



**A new sampling strategy for forest inventories applied to  
the temporary clusters of the Swedish NFI**

Journal:	<i>Canadian Journal of Forest Research</i>
Manuscript ID	cjfr-2017-0095.R1
Manuscript Type:	Article
Date Submitted by the Author:	12-May-2017
Complete List of Authors:	Grafström, Anton; Swedish University of Agricultural Sciences, Department of Forest Resource Management Zhao, Xin; Swedish University of Agricultural Sciences, Department of Forest Resource Management Nylander, Martin; Swedish University of Agricultural Sciences, Department of Forest Resource Management Petersson, Hans; Swedish University of Agricultural Sciences, Department of Forest Resource Management
Keyword:	Continuous population, Double sampling, Local pivotal method, Remote sensing, Sampling design
Is the invited manuscript for consideration in a Special Issue? :	N/A

SCHOLARONE™  
Manuscripts

1 **A new sampling strategy for forest inventories applied to**  
2 **the temporary clusters of the Swedish NFI**

3  
4 Anton Grafström<sup>1,2</sup>, Xin Zhao<sup>1,3</sup>, Martin Nylander<sup>1,4</sup>, Hans Petersson<sup>1,5</sup>

5  
6 <sup>1</sup>*Swedish University of Agricultural Sciences SLU, Department of Forest Resource Management,*  
7 *Skogsmarksgränd, SE-901 83 Umeå, Sweden.*

8 E-mail: <sup>2</sup>[anton.grafstrom@slu.se](mailto:anton.grafstrom@slu.se), <sup>3</sup>[xin.zhao@slu.se](mailto:xin.zhao@slu.se), <sup>4</sup>[martin.nylander@slu.se](mailto:martin.nylander@slu.se), <sup>5</sup>[hans.petersson@slu.se](mailto:hans.petersson@slu.se)

9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21

Draft

22

23

24

25

26

27 **Abstract**

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

42

43 **Keywords:** Continuous population; Double sampling; Local pivotal method; Remote sensing;  
44 Sampling design

## 45 **1 Introduction**

46 National forest inventories (NFIs) have evolved and developed, in some cases more than one hundred  
47 years, and the need for accurate national level information is more requested than ever (Tomppo *et al.*,  
48 2010, ch.1). Still the NFI-designs normally rest on traditional area based sampling, which spreads the  
49 sample units over the landscape. Often the sample units are systematically distributed and organised  
50 in clusters of circular plots. NFIs in general have a very low sampling intensity due to the large areas  
51 that need to be covered. In such a situation, it is inevitable that forest attributes vary rapidly across the  
52 landscape with respect to the low sampling intensity. This means that spreading the sample only  
53 geographically is not sufficient to ensure that the sample is representative of the population. With the  
54 intention of providing a more effective sampling design and thereby increasing the precision of  
55 estimates of forest attributes, we present a strategy for obtaining a more representative sample by  
56 using auxiliary information from remote sensing in the planning phase of a forest inventory. In recent  
57 years, e.g. assessments using LiDAR-techniques (Light Detection and Ranging) can provide quite up  
58 to date wall-to-wall coverage of remote sensing data. In some countries, such data is available even at  
59 national scale and may be used for distributing sample units efficiently for NFIs.

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

65 Auxiliary variables can be used in different ways in a sampling design. Common use includes  
66 stratification (e.g. Särndal *et al.*, 2003, ch.3 and ch.12), balancing (e.g. Deville and Tillé, 2004), using  
67 unequal probabilities or achieving a good spread of the sample (e.g. Stevens and Olsen 2004).  
68 Including the auxiliary variables in the design normally reduces the need for including the same  
69 variables in the estimators and can allow for a simpler analysis. A sampling design which uses  
70 auxiliary variables to spread the sample is particularly useful for multi-purpose inventories, such as  
71 NFIs (Grafström and Schelin, 2014). When a multi-purpose inventory is planned the choice of a  
72 robust design is especially important. Tillé and Wilhelm (2016) discusses principles for choice of

73 sampling design and state that “Indeed, if the response variable is correlated with the auxiliary  
74 variable, then spreading the sample on the space of auxiliary variables also spreads the sampled  
75 response variable. It also induces an effect of smooth stratification on any convex set of the space of  
76 variables. The sample is thus stratified for any domain, which can be interpreted as a property of  
77 robustness.” As demonstrated by e.g. Grafström and Ringvall (2013), use of auxiliary variables in an  
78 estimator can only partly compensate for neglecting the use of the same variables in the design.

