-order inclusion probabilities for plots i, j to be included into S_{2a} given S_1 are denoted as

$$\pi_i = \Pr(i \in S_{2a} | S_1), \qquad \pi_{ij} = \Pr(i \in S_{2a}, j \in S_{2a} | S_1).$$

Let M_i be the number of times a plot i is included through any cluster sampled by S_{2a} conditional on S_1. M_i can thus be expressed as

$$M_i = \sum_{k \in S_{2a}|S_1} I(i \in C_k) = \sum_{k \in C_i} I(k \in S_{2a}|S_1).$$

since $i \in C_k$ implies $k \in C_i$.

Let S_2 denote the set of plots directly or indirectly included by S_{2a}, that is, $S_2 = \cup_{k \in S_{2a}} C_k$. The expected number of direct or indirect inclusions of a plot i through S_{2a}, given the first sample S_1, is

$$E_i := E(M_i | S_1) = \sum_{k \in C_i} \pi_k, \tag{3}$$

that is, the sum of the individual inclusion probabilities of the plots in the cluster to which plot i belongs.

An estimator of Y is thus given as

$$\hat{Y}_2 = \sum_{i \in S_2} \frac{y_i}{E_i} M_i, \tag{4}$$

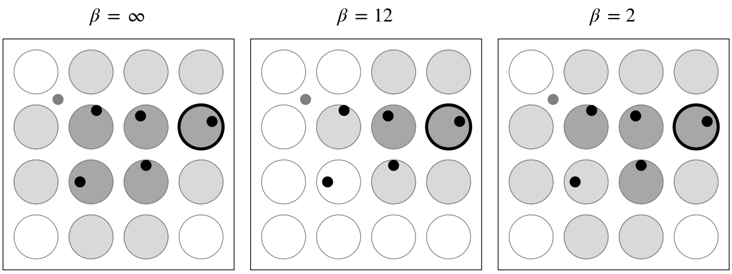


FIGURE 3 Sampling effort after a plot has been selected by S_{2a} for unrestricted (left) and restricted expansion (middle, right), with a criterion for inclusion $y_i > 0$. Green plots are inventoried completely, gray plots are only recorded on their criteria for inclusion.

which can be shown to be unbiased through the law of total expectation

$$E\left(\hat{Y}_2\right) = E_1\left(E_2\left(\hat{Y}_2|S_1\right)\right) = E_1\left(\hat{Y}_1\right) = Y.$$

By using a design with independent observations, it is possible to construct $\hat{Y}_2$ as a HT estimator (see Appendix). For a general design it becomes more difficult to derive inclusion probabilities of first and second order.

The sampling process of unrestricted and restricted sampling will record the variable of interest of all plots in C_k : $k \in S_{2a}$, while also needing to explore (however, not necessarily recording) all clusters defined around the plots in C_k, see Figure 3. The process can be summarized as the following:

(a) Randomly place a systematic grid of plots (S_1).
(b) Through some design with known first-order inclusion probabilities, select sample S_{2a} without replacement from S_1.
(c) For each plot in S_{2a} (the initial plot):
(d) Record if the criteria of inclusion are fulfilled;
(e) If (d) is true, or if the plot is the initial plot, record the variable of interest;
(f) If (d) is true, and the distance to the initial plot is less than β, go to all neighboring plots and repeat from (d);
(g) If the distance to the initial plot is equal to β (this plot is a max-dist plot), then:
    i. if the criteria of inclusion is fulfilled, go to all neighbors;
    ii. record if the criteria of inclusion is fulfilled;
    iii. if the distance to the max-dist plot is less than β, repeat from *i*.

