



A sample coordination method to monitor totals of environmental variables

Xin Zhao | Anton Grafström

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

Correspondence

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

Abstract

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

KEYWORDS

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

1 | INTRODUCTION

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

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

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *Environmetrics* published by John Wiley & Sons Ltd.

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

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

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

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

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

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

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

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

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

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

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

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

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

2 | SAMPLING FRAMEWORK

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

2.1 | Double sampling approach

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

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

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

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

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

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

$$\text{var}(\hat{Y}(t)) = \sum_{i=1}^N \sum_{j=1}^N (\pi_{ij}(t) - \pi_i(t)\pi_j(t)) \frac{y_i(t)}{\pi_i(t)} \frac{y_j(t)}{\pi_j(t)}, \quad (2)$$

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

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

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

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

$$\text{var}(\hat{\Delta}_Y(1, 2)) = \text{var}(\hat{Y}(2)) + \text{var}(\hat{Y}(1)) - 2 \text{cov}(\hat{Y}(1), \hat{Y}(2)). \quad (4)$$

The covariance term in (4) is given by

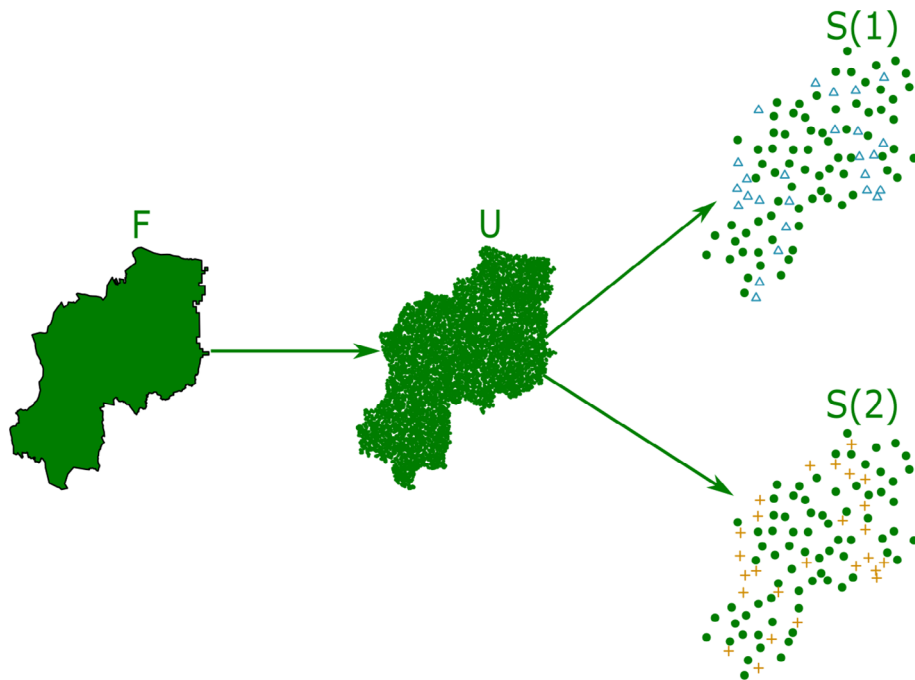
$$\text{cov}(\hat{Y}(1), \hat{Y}(2)) = \sum_{i=1}^N \sum_{j=1}^N (\pi_{ij}(1, 2) - \pi_i(1)\pi_j(2)) \frac{y_i(1)}{\pi_i(1)} \frac{y_j(2)}{\pi_j(2)}, \quad (5)$$

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

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

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

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



2.2 | Sample coordination for a finite population

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

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

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

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

2.3 | Spatial balance

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

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

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

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

3 | SAMPLE COORDINATION FOR SPATIALLY BALANCED SAMPLES

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

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

3.1 | Spatially correlated Poisson sampling

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

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

3.2 | Algorithm of SCPS

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

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

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

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

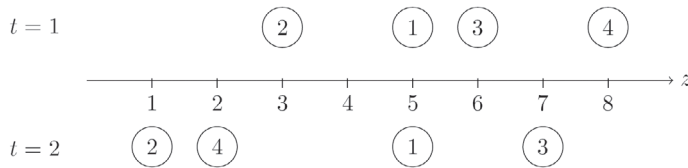


FIGURE 2 The distances between units of the auxiliary variable for two time occasions in Example 1