79 Grafström and Ringvall (2013) and Grafström *et al.* (2014) have recently introduced different  
80 sampling designs for forest inventories that are able to select spatially balanced samples, which means  
81 that the samples are well spread in some space. We have now developed this theoretical framework  
82 further to meet the specific needs of forest inventories. Our framework includes using the continuous  
83 population approach which was first proposed for forest inventories by Mandallaz (1991), see  
84 also Eriksson (1995), Barabesi (2003, 2004), Mandallaz (2007, ch.4) and Gregoire and Valentine,  
85 (2008, ch.10). Following Cordy (1993), we can in this framework use a general sampling design for  
86 selection of clusters of any shape and with any prescribed sampling intensity function. However, we  
87 focus on the selection of representative samples which means that we match as closely as possible the  
88 sample distribution of a set of auxiliary variables to the population distribution. This is achieved  
89 through a double (or two-phase) sampling, where auxiliary responses are extracted for a very large  
90 first phase sample of clusters. For the second phase sample selection, we use the local pivotal method  
91 (LPM) by Grafström *et al.* (2012) to spread the sample. When using a constant sampling intensity, the  
92 LPM produces representative samples (Grafström and Schelin, 2014). Different implementations of  
93 the LPM can be found in the R package ‘BalancedSampling’ (Grafström and Lisic, 2016).

94 The new strategy is illustrated with an application, where we select the temporary clusters for the  
95 Swedish NFI. As auxiliary variables, we use a digital elevation model (DEM) and a recent nationwide  
96 forest attribute map of Sweden predicted using airborne laser scanning data and field data from the  
97 NFI (Nilsson *et al.*, 2016). When compared with two reference strategies (independent observations  
98 and geographically well-spread observations), through a Monte-Carlo simulation, it is evident that the  
99 new strategy succeeds in producing representative samples.

## 100 2 The new sampling strategy

101 For the new sampling strategy, a continuous population approach with double sampling is employed.  
102 In the first phase sample, a very large number  $N$  of clusters is selected by randomly and independently  
103 placing cluster centers in the region. For each cluster, the auxiliary information of the cluster mean is  
104 derived. According to the Glivenko-Cantelli theorem and its multivariate generalisations, the  
105 empirical distribution of the auxiliary variables in the first phase sample converges uniformly almost  
106 surely to the population distribution as the size of the sample increases (Wolfowitz, 1954; Dehardt,  
107 1971). Then, a smaller sample of size  $n$  is selected from the  $N$  clusters by the LPM in such a way that  
108 the distribution of the auxiliary variables in the second phase sample matches the distribution in the  
109 large first phase sample very closely. Thus, by using a very large first phase sample we make sure that  
110 the distribution of the auxiliary variables in the second phase sample is very close to the  
111 corresponding distribution in the population, which means that we obtain a sample that is  
112 representative of the auxiliary variable space. In this section, the new strategy as well as an example  
113 to illustrate the superiority of the new strategy to the reference strategies are presented. Section 2.1  
114 provides the general framework and the notation of a sampling strategy for continuous populations.  
115 Section 2.2 shows the framework and the notation of using auxiliary information in a double sampling  
116 approach. Section 2.3 introduces the definition of spatial balance. Section 2.4 focus on the LPM  
117 which we employ for the second phase sample selection. Finally, section 2.5 provides an illustrative  
118 example of the proposed strategy.

### 119 2.1 A sampling strategy for continuous populations

120 Consider a surface  $F$  which is assumed to be a subset of the Euclidean plane  $\mathbb{R}^2$  with its surface  
121 area  $\ell(F)$ . For a finite population consisting of  $N_T$  objects (e.g. trees) located in  $F$ , the  $N_T$  objects are  
122 represented by points. Let  $U = \{1, \dots, i, \dots, N_T\}$  be the identifiers for the  $N_T$  objects, and let  $S_T \subset$   
123  $U$  denote the probability sample of identifiers for the selected objects. The inclusion probability of  
124 object  $i$  to be sampled is defined as  $\pi_i = \Pr(i \in S_T)$ . The variable of interest, which is generally non-  
125 negative and bounded, is denoted by  $y_i$ . An important objective of a forest inventory is the estimation

126 of the population total  $Y = \sum_{i \in U} y_i$ . For forest inventories, since the sampling frame is indeterminable  
 127 for the units in  $U$ , the objects cannot be sampled directly. Instead, we select our sample from a  
 128 continuous population on  $F$  as described in e.g. Mandallaz (2007).

