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 representativeness 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.

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 over time, 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.


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

Copyright remains with the author(s) or their institution(s). Permission for reuse (free in most cases) can be obtained from RightsLink.
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 NFI s (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 Liscic 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 (Wolffowitz 1954; Deharty 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 employ for the second-phase sample selection, and finally, provide an illustrative example of the proposed strategy.

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 \( N_f \) objects (e.g., trees) located in \( F \), the \( N_f \) objects are represented by points. Let \( U = \{1, \ldots, i, \ldots, N_f\} \) be the identifiers for the \( N_f \) objects, and let \( S_j \subseteq U \) denote the probability sample of identifiers for the selected objects. The inclusion probability of object \( i \) to be sampled is defined as \( \pi_i = \Pr(i \in S_j) \). 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 \( S_n = \{X_1, X_2, \ldots, X_n\} \). The (prescribed) sampling intensity is \( \pi(X) = \sum_{i \in F} f_i(X) \), where \( f_i(X) \) is the marginal probability density function of \( X_i \) and moreover, \( \pi(X) > 0 \) for \( X \in F \) and \( \pi() = 0 \) outside \( F \). The sampling intensity plays the same role as the inclusion probabilities play in finite population sampling. We have

\[
 n = \int_F \pi(X) dX
\]

for a design of a fixed size \( n \).

When using clusters with a given configuration and a fixed orientation, the inclusion zone \( K_i \subseteq F \) for a tree \( i \) on location \( X_i \) can be expressed as \( K_i = K(X_i) = \{X \in F : X_i \in C(X_i)\} \), where \( C(X) \) is a cluster centered on \( X \). 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 \( Y(\mathcal{X}) \) of the target variable. For this article, we define the density function as a weighted sum of \( y_s \) over the objects that are selected:

\[
Y(\mathcal{X}) = \sum_{i \in S} \frac{I(\mathcal{X})y_i}{\ell(K)}
\]

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

\[
Y = \int Y(\mathcal{X})d\mathcal{X} = \int \sum_{i \in S} \frac{I(\mathcal{X})y_i}{\ell(K)}d\mathcal{X} = \sum_{i \in S} \frac{y_i}{\ell(K)} \int I(\mathcal{X})d\mathcal{X} = \sum_{i \in S} y_i
\]

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

\[
\tilde{Y} = \sum_{\mathcal{X} \in \mathcal{E}} \frac{Y(\mathcal{X})}{\pi(\mathcal{X})}
\]

\[
V_{SY}(\tilde{Y}) = \frac{1}{2} \int_{\mathcal{F}} \left[ \frac{\pi(\mathcal{X})}{\pi(\mathcal{X}')} - \frac{\pi(\mathcal{X})}{\pi(\mathcal{X}')} \right]^2 d\mathcal{X} d\mathcal{X}'
\]

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

**Double sampling approach to achieve spatial balance and select representative samples**

If \( Y(\mathcal{X}) \) 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 \( \pi(\mathcal{X}) = n/\ell(F) \) and a spatially balanced second-phase sample of size \( n \), we first select a large sample \( S_n \) of size \( N \) with independent observations over \( F \), where \( N \gg n \), with the sampling intensity \( \pi(\mathcal{X}) = n/\ell(F) \). 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 \( \pi(\mathcal{X}) \).

Suppose we have \( p \) auxiliary variables available from any source that provides wall-to-wall data. They are defined as \( Z(\mathcal{X}) = \{Z_1(\mathcal{X}), ..., Z_p(\mathcal{X})\} \in \mathbb{R}^p \). Let \( Z^*(\mathcal{X}) \) 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 \( Y(\mathcal{X}) \).