3.1 | Conditional variance

The conditional variance of the estimator (4) is

$$V_2\left(\hat{Y}_2|S_1\right) = \sum_{i \in S_1} \sum_{j \in S_1} \frac{y_i}{E_i} \frac{y_j}{E_j} (E_{ij} - E_i E_j), \tag{5}$$

where the second-order expected number of direct or indirect inclusions of two plots i, j through S_{2a} is

$$E_{ij} := E\left(M_i M_j | S_1\right) = \sum_{k \in C_i} \sum_{l \in C_j} \pi_{kl}.$$

The conditional variance (5) can be unbiasedly estimated through

$$\hat{V}_2\left(\hat{Y}_2|S_1\right) = \sum_{i \in S_1} \sum_{j \in S_1} \frac{y_i}{E_i} \frac{y_j}{E_j} \frac{E_{ij} - E_i E_j}{E_{ij}} M_i M_j, \tag{6}$$

if $E_i > 0, E_{ij} > 0$ for all plots i, j.

If the sampling design used for S_{2a} does not have known or universally positive second-order expected number of inclusions, such as a systematic design or a design using auxiliary information, a more conservative variance estimator can in many cases be found by assuming a proportional to size (PPS) or simple random sample with replacement (SRSWR) design (Stevens & Olsen, 2003). In such cases, the following replacements

$$\begin{aligned} M_i^* &= \sum_{k \in C_i} \sum_{l \in S_{2a}|S_1} I(k = l), \\ E_i^* &= n \sum_{k \in C_i} p_k, \\ E_{ij}^* &= \frac{n-1}{n} E_i^* E_j^* + n \sum_{k \in C_i \cap C_j} p_k, \end{aligned}$$

can be used in (6), where p_k denotes the single-draw probability of selecting plot k into S_{2a}, that is, $p_k = N^{-1}$ in case of SRSWR.

3.2 | Variance

The variance of the estimator $\hat{Y}_2$ can be partitioned into two parts

$$V\left(\hat{Y}_2\right) = E_1\left(V_2\left(\hat{Y}_2|S_1\right)\right) + V_1\left(\hat{Y}_1\right), \tag{7}$$

where E_d, V_d denotes the expectation and variance under respective phase d, which can be considered as the variance stemming from the two sampling phases. When the second-order sampling intensity for two points $\mathbf{x}, \mathbf{x}^*$ is strictly positive, an unbiased estimator of $V_1(\hat{Y}_1)$ is given by

$$\hat{V}_1\left(\hat{Y}_1\right) = \sum_{i \in S_1} y_i^2 + \sum_{i \in S_1} \sum_{j \in S_1 \backslash i} y_i y_j \frac{\pi(\mathbf{x}_i, \mathbf{x}_j) - \pi(\mathbf{x}_j)\pi(\mathbf{x}_j)}{\pi(\mathbf{x}_i, \mathbf{x}_j)}. \tag{8}$$

As not all plots in S_1 are observed in the proposed sampling scheme, a conditional estimator must be used

$$\hat{\hat{V}}_1\left(\hat{Y}_1\right) = \sum_{i \in S_2} \frac{y_i^2}{E_i} M_i + \sum_{i \in S_2} \sum_{j \in S_2 \backslash i} y_i y_j \frac{\pi(\mathbf{x}_i, \mathbf{x}_j) - \pi(\mathbf{x}_j)\pi(\mathbf{x}_j)}{\pi(\mathbf{x}_i, \mathbf{x}_j) E_{ij}} M_i M_j. \tag{9}$$

If indeed $\pi(\mathbf{x}_i, \mathbf{x}_j) > 0$ for all pairs of points, the estimator

$$\hat{V}\left(\hat{Y}_2\right) = \hat{V}_2\left(\hat{Y}_2|S_1\right) + \hat{\hat{V}}_1\left(\hat{Y}_1\right), \tag{10}$$

is an unbiased estimator of (7), as

$$\begin{aligned} E_1\left(E_2\left(\hat{V}_2\left(\hat{Y}_2|S_1\right)\middle|S_1\right)\right) &= E_1\left(V_2\left(\hat{Y}_2|S_1\right)\right), \\ E_1\left(E_2\left(\hat{\hat{V}}_1\left(\hat{Y}_1\right)\middle|S_1\right)\right) &= V_1\left(\hat{Y}_1\right). \end{aligned}$$

