Received: 17 July 2021 | Revised: 8 August 2023 | Accepted: 3 September 2023

DOI: 10.1002/env.2825

RESEARCH ARTICLE

WILEY

Estimation of change with partially overlapping and spatially balanced samples

Xin Zhao[1,2] | **Anton Grafström**[2]

[1]Division for Statistics and Data Collection, Swedish Forestry Agency, Umeå, Sweden

[2]Department of Forest Resource Management, Swedish University of Agriculture Sciences, Umeå, Sweden

Correspondence
Xin Zhao, Division for Statistics and Data Collection, Swedish Forestry Agency, Box 284, 90106 Umeå, Sweden.
Email: xin.zhao@skogsstyrelsen.se

Abstract

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

KEYWORDS

overlapping samples, repeated surveys, spatially correlated Poisson sampling, well-spread samples

1 | INTRODUCTION

In repeated surveys, the monitoring of changes in population totals over time is a common focus in various fields, including environmental and ecological research see, for example, Kalton (1983), Wang and Zhu (2019). Accurately estimating the variance of an estimator of change is crucial for assessing the statistical significance of observed changes (Berger & Priam, 2016). In the field of environmental metrics, understanding and monitoring changes in natural systems are of paramount importance. Environmental variables, such as water quality, air pollution levels, or habitat suitability, play critical roles in ecosystem health and human well-being. By monitoring changes in these variables over time, researchers and policymakers can gain insights into the impacts of human activities, climate change, and other ecological factors on the environment Foss et al. (2022), Lowther et al. (2023). Additionally, tracking changes in environmental variables helps identify areas of concern, guide conservation efforts, and support sustainable resource management practices.

Considering the broader environmental and ecological context, it becomes evident that spatial sampling and the spatial distribution of populations play crucial roles in accurately estimating changes in environmental variables. The spatial arrangement of sample units can provide valuable insights into the underlying ecological processes and help identify spatial patterns or hotspots of change. For example, nearby units in the spatial domain often exhibit similar values due

to shared environmental conditions or underlying ecological interactions. Incorporating the geographical locations of populations and utilizing spatially balanced sampling techniques have been recognized as essential components of effective survey designs in environmental and ecological research Stevens and Olsen (2003).

Spatially balanced sampling designs, such as systematic spatial grids or randomized stratified sampling based on auxiliary variables, ensure that sample units are distributed evenly across the study area, accounting for the spatial dependence of environmental variables. These designs offer advantages over traditional random sampling methods by capturing a representative range of environmental conditions and reducing the potential bias introduced by clustering or spatial autocorrelation. Furthermore, recent studies by Zhao and Grafström (2020) have demonstrated the benefits of employing spatially balanced and partially overlapping samples for monitoring changes in environmental variables, leading to improved efficiency and reduced variance in state and change estimators.

Despite the advantages of using spatially balanced and partially overlapping samples for monitoring changes, there remains a critical knowledge gap in understanding the estimation of variance for the estimator of change under these sampling designs. This study aims to address this important issue by developing novel estimation techniques that account for the unique characteristics of spatially balanced sampling designs and provide reliable variance estimates for change estimators.

It is well known that, when estimating the variance of an estimator of change, we need to estimate the variance of the two state estimators as well as the covariance between them, for example, Kish (1965, ch. 12). To reduce the variance of the estimator of change, we can either make the variance of the two state estimators smaller or attempt to create a high positive covariance between the two state estimators or try both of them. The question of whether we should use independent samples, a permanent sample or partially overlapping samples over time arises (De Leeuw et al., 2008, ch. 25).

For independent samples (where a new sample is taken independently of previous samples at each survey time), we do not need to consider the covariance. Then, the variance of change depends only on the variance of the estimators at each time occasion. This simplifies the estimation problem. However, it will not be the best strategy to use independent samples when estimating changes over time. This is because the variance of the change estimator becomes the sum of the variances of the state estimators when using independent samples. When the time between surveys is short and the values of the target variables have not changed much, a permanent sample (where the same sample units would be revisited at each survey time) might be employed to reduce the variance of an estimator of change. However, at the following time occasion, the permanent sample may not be as representative as it used to be as the population changes over time. If the sample changes in a different way than the population, which is out of our control, then there is a risk of a much larger variance of the state estimator at the following time occasion (Scott, 1998). Thus, even if the covariance between the two state estimators becomes large by having fully overlapping samples, it is not guaranteed that the variance of change will be reduced. There is a need for updating the sample at the next time occasion to account for changes while retaining as many units as possible from the old sample see, for example, Patterson (1950), Breidt and Fuller (1999).

