NRC Research Press

ARTICLE

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

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

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

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

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

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

Introduction

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

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

Received 6 March 2017. Accepted 12 May 2017.

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

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

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

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

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

The new sampling strategy

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

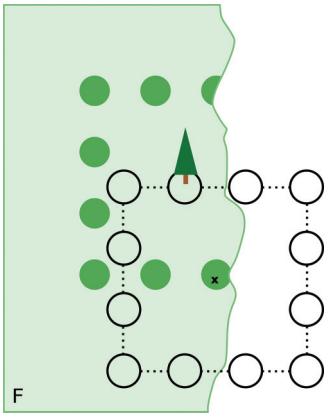


Fig. 1. An example of an inclusion zone. The inclusion zone K for the tree consists of the darker circles intersected by the surface of the forest; the circles connected with dots represent a cluster. Any cluster $\mathcal{C}(\mathbb{X})$ with its center $\mathbb{X}$ within K, such as the one in the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

A sampling strategy for continuous populations

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

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

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

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

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

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

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

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

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

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

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

Double sampling approach to achieve spatial balance and select representative samples

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

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

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

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

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

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

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

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

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

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

Measuring the spatial balance for continuous populations

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

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

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

Local pivotal method

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

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

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

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

Example for a one-dimensional auxiliary space

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

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

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

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

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

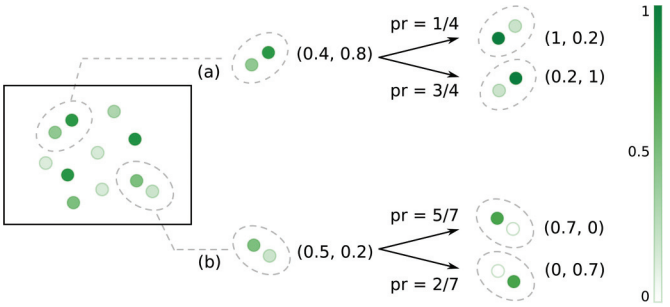
$$\hat{V}_{\text{LPM}}(\hat{Y}) = \frac{1}{2}\sum_{\mathbb{X}\in S_F}\left[\frac{Y(\mathbb{X})}{\pi(\mathbb{X})} - \frac{Y(\mathbb{X}')}{\pi(\mathbb{X}')}\right]^2$$

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

Swedish NFI and the current sampling strategy

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

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



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

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

Implementation of the new strategy in Sweden

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

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

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

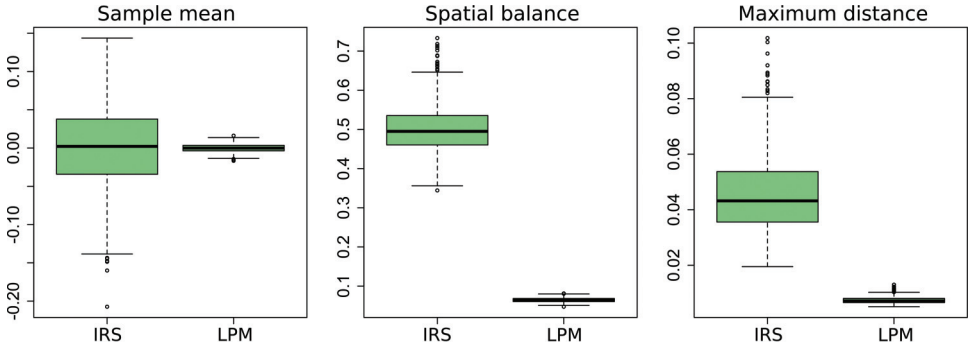


Fig. 4. Illustration of the selected region. [Colour online.]

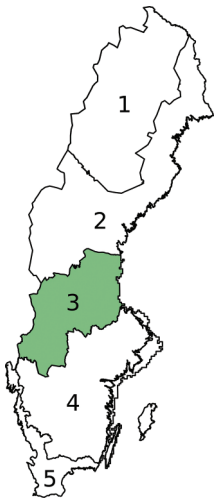
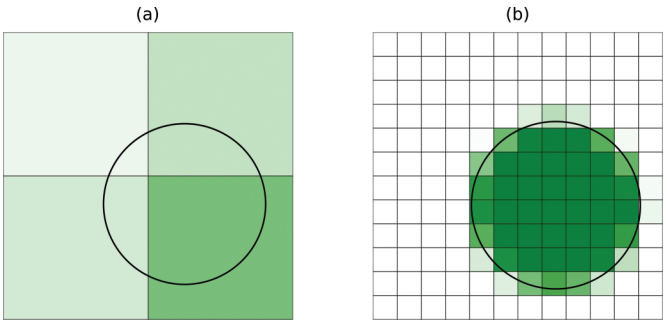


Fig. 5. Illustration of how we derive the plot total of auxiliaries for a 7 m radius plot. Each cell receives a weight proportional to the area of its intersection with the plot, which correlates with the intensity of the colour in the figure. (*a*) An example for the tree height and the basal area,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



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

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

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

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

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

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

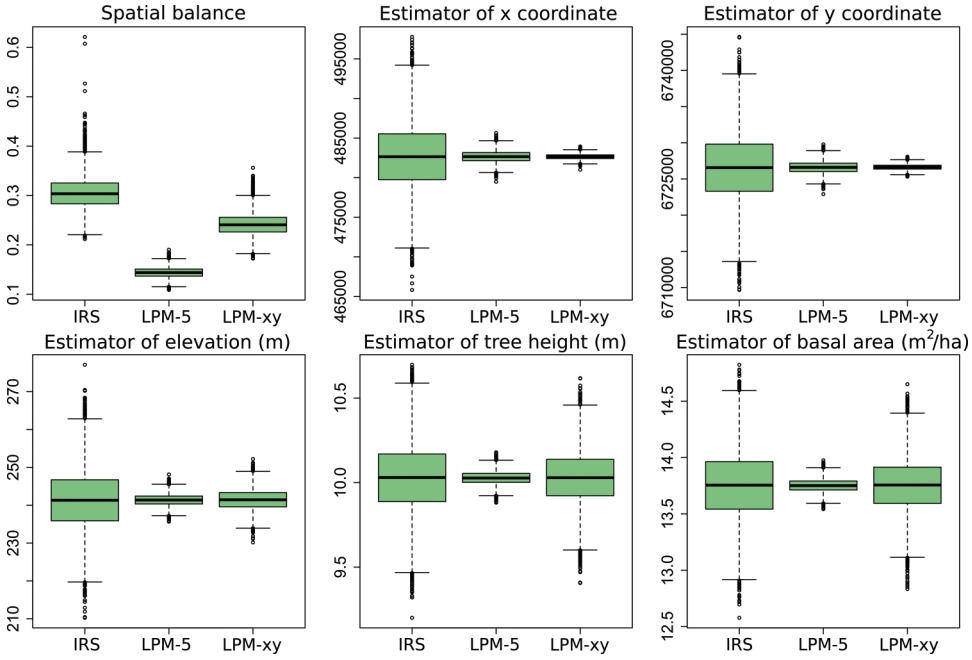
Note: First-phase sample size is 100 000, second-phase sample size is 360, and 10 000 samples were generated. LPM-5, local pivotal method with all five auxiliary variables; LPM-*xy*, local pivotal method with only *xy*-coordinates; IRS, independent random sampling. The variance ratios presented are called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

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

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

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

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



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

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

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

Conclusion and discussion

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

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

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

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

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

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

Acknowledgements

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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