However, as S_1 is selected as a systematic sample, the second-order sampling intensity for two points $\mathbf{x}, \mathbf{x}^*$ will not be universally positive if $\lambda > 2^{1.5} r$. By assuming a SRSWR design, a most likely conservative estimator can be found by using

$$\pi^*(\mathbf{x}, \mathbf{x}^*) = \frac{N(N-1)}{(N\lambda^2)^2}, \qquad \pi^*(\mathbf{x}) = \pi(\mathbf{x}),$$

thus in (10) substituting $\hat{\hat{V}}_1\left(\hat{Y}_1\right)$ for

$$\hat{\hat{V}}_1^*\left(\hat{Y}_1\right) = \sum_{i \in S_2} \frac{y_i^2}{E_i} M_i - \frac{1}{N-1} \sum_{i \in S_2} \sum_{j \in S_2 \backslash i} \frac{y_i y_j}{E_{ij}} M_i M_j. \tag{11}$$

3.3 | Cost of sampling effort

Thompson (1990) propose a cost equation that is linear, cost = $c_0 + c_1 n + c_2 n_2$, where c_0 denotes a fixed cost, c_1, c_2 and n, n_2 denotes the marginal costs and sizes of the initial second-phase sample S_{2a} and the subsequently sampled plots respectively. Furthermore, Thompson (1990) notes that the marginal cost for sampling a plot not satisfying the criteria of inclusion might be lower compared to conducting a full inventory of a plot.

The variable costs primarily associated with environmental surveys are mostly related to labor costs. As such, the marginal costs for an ACS design could be divided into time spent travelling to a survey site (i.e., a plot), time spent identifying if the criteria of inclusion of a plot are fulfilled, and time spent recording the variable of interest. For ACS, travelling between plots in and around a cluster could be considered negligible or included in the cost of checking the criteria of a plot. Thus, the cost for ACS could be modified as

$$\text{cost} = c_0 + c_{2a} n + c_I n_I + c_c n_c, \tag{12}$$

where c_{2a} is the cost associated with travelling to each plot in S_{2a}, c_I is the cost associated with recording the variable of interest on the n_I plots that need to be inventoried, and c_c is the cost associated with resolving the criteria of inclusion for the n_c plots visited.

Given a fixed initial ACS sample n_{2a}, the sample size of a SRS design can with equal expected cost, can be derived as

$$n_{SRS} = \frac{c_{2a} n + c_I E(n_I) + c_c E(n_c)}{c_{2a} + c_I E(p) + c_c},$$

where $E(p)$ is the expected proportion of plots fulfilling the criteria of inclusion in a sample S_2 (see Appendix for further details).

4 | SIMULATION

To evaluate the proposed methods, a simulation was conducted. Six populations were simulated according to a Poisson cluster process similar to those used by both Thompson (1990) and Christman (1997). In this process, λ_p parent locations