A large number of variance estimators (approximations) have been proposed under different sampling designs (e.g., Berger, 2004; Hájek, 1964; Hartley & Rao, 1962; Horvitz & Thompson, 1952; Yates & Grundy, 1953). For repeated surveys, researchers have also paid a lot of attention to the estimation of covariance. Tam (1984) was one of the earliest studies that considered covariance estimations from overlapping samples. Qualité (2009, ch. 5) derived covariance estimators based on two overlapping samples by considering sampling designs that are essentially applicable to obtain rotating panels, that is, panels where only a part of the sample at a previous time occasion is maintained, and the rest of the units in the sample are replaced by new units at a next time occasion.

Due to the improved spread of samples achieved with spatially balanced sampling designs, it may not be appropriate to use conservative variance or covariance estimators typically employed in simple random sampling. Grafström and Schelin (2014) introduced a novel local mean variance estimator tailored for spatially balanced sampling designs. Instead of using a global mean, this method adopts a local mean in the variance estimator. It considers only the nearest neighbors of a sample unit including the unit itself in the computation of the local mean. Additionally, the authors addressed situations where the distances between units are equal, resulting in the local neighborhood size of sample units being subject to variation and not remaining constant in the variance estimator. In our work, we enhance this variance estimator by introducing a fixed size for the local neighborhoods of all sample units. Furthermore, following the settings in Qualité (2009, ch. 5), we develop a local mean covariance estimator and derive a variance estimator for the estimator of change. Through simulations, we demonstrate that the proposed local mean estimators exhibit stability and lower bias compared to estimators that do not utilize local means. Consequently, these local mean variance and covariance estimators can be effectively employed when estimating the variance of the estimator of change with partially overlapping and spatially balanced samples.

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

2 | ESTIMATION OF CHANGE WITH GENERAL DESIGNS

Suppose we have a shared list frame $U = \{1, \dots, i, \dots, N\}$ for N objects (e.g., field plots) over time. A list frame is a finite list of labels used to sample and identify units. From U, a sample S_t of n_t labels of corresponding plots can be selected at time t. Denote the target variable for unit i at time t as y_{it}, it can be total number of trees in plot i at time t. The total can be expressed as $Y_t = \sum_{i \in U} y_{it}$. Let $\pi_{it} = \Pr(i \in S_t)$ be the prescribed inclusion probability of unit i at time t, that is, the probability that unit i is selected to the sample S_t. The Horvitz–Thompson (HT) estimator of Y_t can be expressed as

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

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

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

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

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

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

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

Estimator (4) is unbiased for (3) if all second-order inclusion probabilities π_{ijt} are strictly positive. Otherwise, it is impossible to obtain unbiased variance estimators.

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

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

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

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

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

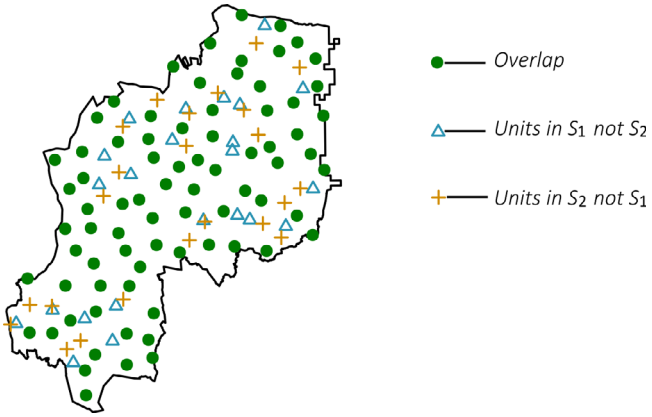

FIGURE 1 Illustration of two samples selected by the strategy.

3 | AN EFFICIENT SAMPLING STRATEGY TO MONITOR THE CHANGE OF ENVIRONMENTAL VARIABLES

In environmental surveys, the spatial pattern of units is important because the units themselves are defined using spatial criteria. Since nearby units are more similar than units that are farther apart, more information could then be obtained if the random sample avoids the selection of nearby units. To achieve good estimates of population characteristics, the spatial pattern of the sample should be similar to the spatial pattern of the population. Often, we do not know the spatial pattern of the target variable before the sample is selected. Instead, we have full access to some auxiliary variables that are related to the target variables. Stevens and Olsen (2004) introduced the generalized random tessellation stratified (GRTS) design and coined the phrase "spatially balanced sampling". They also proposed a statistic that measures the spatial balance of a sample using Voronoi polygons. The local pivotal method (LPM) and spatially correlated Poisson sampling (SCPS) proposed by Grafström et al. (2012) and Grafström (2012) are two spatially balanced sampling designs that employ auxiliary variables (often including geographical coordinates plus several other attribute variables) to spread the samples based on distances. Grafström and Lundström (2013) illustrated that when the target variables are smooth functions of auxiliary variables, we can get improved estimators if the samples are spread in the auxiliary variables. It has been confirmed that, by using these designs, we gain in efficiency of design–based estimators of the totals of target variables (see e.g., Benedetti et al., 2017).