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

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

138 [Figure 1 about here]

139  
 140 There exist several ways to formulate the density function  $Y(\mathbf{x})$  of the target variable. For this article  
 141 we define the density function as a weighted sum of  $y_i$ s over the objects that are selected,

$$Y(\mathbf{x}) = \sum_{i \in U} \frac{I_i(\mathbf{x}) y_i}{\ell(K_i)}, \quad (1)$$

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

$$Y = \int_F Y(\mathbf{x}) d\mathbf{x} = \int_F \sum_{i \in U} \frac{I_i(\mathbf{x}) y_i}{\ell(K_i)} d\mathbf{x} = \sum_{i \in U} \frac{y_i}{\ell(K_i)} \int_F I_i(\mathbf{x}) d\mathbf{x} = \sum_{i \in U} y_i. \quad (2)$$

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

$$\hat{Y} = \sum_{\mathbf{x} \in S_F} \frac{Y(\mathbf{x})}{\pi(\mathbf{x})},$$

$$V_{SYG}(\hat{Y}) = \frac{1}{2} \iint_F (\pi(\mathbf{x})\pi(\mathbf{x}') - \pi(\mathbf{x}, \mathbf{x}')) \cdot \left( \frac{Y(\mathbf{x})}{\pi(\mathbf{x})} - \frac{Y(\mathbf{x}')}{\pi(\mathbf{x}')} \right)^2 d\mathbf{x}d\mathbf{x}',$$

148

149 where  $\pi(\mathbf{x}, \mathbf{x}')$  is the second order sampling intensity for a pair of points  $(\mathbf{x}, \mathbf{x}')$ .

## 150 2.2 Double sampling approach to achieve spatial balance and select representative samples

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

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

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

164 Suppose we have  $p$  auxiliary variables available from any source that provides wall-to-wall data.

165 They are defined as  $\mathbf{Z}'(\mathbf{x}) = (Z'_1(\mathbf{x}), \dots, Z'_p(\mathbf{x}))^T \in \mathbb{R}^p$ . Let  $Z'(\mathbf{x})$  be the single point response for the  
 166 auxiliary variables (i.e. the value for the grid-cell that contain the point). Thus, all single point  
 167 responses within one grid-cell have the same value for the auxiliary variable. To preserve the

168 relationship between the auxiliary and the target variables, it is ideal to derive the auxiliary response  
 169 in a similar way as  $Y(\mathbf{x})$ .

170 The point response of the cluster  $\mathcal{C}(\mathbf{x})$  is here defined as

$$Z^*(\mathbf{x}) = \int_{\mathbf{x}' \in F} \frac{I(\mathbf{x}' \in \mathcal{C}(\mathbf{x}))Z'(\mathbf{x}')}{\ell(K(\mathbf{x}'))} d\mathbf{x}'. \quad (3)$$

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

172

173

$$\begin{aligned} \int_{\mathbf{x} \in F} Z^*(\mathbf{x}) d\mathbf{x} &= \int_{\mathbf{x} \in F} \int_{\mathbf{x}' \in F} \frac{I(\mathbf{x}' \in \mathcal{C}(\mathbf{x}))Z'(\mathbf{x}')}{\ell(K(\mathbf{x}'))} d\mathbf{x}' d\mathbf{x} \\ &= \int_{\mathbf{x}' \in F} \frac{Z'(\mathbf{x}')}{\ell(K(\mathbf{x}'))} \int_{\mathbf{x} \in F} I(\mathbf{x}' \in \mathcal{C}(\mathbf{x})) d\mathbf{x} d\mathbf{x}' = \int_{\mathbf{x}' \in F} Z'(\mathbf{x}') d\mathbf{x}' \end{aligned} \quad (4)$$

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

### 175 2.3 Measuring the spatial balance for continuous populations

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

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

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

#### 189 **2.4 Local pivotal method**

190 The local pivotal method (LPM) has been shown to be one of the most effective methods in regards to  
 191 spreading the sample in auxiliary space, e.g. Benedetti *et al.* (2015, ch.7). By employing the LPM, we  
 192 can select samples whose empirical distribution matches the population distribution of the auxiliary  
 193 variables. Such samples are spatially balanced in the auxiliary space, leading to an approximate  
 194 balance for any target  $Y(\mathbf{x})$  well explained by those auxiliary variables (Grafström and Lundström,  
 195 2013). Thus, for such targets we achieve  $\hat{Y} \approx Y$ . When applying the LPM, spatial balance is achieved  
 196 by successively updating the inclusion probabilities for nearby units until they become inclusion  
 197 indicators, i.e. 0's and 1's, where the 0's indicate exclusions of the units and the 1's indicate  
 198 inclusions of the units.