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

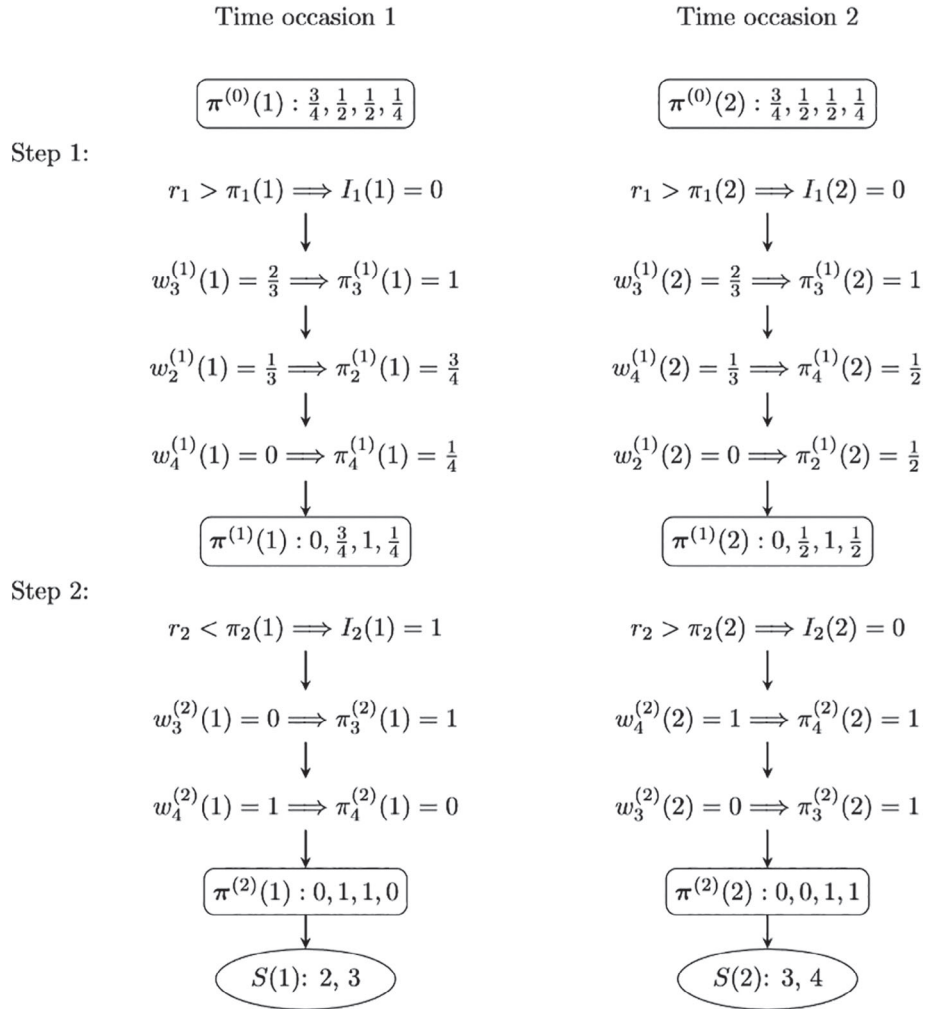
When it comes to estimation under positively coordinated samples selected with SCPS, we may use the unbiased HT-estimator (1). However, for designs that produce well-spread samples, many of the second-order inclusion probabilities may be zero. Hence, it is not possible to have a design-based unbiased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