Regarding the monitoring of change, we need to be cautious about the determination of whether a sampling strategy is an efficient strategy or not. Zhao and Grafström (2020) proposed an efficient sampling strategy for monitoring the change of environmental variables. In this strategy, the concept of spatially balanced samples and positive sample coordination are combined. The spatially balanced samples are selected by the SCPS. When applying the SCPS, a set of auxiliary variables that are related to the target variables should be used to spread the sample. We choose the same set of auxiliary variables (with different values) at different time occasions. The positive sample coordination is achieved by assigning the same random number to each unit in the algorithm of SCPS. In this way, we will get partially overlapping and spatially balanced samples. Figure 1 illustrates two such samples selected by SCPS.

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

4 | ESTIMATION OF CHANGE WHEN SAMPLES ARE OVERLAPPING AND WELL SPREAD

Under a spatially balanced sampling design, it is often difficult to obtain π_{ijt} and π_{ij}^{12}. Moreover, many second-order inclusion probabilities may be zero. It will likely not be possible to use design-based unbiased variance estimators such as

(4) and (6) under spatially balanced sampling designs. Even if it is possible, it will generally not be recommended as such variance estimators can become highly unstable when some second-order inclusion probabilities are very small.

4.1 | Variance estimators for spatially balanced samples

Matérn (1947) introduced a variance estimator for systematic sampling from a regular grid of sample locations. In Matérn's variance estimator, the sample locations are split into several nonoverlapping groups of neighbors. A local variance is first constructed for each group, then an average over groups is calculated as the variance estimator. Motivated by this estimator, Grafström and Schelin (2014) also proposed a local mean variance estimator which was shown to perform well under spatially balanced sampling. In their variance estimator, the local neighborhood for each sample unit i depends only on i and its nearest neighbors. The size of the local neighborhood is calculated by one (the sample unit i) plus the number of nearest neighbors of i in the sample. For example, if a sample unit i has two nearest neighbors, then the size of local neighborhood is three. Their variance estimator can be applied in situations where units have many nearest neighbors. For well-spread environmental samples, it is rare for a sample unit to have equidistant neighbors. In principle, by only including the unit i and its nearest neighbor in the local neighborhood, we often have two units in the local neighborhood when estimating the variance with the estimator proposed by Grafström and Schelin (2014). In Stevens and Olsen (2003), the authors recommended using four sample units in the local neighborhood. This is because they found that their local mean variance estimator became unstable when including fewer sample units in the local neighborhood. We consider their suggestion and modify the local mean variance estimator in Grafström and Schelin (2014) by using a neighborhood size proportional to the sample size. For $V\left(\hat{Y}_t\right)$, the local mean variance estimator can be expressed as

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

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

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

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

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

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

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

4.2 | Covariance estimator for partially overlapping and spatially balanced samples

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

of the covariance. As a starting point, we introduce the setting with independent observations. First n_1 independent observations are drawn from U to S_1 according to the drawing probabilities p_i, and a subsample S_{12} of S_1 is retained as a part of S_2, with $n_{12} \geq 2$ observations. Next, an additional number of $n_2 - n_{12}$ independent observations are drawn from U to S_2 according to p_i. Now, the two samples S_1 and S_2 share n_{12} observations in the sample S_{12}. In this setting, we estimate the total $Y_t = \sum_{i \in U} y_{it}$ with $\hat{Y}_t = \sum_{i \in S_t} y_{it} n_t^{-1} p_i^{-1}$ for $t = 1, 2$. As this is a special case of the bidimensional sampling design described in Qualité (2009, ch. 5), the covariance between $\hat{Y}_1$ and $\hat{Y}_2$ follows from equation (5.5) in Qualité (2009, ch. 5). For sampling with replacement, it can be expressed as