199 In one step of the LPM, we randomly select one unit  $i$  and find its nearest neighbor  $j$ . The pair of  
 200 nearby units will compete with the (possibly updated) inclusion probabilities  $0 < \pi_i < 1$  and  $0 <$   
 201  $\pi_j < 1$ . The winner takes as much inclusion probability as possible from the loser. Thereafter, the  
 202 winner has an updated inclusion probability  $\pi_W = \min(1, \pi_i + \pi_j)$ , while the loser has the new  
 203 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  
 204 included in the sample. If  $\pi_i + \pi_j < 1$ , then  $\pi_L = 0$ , and the loser is excluded from the sample. A final  
 205 decision is made for at least one unit each step. The procedure for the competition is given by

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

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

210

211 [ Figure 2 about here ]

212

### 213 **2.5 Example for a one-dimensional auxiliary space**

214 To illustrate the proposed strategy, we provide an example for a one-dimensional auxiliary variable  
 215 space. Let the auxiliary variable distribution be  $Z \sim N(0,1)$ . We perform a simulation of 1000 random  
 216 samples of size  $n = 350$  with independent observations, and compared with 1000 first phase samples  
 217 of size  $N = 100000$  with independent observations, followed by a selection of second phase samples  
 218 of size  $n = 350$  using the LPM with probabilities  $\pi_i = n/N, i = 1, 2, \dots, N$ .

219 The results of the comparisons are presented in Figure 3 for variation of sample mean, spatial balance,  
 220 and maximum distance. The maximum distance is the maximum distance between the empirical  
 221 distribution function and the reference distribution, which was calculated by employing the One-  
 222 sample Kolmogorov-Smirnov test.

223

224 [ Figure 3 about here ]

225

226 For the LPM with a second phase sample of size 350, the variance of the sample mean corresponded  
 227 approximately to the variance of the sample mean of 35000 independent observations. Thus, for the

228 mean of the auxiliary variables, such balanced samples of size 350 are as good samples of size 35000  
 229 with independent observations. The mean of the spatial balance of the LPM was 0.065 and the mean  
 230 of the maximum distance was 0.007, compared with 0.499 and 0.046 for independent random  
 231 sampling (IRS), respectively.

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

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

$$\hat{V}_{LPM}(\hat{Y}) = \frac{1}{2} \sum_{\mathbf{x} \in S_F} \left( \frac{Y(\mathbf{x})}{\pi(\mathbf{x})} - \frac{Y(\mathbf{x}')}{\pi(\mathbf{x}')} \right)^2.$$

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

242

### 243 3 Swedish NFI and the current sampling strategy

244 The current Swedish NFI follows the strategy developed by Ranneby *et al.* (1987). The country was  
 245 divided into five strata with decreasing sampling intensities towards the north. Within each stratum,  
 246 clusters of circular plots are sampled. The clusters were quadratic or rectangular in shape, with a side  
 247 length varying from 300 to 1800 *m* between different parts of the country. The circular plots were  
 248 located along the sides of the cluster with fixed distance between plots within stratum. The within  
 249 stratum fixed distance between plots increased by latitude. The design was motivated by assumed  
 250 autocorrelation for relevant forest variables such as stem volume. In other words, the landscape

251 changes more rapidly in the south with mixed species forests while the boreal conifer forests in the  
252 north are more homogenous and often dominated by one species. Thus, longer distances between  
253 plots was needed in the north to obtain new information.

254 Two kinds of clusters are used: temporary ones and permanent ones. The temporary clusters are  
255 mainly intended to capture the current state of the forest and are only surveyed once, whereas  
256 permanent clusters primarily aim to capture changes and are resurveyed regularly (Tomppo et al.,  
257 2010 ch.35). The selections in different strata are independent, and the estimation for target variables  
258 is required at the stratum level. A sample of the survey clusters, systematically distributed over the  
259 whole country, is measured annually from early May to mid-October. A five-year inventory cycle is  
260 used, using five consecutive yearly inventories, and the estimates are calculated as a five-year moving  
261 average. Separate estimators are used for each year and each cluster type, and a weighting is used to  
262 calculate averages of both cluster types. Details about the estimators used in the Swedish NFI can be  
263 found in (Ranneby *et al.*, 1987) and (Fridman *et al.*, 2014, appendix A-C).

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

267

#### 268 **4 Implementation of the new strategy in Sweden**

269 To evaluate the potential improvement in efficiency by introducing the new sampling strategy in  
270 Sweden, a simulation was performed for selecting the positions of temporary clusters of the Swedish  
271 NFI. The efficiency of alternatively using two reference sampling strategies were compared with the  
272 new sampling strategy. The new sampling strategy, denoted LPM-5 (local pivotal method using five  
273 auxiliary variables), is in many ways similar to the previous strategy. We use the same geographical  
274 stratification and the same number of clusters. The main difference is that the new strategy uses  
275 auxiliary information in the sampling design to ensure that the selected clusters are more  
276 representative. As the first reference sampling strategy we use IRS where the clusters are randomly

277 and independently distributed over the area. The second reference sampling strategy (LPM- $xy$ ) is the  
278 local pivotal method with geographical spread, which represent a proxy for the current strategy. The  
279 reason for including IRS is that we then can see also the effect of geographical spread.