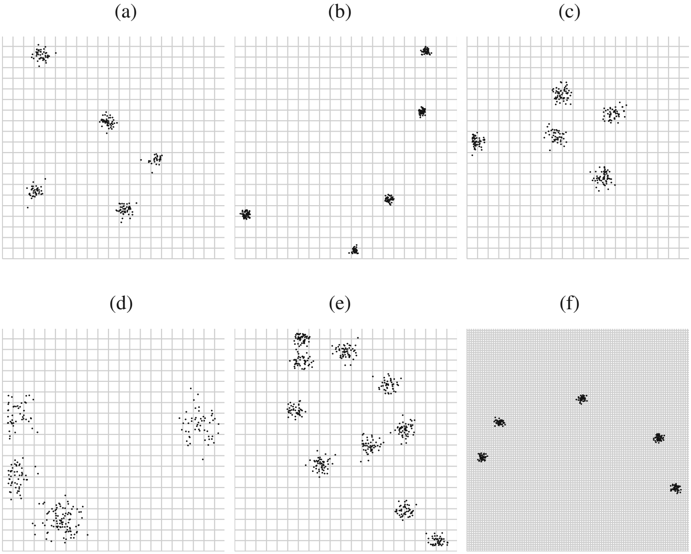


FIGURE 4 Artificial populations on area frames of (200 m)^2 (a–e) and (1000 m)^2 (f). Reference grids are shown at 10 m intervals. Superpopulation parameters as ($\lambda_p, \lambda_c, \sigma$). (a) (5, 40, 4); (b) (5, 50, 2); (c) (5, 50, 5), (d) (5, 50, 10), (e) (10, 50, 5), (f) (5, 50, 10)

are randomly placed uniformly over a square area frame. Each parent location spawns a random number of children according to a Poisson distributed with mean λ_c, where each child is placed relative to its parent following a bivariate normal distribution with variance $\sigma^2\mathbf{I}$.

Any objects falling outside of the area frame were mirrored back onto the frame, and a buffer of size $r = 5$ m was placed around the frame to ensure equal inclusion probabilities for all objects. In Figure 4, the six artificial populations are shown. The artificial populations were created using R 4.0.4, and the simulations were conducted using Go 1.17.6.

The parameters defining the superpopulations, see Figure 4, were chosen similar to those used in Thompson (1990) and Christman (1997), scaled to the areas used.

For each population, a systematic grid of square plots was randomly placed over the area frame with distance $\lambda = 10$ m as the sampling units for the ACS. Note that this is slightly different from Thompson (1990), as the primary sampling units are random. However, with little importance in practice, as most grids for a single survey could probably be considered random. The area frame was extended by a buffer of λ in both directions, in order to ensure equal inclusion probabilities for all units. The center point of each square were also considered as the center point for the circular field plot of radius $r = 5$ m, which constitutes the first sample S_1 for the Two-phase ACS with circular field plots (2PACS(β)). This process was repeated 100,000 times.

For each iteration, samples were selected for ACS, 2PACS(∞), and 2PACS(1). Initial samples S_{2a} for the designs were selected using SRSWOR and the Local Pivotal Method (LPM), where LPM is used to achieve geographically well-spread samples (Grafström et al., 2012). The size n of the initial samples S_{2a} were chosen so that around 90% of the initial samples had at least one plot satisfying the criteria of inclusion. The mean number of inventoried and visited plots for each of the methods are provided in Table 1, together with the mean of the variance estimates relative to the empirical variance, over the 100,000 iterations.

TABLE 1 The mean number of inventoried (inv.) and visited (vis.) plots, and the mean of the variance estimates relative to the empirical variance (RV) for ACS and 2PACS (unrestricted and restricted), using two strategies (SRS, LPM) for selecting the initial sample S_{2a} of size n.

			SRS			LPM		
Pop.	**Method**	***n***	**Inv.**	**Vis.**	**RV**	**Inv.**	**Vis.**	**RV**
(a)	ACS	35	12.3	62.2	1.00	14.0	67.0	1.09
	2PACS(∞)	35	9.8	57.4	1.18	11.1	61.2	1.30
	2PACS(1)	35	5.8	50.6	1.16	6.2	52.9	1.25
(b)	ACS	65	6.2	79.4	1.00	7.0	82.2	1.11
	2PACS(∞)	65	4.7	76.4	1.27	5.1	78.4	1.38
	2PACS(1)	65	4.0	75.5	1.27	4.3	77.2	1.38
(c)	ACS	30	15.2	61.8	1.00	17.4	67.4	1.11
	2PACS(∞)	30	12.4	57.0	1.14	14.1	61.5	1.26
	2PACS(1)	30	6.7	47.2	1.13	7.1	49.6	1.22
(d)	ACS	15	32.4	70.2	1.01	37.7	79.6	1.19
	2PACS(∞)	15	26.2	61.5	1.05	30.3	69.2	1.23
	2PACS(1)	15	7.1	32.5	1.04	7.4	34.2	1.23
(e)	ACS	15	20.2	56.4	1.01	22.6	61.8	0.96
	2PACS(∞)	15	16.0	48.8	1.07	17.6	53.0	1.05
	2PACS(1)	15	6.9	33.4	1.06	7.1	35.0	1.05
(f)	ACS	250	32.4	310.0	1.00	37.4	320.0	1.14
	2PACS(∞)	250	26.4	300.5	1.04	30.1	308.2	1.17
	2PACS(1)	250	6.8	268.1	1.03	7.1	269.5	1.20

Note: See Figure 4 for the populations.