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

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

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

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



4 | A FOREST INVENTORY APPLICATION

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

TABLE 2 Example National Forest Inventory

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

Note: The first-phase sample size is $N = 10,000$, the second-phase sample size is $n(1) = n(2) = 100$, the repetition time is 10,000. For tree height, the sample mean for the two occasions is 100.27 and 101.50 dm, respectively. The percentage of change for the auxiliary is 95.17% and correlation of the two time occasions for the auxiliary is .9186. For basal area, the sample mean is 13.76 and 14.64 m²/ha, respectively, for the two time occasions. The percentage of change is 95.17% and correlation coefficient of the two time occasions for the auxiliary variable is .9205. $B(i)$ is the mean of spatial balance for the sample selected at time occasion i . Overlap is the mean of the percentage of overlap for the sample units selected at two time occasions. $V\left(\widehat{Z}_i(t)\right)$ represents the empirical variance of the estimator for the mean of the auxiliary variable. $V\left(\widehat{\Delta}_{Z_i(1,2)}\right)$ stands for the variance of the estimator of change. P_{ij} is the split panel design, the value of i corresponding to the percentage of the permanent sample and value of j corresponding to the percentage of temporary sample. For example, P_{28} means the permanent panel is 20% and the temporary panel is 80%.

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

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

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

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

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

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

5 | CONCLUSION AND DISCUSSION

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

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

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

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

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

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

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

ACKNOWLEDGEMENTS

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

REFERENCES

- Artiola, J. F., Pepper, I. L., & Brusseau, M. L. (2004). *Environmental monitoring and characterization*. London, UK: Elsevier.
- Awange, J. L. (2012). *Environmental Monitoring Using GNSS: Global navigation satellite systems*. Berlin/Heidelberg, Germany: Springer.
- Benedetti, R., Piersimoni, F., & Postiglione, P. (2015). *Sampling spatial units for agricultural surveys*. Berlin/Heidelberg, Germany: Springer.
- Bondesson, L., & Thorburn, D. (2008). A list sequential sampling method suitable for real-time sampling. *Scandinavian Journal of Statistics*, 35, 466–483. <https://doi.org/10.1111/j.1467-9469.2008.00596.x>
- Brewer, K. R. W., Early, L. J., & Joyce, S. F. (1972). Selecting several samples from a single population. *Australian Journal of Statistics*, 14(3), 231–239. <https://doi.org/10.1111/j.1467-842X.1972.tb00899.x>
- Cotton, F. & Hesse, C. (1992). Tirages coordonnés d'échantillons. Technical report E9206, Direction des Statistiques Économiques. Paris, France: INSEE.
- Dehardt, J. (1971). Generalizations of the Glivenko-Cantelli theorem. *The Annals of Mathematical Statistics*, 46(2), 2050–2055. <https://doi.org/10.1214/aoms/1177693073>
- Dobbie, M. J., Henderson, B. L., & Stevens, D. L. (2008). Sparse sampling: Spatial design for monitoring stream networks. *Statistics Surveys*, 2(2008), 113–153. <https://doi.org/10.1214/07-SS032>
- Duncan, G. J., & Kalton, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, 55(1), 97–117. <https://doi.org/10.2307/1403273>
- Ernst, L. R. (1999). The maximization and minimization of sample overlap: A half century of results. *Bulletin of the International Statistical Institute*, 57, 293–296.
- Fridman, J., Holm, S., Nilsson, M., Ringvall, A. H., & Ståhl, G. (2014). Adapting national forest inventories to changing requirements – the case of the Swedish national forest inventory at the turn of the 20th century. *Silva Fennica*, 48(3), 1–29. <https://doi.org/10.14214/sf.1095>
- Grafström, A. (2012). Spatially correlated Poisson sampling. *Journal of Statistical Planning and Inference*, 142(1), 139–147. <https://doi.org/10.1016/j.jspi.2011.07.003>
- Grafström, A. & Liscic, J. (2019). Balanced sampling: Balanced sampling and spatially balanced sampling. *R package version 1.5.5*.
- Grafström, A., Lundström, N. L. P., & Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, 68(2), 514–520. <https://doi.org/10.1111/j.1541-0420.2011.01699.x>
- Grafström, A., & Matei, A. (2015). Coordination of conditional poisson samples. *Journal of Official Statistics*, 31(4), 649–672. <https://doi.org/10.1515/jos-2015-0039>
- Grafström, A., & Matei, A. (2018a). Spatially balanced sampling of continuous populations. *Scandinavian Journal of Statistics*, 45(3), 792–805. <https://doi.org/10.1111/sjos.12322>
- Grafström, A. & Matei, A. (2018b). Coordination of spatially balanced samples. *Survey Methodology*, 44(2), 215–238 (Statistics Canada, Catalogue No. 12-001-X).
- Grafström, A., & Schelin, L. (2014). How to select representative samples. *Scandinavian Journal of Statistics*, 41(2), 277–290. <https://doi.org/10.1111/sjos.12016>
- Grafström, A., Schnell, S., Saarela, S., Hubbell, S. P., & Condit, R. (2017). The continuous population approach to forest inventories and use of information in the design. *Environmetrics*, 28(8), 1–12. <https://doi.org/10.1002/env.2480>
- Grafström, A., Zhao, X., Nylander, M., & Petersson, H. (2017). A new sampling strategy for forest inventories applied to the temporary clusters of the Swedish national forest inventory. *Canadian Journal of Forest Research*, 47(9), 1161–1167. <https://doi.org/10.1139/cjfr-2017-0095>
- Keyfitz, N. (1951). Sampling with probabilities proportional to size: Adjustment for changes in the probabilities. *Journal of the American Statistical Association*, 46(253), 105–109. <https://doi.org/10.1080/01621459.1951.10500773>
- Kish, L. (1983). *Data collection for details over space and time*. In T. Wright (Ed.), *Statistical methods and the improvement of data quality* (pp. 73–84). Cambridge, MA: Academic Press.