280 We selected Region 3 in the middle of Sweden as our study region, see Figure 4. In this region, the  
281 clusters consist of 12 circular plots of 7 m radius. The plots in a cluster are placed along a square  
282 formation with a side-length of 1500 m and with 500 m between plots. Five auxiliary variables were  
283 used simultaneously with equal weights to spread the sample for the new strategy. These variables  
284 were geographical coordinates of the cluster centre, the mean elevation of the cluster, the cluster mean  
285 tree height and the mean basal area. Elevation was derived from a digital elevation model (DEM),  
286 while tree height and basal area were derived from remote sensing information from airborne laser  
287 scanning (ALS) data, which were collected between 2009 and 2015. The forest variables were  
288 estimated by regression models combining NFI plot data with ALS data metrics, and were available  
289 on a nationwide map (Nilsson *et al.*, 2016).

290 For the first phase sample, a number of 100000 clusters were independently selected. For each such  
291 cluster of plots, the cluster response of the five auxiliary variables was derived. Then a subset of size  
292 360 of clusters was selected by the LPM-5 and the two reference designs, respectively. Spatial  
293 balance, design effects and estimators for the auxiliary variables were compared by a Monte-Carlo  
294 simulation.

295 [ Figure 4 about here ]

296  
297 Equation (3) can be employed to calculate the value of auxiliaries for the point response of a cluster.  
298 However, it is unpractical to use the expression of  $Z^*(\mathbf{x})$  directly, since it is difficult to integrate the  
299 function in the equation. As we match the distribution of the derived auxiliary response we are free to  
300 introduce any approximation to the auxiliary response.

301 The inclusion zones for a point within a plot vary less than they vary within a cluster. Hence, it is  
302 natural to set an equal value of the area of the inclusion zone for all points in the same plot. Then, the

303 response of the cluster can be calculated by a weighted sum over the plots. To achieve this, we  
 304 introduce an approximation by assuming all points in a plot have the same inclusion zone as the plot  
 305 center. The cluster response (3) can then be approximated as

$$\mathbb{Z}^*(\mathbf{x}) = \sum_{i=1}^{n_c} \int_{\mathbf{x}' \in \mathcal{C}_i(\mathbf{x}) \cap F} \frac{\mathbb{Z}'(\mathbf{x}')}{\ell(K(\mathbf{x}'))} d\mathbf{x}' \approx \sum_{i=1}^{n_c} \frac{1}{\ell_i(\mathbf{x})} \int_{\mathbf{x}' \in \mathcal{C}_i(\mathbf{x}) \cap F} \mathbb{Z}'(\mathbf{x}') d\mathbf{x}' = \mathbb{Z}(\mathbf{x}),$$

306 where  $\mathcal{C}_i(\mathbf{x})$  is plot  $i$  in the cluster centered at  $\mathbf{x}$ ,  $n_c$  is the number of plots in a cluster,  $\ell_i(\mathbf{x})$  is the  
 307 surface area of the inclusion zone of the center point of plot  $i$  in the cluster. The integral

$$\int_{\mathbf{x}' \in \mathcal{C}_i(\mathbf{x}) \cap F} \mathbb{Z}'(\mathbf{x}') d\mathbf{x}', \quad (5)$$

308 is the total of the single point response on plot  $i$  in the cluster. We get this plot total if we multiply cell  
 309 values with respect to intersected area of the plot. Figure 5 is an example of how we weight the grid  
 310 cells to calculate equation (5) of auxiliary variables derived from ALS and DEM, respectively. The  
 311 values of auxiliary variables for each grid-cell were available beforehand see, e.g., Nilsson *et al.*  
 312 (2016). The resolution of the grid-cell is  $12.5m \times 12.5m$  for the ALS data,  $2m \times 2m$  for the  
 313 elevation. The radius of each plot is 7 meters.

314

315 [ Figure 5 about here ]

316

317 Table 1 and Figure 6 demonstrate variance for the estimator of the five auxiliary variables with  
 318 respect to the three designs. Compared with IRS, the reduction of the variance was more than 95% for  
 319 all five auxiliary variables while using LPM-5. We have also reduced variance by more than 90% for  
 320 mean tree height and mean basal area, even compared to the design that spreads geographically  
 321 (LPM-xy). We can clearly see from the table, if we just spread the samples geographically, the  
 322 reduction of the variance was less than 45% of mean tree height and mean basal area compared with

323 IRS. The mean of the spatial balance was 0.144, 0.242, and 0.306 for LPM-5, LPM- $xy$  and IRS,  
324 respectively.

325

326 [ Table 1 about here ]

327

328 [ Figure 6 about here ]