TABLE 2 Efficiencies (Eff), MSE(*)/MSE(SRS), of the methods when compared to cost-relative sample sizes of an ordinary SRS, using two strategies (SRS, LPM) for selecting the initial sample S_{2a} of size n.

					SRS		LPM	
Pop.	**Method**	n	n_1	n_2	**Eff**$_1$	**Eff**$_2$	**Eff**$_1$	**Eff**$_2$
(a)	ACS	35	63	47	0.90	0.64	0.67	0.48
	2PACS(∞)	35	58	44	0.84	0.62	0.65	0.48
	2PACS(1)	35	51	40	0.83	0.64	0.66	0.50
(b)	ACS	65	80	71	0.75	0.65	0.60	0.52
	2PACS(∞)	65	77	69	0.77	0.68	0.66	0.58
	2PACS(1)	65	76	68	0.78	0.67	0.67	0.58
(c)	ACS	30	62	44	1.12	0.76	0.83	0.56
	2PACS(∞)	30	57	41	1.04	0.73	0.79	0.55
	2PACS(1)	30	48	36	0.99	0.72	0.78	0.56
(d)	ACS	15	71	44	3.07	1.77	2.02	1.17
	2PACS(∞)	15	62	39	2.81	1.67	1.92	1.14
	2PACS(1)	15	33	21	1.82	1.12	1.30	0.80
(e)	ACS	15	57	33	2.02	1.12	1.70	0.94
	2PACS(∞)	15	49	29	1.77	1.00	1.52	0.86
	2PACS(1)	15	34	21	1.38	0.83	1.22	0.74
(f)	ACS	250	310	283	0.77	0.70	0.54	0.49
	2PACS(∞)	250	301	277	0.77	0.71	0.55	0.51
	2PACS(1)	250	269	256	0.86	0.82	0.64	0.61

Note: Eff$_1$ compares against an SRS of size n_1, using marginal cost parameters $c_{2a} = c_I = 0, c_c = 1$, that is, n_1 is decided solely by the number of visited sample units. Eff$_2$ compares against an SRS of size n_2, using marginal cost parameters $c_{2a} = c_I = 1, c_c = 0.1$. See Figure 4 for the populations.

The results from the simulations were then evaluated against an ordinary SRS of comparable expected cost, and the relative efficiency of each strategy is presented in Table 2. The cost function of Thompson (1990) considered each visited unit as equally expensive, independent of whether or not the plot was inventoried. However, as is mentioned in Section 3.3, it is probably reasonable to expect that it is less expensive to sample an empty sample unit. Thus, two cost functions were used, one reflecting the conservative cost function of Thompson (1990), whereas the other considered (12) with marginal cost parameters $c_{2a} = c_I = 1, c_c = 0.1$.

Results in Table 2 highlight the effects of the cost function used to compare, as the 2PACS(β) methods are increasingly competitive as the cost of just visiting a plot goes down. Furthermore, the results show that as clusters become larger or more spread out, the usefulness of any ACS design decreases. This effect can be countered to an extent by restricting the sizes of the sampled clusters.

5 | DISCUSSION

As is often the case with sampling, the optimal strategy is heavily dependent on the population of interest. Cluster sampling can be shown to be beneficial if one is expecting the population to be grouped into small and relatively few clusters. As the clusters increase in size or abundance, cluster sampling ceases to be useful.