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

and the covariance (10) can be estimated using S_{12} by the simple expansion

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

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

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

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

where $\bar{y}_{i1} = n_{l12}^{-1} \sum_{j \in S_{i_1}} y_{j1} \pi_{j1}^{-1}, \bar{y}_{i2} = n_{l12}^{-1} \sum_{j \in S_{i_2}} y_{j2} \pi_{j2}^{-1}$ and S_{it} is the local neighborhood for unit i in S_{12} at time t. The neighborhood size n_{l12} is chosen as $\lfloor p_l n_{12} \rceil$ ($p_l n_{12}$ rounded to the nearest integer) and the same proportion p_l as in the local mean variance estimator is recommended. Since $n_{l12} \leq n_t$, it is reasonable to decide the proportion by the size of the overlap when estimating the variance of the estimator of change, that is, $p_l = n_{l12} n_{12}^{-1}$, where n_{l12} can be any integer between two and n_{12}. Then we make sure that $n_{lt} = \lfloor p_l n_t \rceil \geq n_{l12}$. If the size $n_{l12} = n_{12}$, we get back to the estimator (11). The estimator (12) of covariance under spatially balanced sampling is consistent with the estimator $\hat{V}_{SB}\left(\hat{Y}_t\right)$ of variance, that is, $\hat{C}_{SB}\left(\hat{Y}_t, \hat{Y}_t\right) = \hat{V}_{SB}\left(\hat{Y}_t\right)$. This is important for estimating the variance of an estimator of change. Combining (7) and (12), the expression of the variance estimator for the estimator of change with partially overlapping and spatially balanced samples follows.

Different from the sampling plans in Qualité (2009, ch. 5), the size of overlap is random when using the strategy in Zhao and Grafström (2020). For the current algorithm in the strategy, it is not possible to fix the size of the overlap and select a well-spread sample at a second time occasion. That is to say, we do not know how many sample units from S_1 that will also be selected into S_2 before we get the full sample on the second time occasion. The percentage of overlap between two samples depends mainly on the change over time of the auxiliary variables that we use to spread the samples. When the changes in auxiliary variables increase, the overlap between two samples tends to decrease. Similar to the variance estimator (7), the covariance estimator (12) is proposed as a general estimator for spatially balanced samples. In the next section, we study the performance of the proposed variance and covariance estimators, specifically under the strategy in Zhao and Grafström (2020).

5 | EVALUATION OF THE ESTIMATORS

To evaluate the proposed estimators for positively coordinated and spatially balanced samples, two simulation studies are considered. In the first study we evaluate the estimators under different spatial configurations by applying surfaces. For the second study, the estimators are evaluated for a forest inventory as an example of environmental monitoring. In both studies, we take different sizes of neighborhoods into account to check how they will affect the estimators. Estimators

which apply the full samples/overlap in the neighborhoods are incorporated in the simulations as well. It is worth noting that, the samples selected at the two time occasions are well spread and positively coordinated in both examples. For each study, the empirical variance and covariance, the mean of the variance and covariance estimators are presented. We calculate the mean coverage rates for the 95% confidence intervals when using the variance estimators. The relative bias(RB) as well as the empirical relative root mean squared error (RRMSE) for the estimators are also compared for different estimators. Before we go any further in the studies, we will give the explicit expressions of RB and RRMSE.

RB is the ratio between the bias of an estimator and the value we are estimating. Suppose $\widehat{\theta}$ is an estimator of θ, the relative bias of $\widehat{\theta}$ can be denoted as

$$\mathrm{RB}(\widehat{\theta}) = \frac{\mathrm{Bias}(\widehat{\theta})}{\theta} \cdot 100\%. \tag{13}$$

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

$$\mathrm{MSE}(\widehat{\theta}) = E\Big((\widehat{\theta} - \theta)^2\Big) = V(\widehat{\theta}) + \Big(\mathrm{Bias}(\widehat{\theta})\Big)^2. \tag{14}$$

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

$$\mathrm{RMSE}(\widehat{\theta}) = \sqrt{MSE(\widehat{\theta})}. \tag{15}$$

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

$$\mathrm{RRMSE}(\widehat{\theta}) = \frac{\mathrm{RMSE}(\widehat{\theta})}{\theta} \cdot 100\%. \tag{16}$$