329

### 330 **5 Conclusion and discussion**

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

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

345 For the application study of the new strategy in Sweden, the auxiliary variables we used for the  
346 sampling design are related to most of the target variables of NFIs. Therefore, adapting the NFI to the

347 proposed strategy will lead to visible improvements for the estimation of the related target variables.  
348 If a variable is not related with the auxiliaries, the new strategy will not make their estimation worse.  
349 The observed potential of using the new sampling strategy confirms the claims from earlier studies. In  
350 the article by Grafström and Ringvall (2013), another sampling design which is called the local cube  
351 method (LCM) confirmed the advantages from selecting spatially balanced samples. However, the  
352 LPM tend to produce slightly better spread than LCM, and we chose to prioritize a better spread due  
353 to the multipurpose nature of NFIs.

354 According to Henttonen and Kangas (2015), the optimal sampling strategy depends heavily on the  
355 purpose of the inventory, thus, prioritizing the forest characteristics is also needed if an optimal  
356 strategy is to be determined. For multipurpose forest inventories, when the number of characteristics  
357 of interest is large, the task gets more complicated. To choose a proper sampling strategy while using  
358 the auxiliary variables in the design, we need to consider the relationship between the auxiliary  
359 variables and the target variables. E.g. balanced samples are optimal for linear relationships and  
360 spatially balanced samples perform better for non-linear relationships (Grafström and Lundström,  
361 2013). The encouraging results of this study have led to a decision to implement this sampling  
362 strategy in all regions for the selection of temporary tracts within the Swedish NFI, starting from  
363 2018.

364

365

### 366 **Acknowledgements**

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

### 371 **References**

- 372 Barabesi, L. (2003). A Monte Carlo integration approach to Horvitz-Thompson estimation in  
373 replicated environmental designs. *Metron*, 61(3), pp. 355-374.
- 374 Barabesi, L. (2004). Replicated environmental sampling design and Monte Carlo integration methods:  
375 two sides of the same coin. In: *Proceedings of Invited paper, Proceedings of the XLII Meeting of the*  
376 *Italian Statistical Society*.
- 377 Benedetti, R., Piersimoni, F. and Postiglione, P. (2015). *Sampling Spatial Units for Agricultural*  
378 *Surveys*: Springer.
- 379 Cordy, C.B. (1993). An extension of the Horvitz—Thompson theorem to point sampling from a  
380 continuous universe. *Statistics & Probability Letters*, 18(5), pp. 353-362.
- 381 Dehardt, J. (1971). Generalizations of the Glivenko-Cantelli Theorem. *The Annals of Mathematical*  
382 *Statistics*, 42(6), pp. 2050-2055.
- 383 Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91(4),  
384 pp. 893-912.
- 385 Eriksson, M. (1995). Design-Based Approaches to Horizontal-Point-Sampling. *Forest Science*, 41(4),  
386 pp. 890-907.
- 387 Fridman, J., Holm, S., Nilsson, M., Nilsson, P., Ringvall, A.H. and Ståhl, G. (2014). Adapting  
388 National Forest Inventories to changing requirements—the case of the Swedish National Forest  
389 Inventory at the turn of the 20th century. *Silva Fennica*, 48(3), p. article id 1095.
- 390 Grafström, A., Lundström, N.L.P. and Schelin, L. (2012). Spatially Balanced Sampling through the  
391 Pivotal Method. *Biometrics*, 68(2), pp. 514-520.
- 392 Grafström, A. and Lundström, N.L.P. (2013). Why well spread probability samples are balanced.  
393 *Open Journal of Statistics*, 3(1), pp. 36-41.
- 394 Grafström, A. and Ringvall, H. (2013). Improving forest field inventories by using remote sensing  
395 data in novel sampling designs. *Canadian Journal Of Forest Research-Revue Canadienne De*  
396 *Recherche Forestier*, 43(11), pp. 1015-1022.
- 397 Grafström, A. and Schelin, L. (2014). How to Select Representative Samples. *Scandinavian Journal*  
398 *of Statistics*, 41(2), pp. 277-290.
- 399 Grafström, A., Saarela, S., and Ene, L. T. (2014). Efficient sampling strategies for forest inventories  
400 by spreading the sample in auxiliary space. *Canadian Journal of Forest Research*, 44(10), 1156-1164.
- 401 Grafström, A. and Lisic, J. (2016) *BalancedSampling: Balanced and spatially balanced sampling*. R  
402 *package version 1.5.2*. <http://www.antongrafstrom.se/balancedsampling/>.
- 403 Gregoire, T.G. and Valentine, H.T. (2008). *Sampling strategies for natural resources and the*  
404 *environment*. CRC Press.
- 405 Henttonen, H. and Kangas, A. (2015). Optimal plot design in a multipurpose forest inventory. *Forest*  
406 *Ecosystems*, 2(1), pp. 1-14.