In inventories where the marginal cost of visiting empty plots is expected to be relatively small in comparison to the marginal costs of travelling between sites or inventorying nonempty plots, cluster sampling can be competitive for populations with larger or more clusters. However, this does not take into account the increased administrative cost of going for a more complex sampling strategy.

Circular field plots are the standard of many environmental surveys, especially in forestry, as they are easily established by field workers. Compared to a grid of square plots, circular plots naturally exclude part of the population, which has the effect of creating smaller clusters on average. For populations where the clusters are many or large, this can be beneficial, as the maximal sampling effort is reduced, even though the effect is probably rather small. Restricting the sampling process through 2PACS(1) can however have a larger effect, as small clusters will at most only be slightly smaller compared to when using 2PACS(∞), whereas large clusters will not get too costly.

Generally, a HT estimator would be expected to yield lower variance compared to a Hansen-Hurwitz (HH) estimator, using similar designs. This stems from a larger variation of the number of inclusions for the HH estimator. For a non-adaptive sample (i.e., $\beta = 0$), this effect might be quite small, but as the clusters grow larger with increasing β, there is an increased likelihood of including plots multiple times. When the expansion is unrestricted, Thompson (1990) shows that the HT estimator outperforms the HH estimator. However, by using a sampling procedure that spreads the initial sample geographically, such as the LPM, we can reduce the probability of selecting plots from the same clusters multiple times, thus mitigating this effect.

Another effect of 2PACS(1) is that the expected number of inclusions of a plot is likely to have a higher correlation with the variable of interest, compared to 2PACS(∞). A positive correlation between the variable of interest and the expected number of inclusions should yield lower variance, as is the basis for designs using probabilities proportional to size. For a cluster with a uniform distribution of objects, the total of the variable of interest for plots at the edge of a cluster will on average be lower, as plots will encircle fewer objects. At the same time, plots at the edge of a cluster have fewer neighboring plots fulfilling the criteria of inclusion, leading to a lower expected number of inclusions for such plots under 2PACS(1). This has been shown in preliminary simulations, for the populations in this manuscript.

Further research should explore the effects of available auxiliary information, as data on where the clusters might be located may reduce the possible gain of using a variant of ACS.

ORCID

Wilmer Prentius https://orcid.org/0000-0002-3561-290X

REFERENCES

Brown, J. A., & Manly, B. J. F. (1998). Restricted adaptive cluster sampling. *Environmental and Ecological Statistics*, *5*(1), 49–63. https://doi.org/10.1023/A:1009607403647

Christman, M. C. (1997). Efficiency of some sampling designs for spatially clustered populations. *Environmetrics*, *8*(2), 145–166. https://doi.org/10.1002/(SICI)1099-095X(199703)8:2<145::AID-ENV249>3.0.CO;2-T

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. https://doi.org/10.1016/0167-7152(93)90028-H

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

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), e2480. https://doi.org/10.1002/env.2480

Hansen, M. H., & Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, *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. *Journal of the American Statistical Association*, *47*(260), 663–685. https://doi.org/10.2307/2280784

Salehi, M., & Seber, G. A. (1997). Two-stage adaptive cluster sampling. *Biometrics*, *53*(3), 959–970. https://doi.org/10.2307/2533556

Stevens, D. L., Jr., & Olsen, A. R. (2003). Variance estimation for spatially balanced samples of environmental resources. *Environmetrics*, *14*(6), 593–610. https://doi.org/10.1002/env.606

Talvitie, M., Leino, O., & Holopainen, M. (2006). Inventory of sparse forest populations using adaptive cluster sampling. *Silva Fennica*, *40*(1), 101. https://doi.org/10.14214/sf.354

Thompson, S. K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, *85*(412), 1050–1059. https://doi.org/10.2307/2289601

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*(sup1), 234–240. https://doi.org/10.2307/143141

How to cite this article: Prentius, W., & Grafström, A. (2022). Two-phase adaptive cluster sampling with circular field plots. *Environmetrics*, *33*(5), e2729. https://doi.org/10.1002/env.2729