Simulation study 1. The purpose of this study is to examine the behavior of estimators when tracking changes in elevation for positively coordinated and spatially balanced samples. To define the target variable, a surface consisting of 2500 pixels is used on each time occasion. The initial target surface at time 1 is generated by applying a Gaussian kernel function to 20 observations (points) within the range of $[0, 1] \times [0, 1]$. The smoothing parameter, σ, is set to 0.1. The value of each point is randomly generated from a uniform distribution, $U(50, 50)$. At the subsequent time occasion, the target surface is obtained by adding an error surface to the surface from time 1 to ensure a change in the elevation. The error surface is generated similarly to the initial target surface, with 30 points randomly generated from a $U(10, 20)$ distribution. We employ geographical coordinates as the two auxiliary variables to spread the samples to ensure the selection of positively coordinated and spatially balanced samples. At each iteration, we select a sample of size $n_1 = 100$ at the first time occasion, and a smaller sample of size $n_2 = 50$ at time 2 with equal inclusion probabilities of $\pi_{it} = n_t/N$. By varying sample sizes, we aim to prevent the occurrence of permanent samples since geographical coordinates remain constant. The simulations are repeated 10,000 times. The surfaces are presented in Figure 2, and Table 1 displays the simulation results for estimators that utilize varying neighborhood sizes.

Simulation study 2. For this study, a specific area in central Sweden has been chosen as the focus. The population of interest consists of $N = 10{,}000$ clusters of circular plots. The primary objective of this study is to assess the performance of estimators under the conditions of positively coordinated and spatially balanced samples when monitoring changes in the basal area of the population over time. To ensure the selection of such samples, we employed four different auxiliary variables, namely geographical coordinates, elevation, and tree height. Among these, tree height is the only variable that changes over time, whereas geographical coordinates and elevation remain constant at time occasion 2. Our target variable is the basal area of the population changes over time as well. At each iteration, we have sample size $n_1 = n_2 = 100$, equal inclusion probabilities $\pi_t = n_t/N = 0.01$ and the number of repetitions is 10,000. The results are illustrated for the basal area in Table 2.

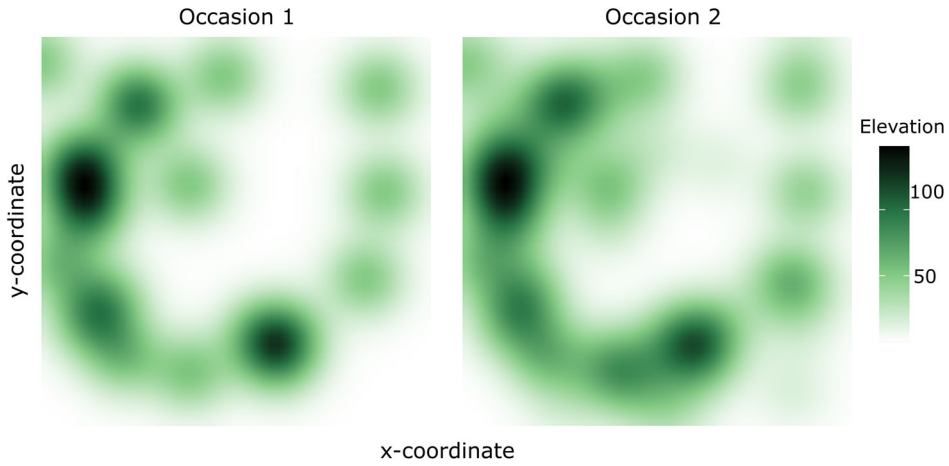

FIGURE 2 Target surfaces in Example 1. Darker colors indicate higher values of elevation.

TABLE 1 Performance of estimators in simulation study 1.

	$n_{l12}=2$	$n_{l12}=4$	$n_{l12}=6$	$n_{l12}=n_{12}$
$V\left(\hat{Y}_1\right)$	5,603,276			
$V\left(\hat{Y}_2\right)$	24,663,288			
$C\left(\hat{Y}_1,\hat{Y}_2\right)$	2,348,328			
$V(\hat{\Delta})$	25,563,549			
$\bar{\hat{V}}_{SB}\left(\hat{Y}_1\right)$	18,307,756 (0.999)	29,398,474 (1)	34,905,003 (1)	51,394,294 (1)
$\bar{\hat{V}}_{SB}\left(\hat{Y}_2\right)$	59,566,510 (0.995)	80,948,404 (0.999)	95,541,122 (0.999)	150,657,134 (1)
$\bar{\hat{C}}_{SB}\left(\hat{Y}_1,\hat{Y}_2\right)$	19,688,631	22,676,798	25,939,517	38,593,385
$\bar{\hat{V}}_{SB}(\hat{\Delta})$	38,497,004 (0.960)	64,993,281 (0.996)	78,567,092 (0.999)	124,864,658 (1)
$RB_{\bar{\hat{V}}_{SB}\left(\hat{Y}_1\right)}$	2.267	4.247	5.229	8.172
$RB_{\bar{\hat{V}}_{SB}\left(\hat{Y}_2\right)}$	1.415	2.282	2.874	5.109
$RB_{\bar{\hat{C}}_{SB}\left(\hat{Y}_1,\hat{Y}_2\right)}$	7.384	8.657	10.046	15.434
$RB_{\bar{\hat{V}}_{SB}(\hat{\Delta})}$	0.506	1.542	2.073	3.885
$RRMSE_{\bar{\hat{V}}_{SB}\left(\hat{Y}_1\right)}$	2.308	4.290	5.270	8.213
$RRMSE_{\bar{\hat{V}}_{SB}\left(\hat{Y}_2\right)}$	1.501	2.350	2.935	5.171
$RRMSE_{\bar{\hat{C}}_{SB}\left(\hat{Y}_1,\hat{Y}_2\right)}$	7.964	9.140	10.535	16.042
$RRMSE_{\bar{\hat{V}}_{SB}(\hat{\Delta})}$	0.782	1.652	2.162	3.963