- 407 Mandallaz, D. (1991). *A unified approach to sampling theory for forest inventory based on infinite*  
 408 *population and superpopulation models*. Diss. Zürich.
- 409 Mandallaz, D. (2007). *Sampling techniques for forest inventories*. CRC Press.
- 410 Nilsson, M., Nordkvist, K., Jonzén, J., Lindgren, N., Axensten, P., Wallerman, J., Egberth, M.,  
 411 Larsson, S., Nilsson, L., Eriksson, J. and Olsson, H. (2016). A nationwide forest attribute map of  
 412 Sweden predicted using airborne laser scanning data and field data from the National Forest  
 413 Inventory. *Remote Sensing of Environment*.
- 414 Ranney, B., Cruse, T., Björn, H., Härje, J. and Johan, S. (1987). Designing a new national forest  
 415 survey for Sweden. *Studia Forestalia Suecica*(177), pp. 1–29.
- 416 Stevens, D.L. and Olsen, A.R. (2004). Spatially Balanced Sampling of Natural Resources. *Journal of*  
 417 *the American Statistical Association*, 99(465), pp. 262-278.
- 418 Särndal, C.-E., Swensson, B. and Wretman, J. (2003). *Model assisted survey sampling*: Springer.
- 419 Tillé, Y. and Wilhelm, M. (2016). Probability Sampling Designs: Principles for Choice of Design and  
 420 Balancing. arXiv preprint arXiv:1612.04965.
- 421 Tomppo, E., Gschwantner, T., Lawrence, M. and McRoberts, R.E. (2010). *National Forest*  
 422 *Inventories Pathways for Common Reporting*. Dordrecht: Springer Netherlands.
- 423 Wolfowitz, J. (1954). Generalization of the Theorem of Glivenko-Cantelli. *The Annals of*  
 424 *Mathematical Statistics*, 25(1), pp. 131-138.

425

426

427

428

429

**Table 1.** Design effect for 5 auxiliary variables with respect to reference designs

Auxiliary variable	Design effect		
	$(\hat{V}_{LPM-5}/V_{IRS})$	$(\hat{V}_{LPM-5}/\hat{V}_{LPM-xy})$	$(\hat{V}_{LPM-xy}/V_{IRS})$
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

430 Note: First phase sample size is 100000, second phase sample size is 360, and 10000 samples were  
 431 generated. LPM-5, the local pivotal method with all five auxiliary variables; LPM-xy, the local  
 432 pivotal method with only xy-coordinates; IRS, independent random sampling. The variance ratios  
 433 presented are called design effects.

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

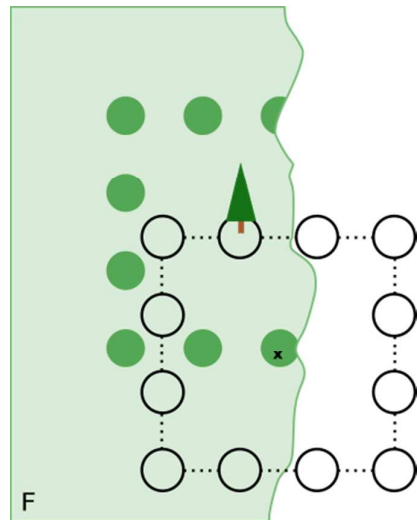
450

451

452

453

Draft

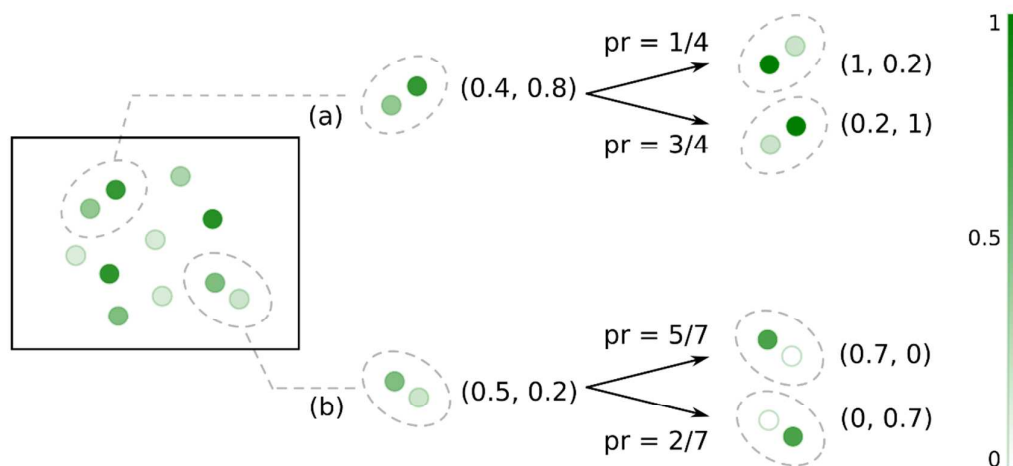


454

455

456 **Figure 1.** An example of inclusion zone. The inclusion zone  $K$  for the tree consists of the darker  
 457 circles intersected by the surface of the forest, the circles connected with dots represent a cluster. Any  
 458 cluster  $\mathcal{C}(x)$  with its center  $x$  within  $K$ , such as the one in the figure, includes the tree in one of the  
 459 plots.