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

\[
Z^*(\mathcal{X}) = \int_{\mathcal{X} \in F} \frac{I(\mathcal{X}') \in \mathcal{O}(\mathcal{X})}{\ell(K(\mathcal{X}'))} d\mathcal{X}'
\]

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

\[
\int_{\mathcal{X} \in F} Z^*(\mathcal{X}) d\mathcal{X} = \int_{\mathcal{X} \in F} \int_{\mathcal{O}(\mathcal{X})} \frac{I(\mathcal{X}')}{\ell(K(\mathcal{X}'))} d\mathcal{X}' d\mathcal{X} = \int_{\mathcal{X} \in F} \int_{\mathcal{O}(\mathcal{X})} I(\mathcal{X}') d\mathcal{X}' = \int_{\mathcal{X} \in F} Z^*(\mathcal{X}) d\mathcal{X}'
\]

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 \( x_i \) in the sample includes all points in the population closer to \( x_i \) than to any other sample point \( x_j \), \( j \neq i \). 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 \( v_i = \int_{\mathcal{X} \in F} \pi(\mathcal{X}) d\mathcal{X} \) is the total probability mass within the polytope \( p_i \). Additionally, all the \( v_s \) 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 \( Y(\mathcal{X}) \) well explained by those auxiliary variables (Grafström and Lundström 2013). Thus, for such targets, we achieve \( \tilde{Y} = Y \). 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 \( 0 < p_i < 1 \) and \( 0 < p_j < 1 \). The winner takes as much inclusion probability as possible from the loser. Thereafter, the winner has an updated inclusion probability \( p_w = \min(1, p_i + p_j) \), while the loser has the new inclusion probability \( p_l = p_i + p_j - p_w \). Thus, if \( p_i + p_j \geq 1 \), then \( p_w = 1 \) and the winner is included in the sample. If \( p_i + p_j < 1 \), then \( p_l = 0 \) 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

\[
\tilde{Y} = \sum_{i \in S} \frac{Y(\mathcal{X})}{\ell(K)}
\]

\[
V_{SY}(\tilde{Y}) = \frac{1}{2} \int_{\mathcal{F}} \left[ \frac{\pi(\mathcal{X})}{\pi(\mathcal{X}')} - \frac{\pi(\mathcal{X})}{\pi(\mathcal{X}')} \right]^2 d\mathcal{X} d\mathcal{X}'
\]

\[
B = \frac{1}{n} \sum_{i \in S} (v_i - 1)^2
\]
tobe included simultaneously, which forces the sample becoming pair. When nearby units compete for inclusion, they are unlikely able distribution be one-dimensional auxiliary variable space. Let the auxiliary variable space is multidimensional, we can use the spatial balance to measure how well a sample represents the sample resulting from the LPM fit the sampling distribution much better than the "best" samples selected by IRS. When the 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.

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 systematic 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 can then 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...
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 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 12.5 m × 12.5 m grid. (c) An example for elevation, which is available on a 2 m × 2 m grid. [Colour online.]

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

<table>
<thead>
<tr>
<th>Auxiliary variable</th>
<th>Design effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>x-coordinate</td>
<td>0.030</td>
</tr>
<tr>
<td>y-coordinate</td>
<td>0.032</td>
</tr>
<tr>
<td>Elevation</td>
<td>0.036</td>
</tr>
<tr>
<td>Tree height</td>
<td>0.036</td>
</tr>
<tr>
<td>Basal area</td>
<td>0.035</td>
</tr>
</tbody>
</table>

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.

Users can 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.]

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 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

\[
Z_{i}^*(x) = \sum_{i=1}^{n_i} \iint_{Z \in D_{i}^*} \frac{Z(x')}{f_{i}(x')} \, dx' = \sum_{i=1}^{n_i} \frac{1}{f_{i}(x)} \iint_{Z \in D_{i}^*} Z(x') \, dx' = Z(x)
\]

where \(C_i(x)\) is plot \(i\) in the cluster centered at \(x\), \(n_i\) is the number of plots in a cluster, and \(f_{i}(x)\) is the surface area of the inclusion zone of the center point of plot \(i\) in the cluster. The integral...
The approximation \( Z(\mathbf{x}) \) 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., \( Z(\mathbf{x}) = Z^*(\mathbf{x}) \). 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

\[
\int_{x' \in H \cap F} Z(\mathbf{x}') d\mathbf{x}'
\]
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