Note: The column headings and their respective descriptions are as follows: n_{l12}: size of local neighborhood of the local mean covariance estimator, V: empirical variance, C: empirical covariance, $\hat{Y}_t$: HT estimator of population total at time t, $\hat{\Delta}$: estimator of change in population total between two time occasions, $\bar{\hat{V}}_{SB}$: mean value of the local mean variance estimator for 10,000 iterations, $\bar{\hat{C}}_{SB}$: mean values of the local mean covariance estimator for 10,000 iterations, *RB*: relative bias, *RRMSE*: relative root mean squared error. The correlation coefficient between the target variables at the two time occasions is 0.9550. The total of the target is 72,861.65 at time 1 and 104,398.8 at time 2. The mean of percentage of overlap $2E(n_{12})/(n_1+n_2)$ between samples at the two time occasions is 43.35%. The numbers in parenthesis are the mean coverage rates for the 95% confidence intervals when using the variance estimators.

TABLE 2 Performance of estimators in simulation study 2.

	$n_{l12}=2$	$n_{l12}=4$	$n_{l12}=6$	$n_{l12}=n_{12}$
$V\left(\hat{Y}_1\right)$	4,450,984			
$V\left(\hat{Y}_2\right)$	4,117,761			
$C\left(\hat{Y}_1,\hat{Y}_2\right)$	1,973,211			
$V(\hat{\Delta})$	4,622,964			
$\bar{\hat{V}}_{SB}\left(\hat{Y}_1\right)$	5,371,933 (0.965)	6,443,658 (0.980)	8,161,177 (0.991)	34,256,433 (1)
$\bar{\hat{V}}_{SB}\left(\hat{Y}_2\right)$	5,131,544 (0.970)	6,209,523 (0.983)	7,935,950 (0.992)	34,776,755 (1)
$\bar{\hat{C}}_{SB}\left(\hat{Y}_1,\hat{Y}_2\right)$	2,152,611	3,225,581	4,338,479	23,229,975
$\bar{\hat{V}}_{SB}(\hat{\Delta})$	6,198,256 (0.975)	6,202,019 (0.974)	7,420,169 (0.986)	22,573,238 (1)
$RB_{\bar{\hat{V}}_{SB}\left(\hat{Y}_1\right)}$	0.207	0.448	0.834	6.696
$RB_{\bar{\hat{V}}_{SB}\left(\hat{Y}_2\right)}$	0.246	0.508	0.927	7.446
$RB_{\bar{\hat{C}}_{SB}\left(\hat{Y}_1,\hat{Y}_2\right)}$	0.091	0.635	1.199	10.773
$RB_{\bar{\hat{V}}_{SB}(\hat{\Delta})}$	0.341	0.342	0.605	3.883
$RRMSE_{\bar{\hat{V}}_{SB}\left(\hat{Y}_1\right)}$	0.278	0.486	0.859	6.716
$RRMSE_{\bar{\hat{V}}_{SB}\left(\hat{Y}_2\right)}$	0.312	0.544	0.952	7.465
$RRMSE_{\bar{\hat{C}}_{SB}\left(\hat{Y}_1,\hat{Y}_2\right)}$	0.334	0.718	1.255	10.858
$RRMSE_{\bar{\hat{V}}_{SB}(\hat{\Delta})}$	0.437	0.432	0.676	3.985