- Kish, L. (1986). Timing of surveys for public policy. *Australian Journal of Statistics*, 28(1), 1–12. <https://doi.org/10.1111/j.1467-842X.1986.tb00579.x>
- Kish, L., & Scott, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66(335), 461–470. <https://doi.org/10.2307/2283509>
- Mach, L., Reiss, P. T., & Şchiopu-Kratina, I. (2006). Optimizing the expected overlap of survey samples via the northwest corner rule. *Journal of the American Statistical Association*, 101(476), 1671–1679. <https://doi.org/10.1198/016214506000000320>
- Mandallaz, D. (2007). *Sampling techniques for forest inventories*. Boca Raton, FL: CRC Press.
- Marker, D. A., & Stevens, D. L., Jr. (2009). *Sampling and inference in environmental surveys*. In D. Pfeffermann & C. R. Rao (Eds.), *Handbook of statistics 29A: Sample surveys: Design, methods and applications* (pp. 487–512). Cambridge, MA: Elsevier.
- McDonald, T. L. (2003). Review of environmental monitoring methods: Survey designs. *Environmental Monitoring and Assessment*, 85(3), 277–292. <https://doi.org/10.1023/A:1023954311636>
- Patterson, H. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Series B (Methodological)*, 12(2), 241–255. <https://doi.org/10.1111/j.2517-6161.1950.tb00058.x>
- Qualité, L., & Tillé, Y. (2008). Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added. *Survey Methodology*, 34(2), 173–181.
- Robertson, B., Brown, J. A., McDonald, T., & Jaksons, P. (2013). BAS: Balanced acceptance sampling of natural resources. *Biometrics*, 69(3), 776–784. <https://doi.org/10.1111/biom.12059>
- Robertson, B., McDonald, T., Price, C., & Brown, J. (2018). Halton iterative partitioning: Spatially balanced sampling via partitioning. *Environmental and Ecological Statistics*, 25(3), 305–323. <https://doi.org/10.1007/s10651-018-0406-6>
- Scott, C. T. (1984). A new look at sampling with partial replacement. *Forest Science*, 30(1), 157–166.
- Scott, C. T. (1998). Sampling method for estimating change in forest resources. *Ecological Applications*, 8(2), 228–233. <https://doi.org/10.2307/2641062>
- Sen, A. R. (1973). Theory and application of sampling on repeated occasions with several auxiliary variables. *Biometrics*, 29(2), 381–385. <https://doi.org/10.2307/2529401>
- Stevens, D. L., & Olsen, A. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, 99(465), 262–278. <https://doi.org/10.1198/016214504000000250>
- Tomppo, E., Gschwantner, T., Lawrence, M., & McRoberts, R. E. (2010). *National forest inventories pathways for common reporting*. Dordrecht, Netherlands: Springer.
- Wolfowitz, J. (1954). Generalization of the theorem of Glivenko-Cantelli. *The Annals of Mathematical Statistics*, 25(1), 131–138.

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