APPENDIX A

A.1 Horvitz–Thompson estimator

Let p_i be the draw probability of plot i being selected into S_{2a}. Let p_{C_i} be the draw probability of any plot in C_i being selected into S_{2a}. Given independent draws, we have

$$\begin{aligned}
\hat{Y}_2 &= \sum_{i \in S_2} \frac{y_i}{\pi_i} \\
p_{C_i} &= \sum_{k \in C_i} p_k \\
\pi_i &= \Pr(i \in S_2) = 1 - \prod_{k=1}^{n} (1 - p_{C_i}) \\
\pi_{ij} &= \Pr(i \in S_2, j \in S_2) = 1 - \Pr(i \notin S_2 \cup j \notin S_2) \\
\Pr(i \notin S_2 \cup j \notin S_2) &= 2 - \pi_i - \pi_j - \Pr(i \notin S_2, j \notin S_2) \\
\Pr(i \notin S_2, j \notin S_2) &= \prod_{k=1}^{n} (1 - p_{C_i \cup C_j}) \\
\pi_{ij} &= \pi_i + \pi_j + \prod_{k=1}^{n} (1 - p_{C_i \cup C_j}) - 1,
\end{aligned}$$

forming the basis for a HT estimator (Horvitz & Thompson, 1952).

A.2 Comparable SRS sample size

We assume the cost function

$$\text{cost} = c_0 + c_{2a} n + c_I n_I + c_c n_c,$$

where c_0 is a fixed cost relating to the survey, c_{2a} is the marginal cost associated with moving to a plot selected in the initial sample S_{2a}, c_I, n_I are the marginal cost and the number of plots which need to be inventoried, and c_c, n_c is the marginal cost and the number of plots that need to be checked regarding to criteria of inclusion.

Using this cost function, we can express the cost for a SRS sample (i.e., a sample without adaptive expansion) as

$$\text{cost}_{\text{SRS}} = c_0 + c_{2a} n_{\text{SRS}} + c_I P n_{\text{SRS}} + c_c n_{\text{SRS}},$$

where P is the proportion of plots fulfilling the criteria of inclusion (a random variable through the first-phase sample). Similarly, we can express the cost for an ACS sample as

$$\text{cost}_{\text{ACS}} = c_0 + c_{2a} n_{\text{ACS}} + c_I n_I + c_c n_c.$$

For a given initial sample size n_{ACS}, we can find the sample size n_{SRS} which gives the same expected cost for both sampling schemes:

$$\begin{aligned}
E(\text{cost}_{\text{SRS}}) &= c_0 + n_{\text{SRS}} (c_{2a} + c_I E(P) + c_c), \\
E(\text{cost}_{\text{ACS}}) &= c_0 + c_{2a} n_{\text{ACS}} + c_I E(n_I) + c_c E(n_c), \\
E(\text{cost}_{\text{SRS}}) &= E(\text{cost}_{\text{ACS}}) \\
\Rightarrow n_{\text{SRS}} &= \frac{c_{2a} n_{\text{ACS}} + c_I E(n_I) + c_c E(n_c)}{c_{2a} + c_I E(P) + c_c},
\end{aligned}$$

getting a rough estimate of n_{SRS} by replacing the expected values with the Monte Carlo estimates.

Acta Universitatis agriculturae Sueciae

Doctoral Thesis No. 2023:80

Environmental monitoring plays a crucial role in guiding climate change and conservation policy decisions. To obtain reliable insights from environmental populations, it is essential to adopt probability sampling. Use of auxiliary variables can greatly enhance the quality by reducing estimator variability. This thesis demonstrates different ways auxiliary information can be used in sampling designs, and introduces new designs to be used when studying environmental populations.

Wilmer Prentius received his doctoral education at the Department of Forest Resource Management, SLU, Umeå. He holds a MSc degree in Statistics from Umeå University, Umeå.

ISSN 1652-6880
ISBN (print version) 978-91-8046-212-9
ISBN (electronic version) 978-91-8046-213-6