Note: The column headings and their respective descriptions are as follows: n_{l12}: size of local neighborhood of the local mean covariance estimator, V: empirical variance, C: empirical covariance, $\hat{Y}_t$: HT estimator of population total at time t, $\hat{\Delta}$: estimator of change in population total between two time occasions, $\bar{\hat{V}}_{SB}$: mean value of the local mean variance estimator for 10,000 iterations, $\bar{\hat{C}}_{SB}$: mean values of the local mean covariance estimator for 10,000 iterations, *RB*: relative bias, *RRMSE*: relative root mean squared error. The total basal area is 137,487.4 and 14,6575.3 m^2/ha, respectively for the two time occasions. Correlation coefficient between basal area at time 1 and 2 is 0.9225. The mean of the overlap is 64.05%. The numbers in parenthesis are the mean coverage rates for the 95% confidence intervals when using the variance estimators.

Simulation results of both studies are also illustrated in Figure 3 for all estimators. From the figure and the tables, we can see that all estimators are generally conservative. Compared to the estimators which apply the full samples/overlap in the neighborhood, we can reduce the bias by using local neighborhood estimators. The smaller the neighborhood size, the less biased the estimator tends to be. The coverage rate of the confidence intervals also increases as the neighborhood size grows. Note that, in each iteration we can fix the neighborhood size of the local mean covariance estimator. The size of the neighborhood of the local mean variance estimator varies according to the size of the overlap. If the aim is to estimate the variance of the estimator of the total at each time occasion, we can use a fixed neighborhood size.

6 | DISCUSSION

Recently, many other studies have been conducted to evaluate different variance estimators in order to find less biased variance estimator of HT estimator using systematic samples, see for example, Babcock et al. (2018), Frank and Monleon (2021). For well-spread samples, we have evaluated our proposed estimators by using and equal inclusion probabilities (representative samples). The use of equal inclusion probabilities is, however, the most common case in multipurpose environmental surveys. As the strength of the relation between different target variables and the auxiliary variables that we use to spread the samples are not the same for different target variables, it is safer to spread the samples with equal inclusion probabilities.

When constructing the local mean covariance estimator, we also tested to use the same number of units in the local neighborhood as the local mean variance estimator. By doing so the overestimation by the local mean covariance estimator will become bigger than the overestimation by the local variance estimators. In such a case, it may produce a negative bias

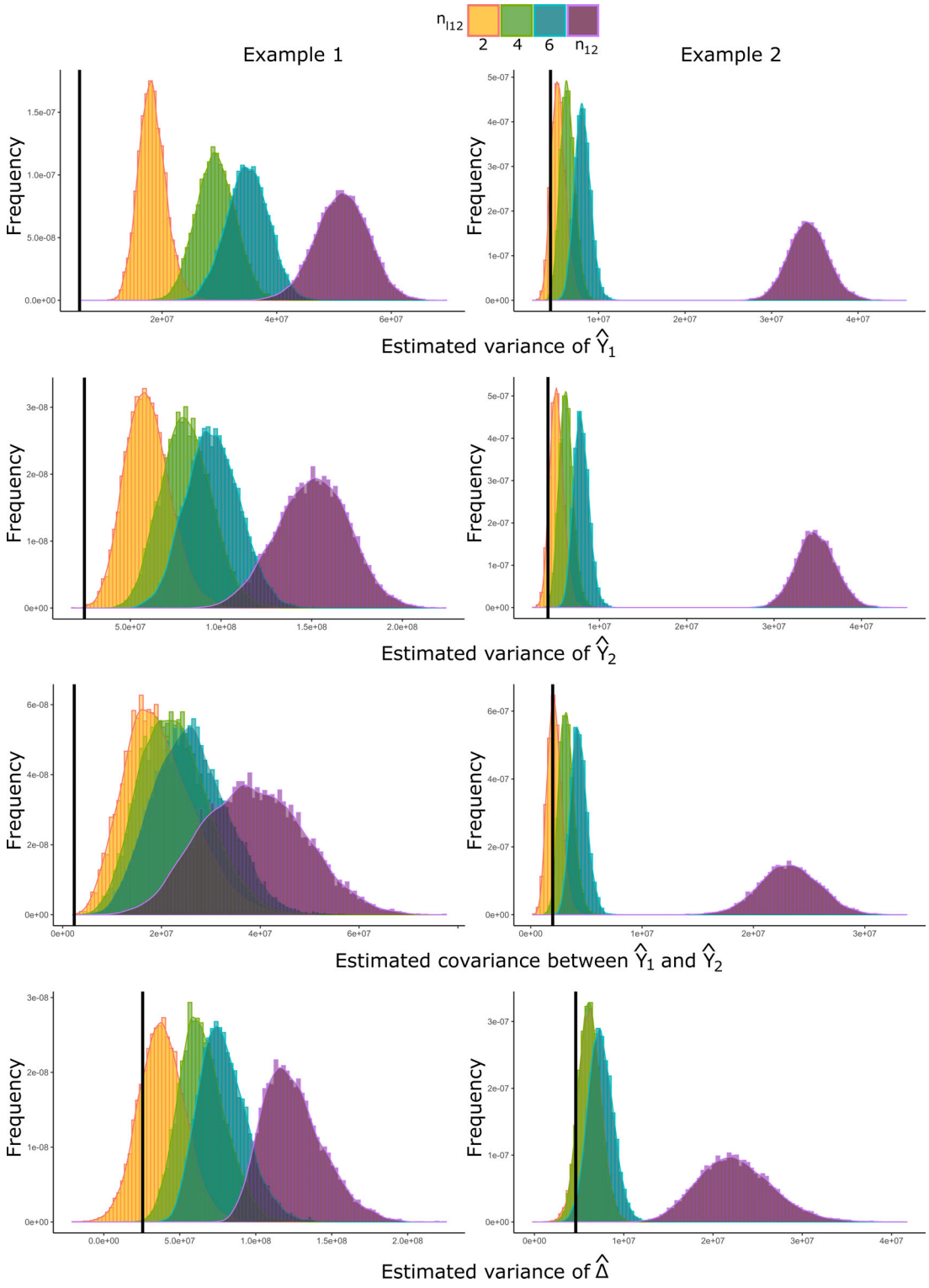

FIGURE 3 Comparing the estimators for different neighborhood sizes. The bold black vertical lines represent the empirical variances/covariances.

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

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

ACKNOWLEDGMENTS

The authors are grateful to an associate editor and three referees for valuable comments that improved the article.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Anton Grafström https://orcid.org/0000-0002-4345-4024

REFERENCES

Babcock, C., Finley, A. O., Gregoire, T. G., & Andersen, H. E. (2018). Remote sensing to reduce the effects of spatial autocorrelation on design-based inference for forest inventory using systematic samples. arXiv:2018 https://doi.org/10.48550/arXiv.1810.08588

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

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

Berger, Y. G., & Priam, R. (2016). A simple variance estimator of change for rotating repeated surveys: an application to the European union statistics on income and living conditions household surveys. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *179*(1), 251–272.

Breidt, F. J., & Fuller, W. A. (1999). Design of supplemented panel surveys with application to the national resources inventory. *Journal of Agricultural, Biological, and Environmental Statistics*, *4*(4), 391–403.

De Leeuw, E. D., Hox, J. J., & Dillman, D. A. (2008). *International handbook of survey methodology*. Taylor & Francis Group/Lawrence Erlbaum Associates.

Foss, K. H., Berget, G. E., & Eidsvik, J. (2022). Using an autonomous underwater vehicle with onboard stochastic advection-diffusion models to map excursion sets of environmental variables. *Environmetrics*, *33*(1), e2702. https://doi.org/10.1002/env.2702

Frank, B., & Monleon, V. J. (2021). Comparison of variance estimators for systematic environmental sample surveys: considerations for post-stratified estimation. *Forests*, *12*(6), 772. https://doi.org/10.3390/f12060772

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

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

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

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

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

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

Kalton, G. (1983). *Introduction to survey sampling (Quantitative applications in the social sciences)*. Sage Publications.

Kish, L. (1965). *Survey sampling*. Wiley.

Lowther, A. P., Killick, R., & Eckley, I. A. (2023). Detecting changes in mixed-sampling rate data sequences. *Environmetrics*, *34*(1), e2762. https://doi.org/10.1002/env.2762

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

Patterson, H. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Series B (Methodological)*, *12*(2), 241–255.

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

Scott, C. T. (1998). Sampling methods for estimating change in forest resources. *Ecological Applications*, *8*(2), 228–233.

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

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

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

Wang, Z., & Zhu, Z. (2019). Spatiotemporal balanced sampling design for longitudinal area surveys. *Journal of Agricultural, Biological, and Environmental Statistics*, *24*(2), 245–263.

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

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

How to cite this article: Zhao, X., & Grafström, A. (2024). Estimation of change with partially overlapping and spatially balanced samples. *Environmetrics*, *35*(1), e2825. https://doi.org/10.1002/env.2825