460



461

462

463 **Figure 2.** One step in the LPM for a pair of nearby units  $i$  and  $j$ . The intensity of the colour correlates  
 464 with the inclusion probability. If  $\pi_i + \pi_j > 1$  (case a), then the winner receives probability 1 and will  
 465 definitely be included. If  $\pi_i + \pi_j < 1$  (case b), then the loser receives probability 0 and will definitely  
 466 be excluded.

467

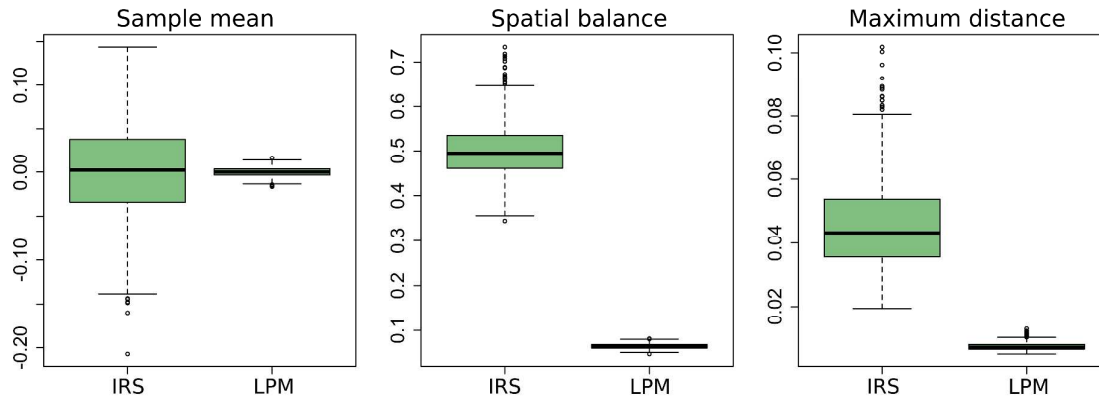
468

469

470

471

472



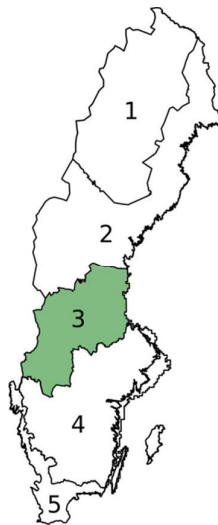
473

474

475 **Figure 3.** Results for the one-dimensional example. Box plots for sample mean, spatial balance and  
 476 maximum distance for IRS and the LPM respectively. All the results are based on a simulation of  
 477 1000 samples of size 350, and for the LPM we used a first phase sample of size  $N = 100000$ .

478

479



480

481

482

**Figure 4.** Illustration of the selected region.

483

484

485

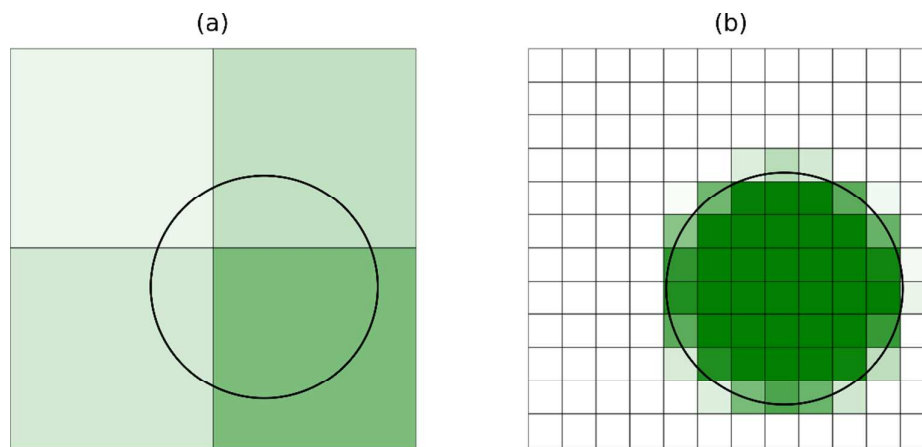
486

487

488

489

490

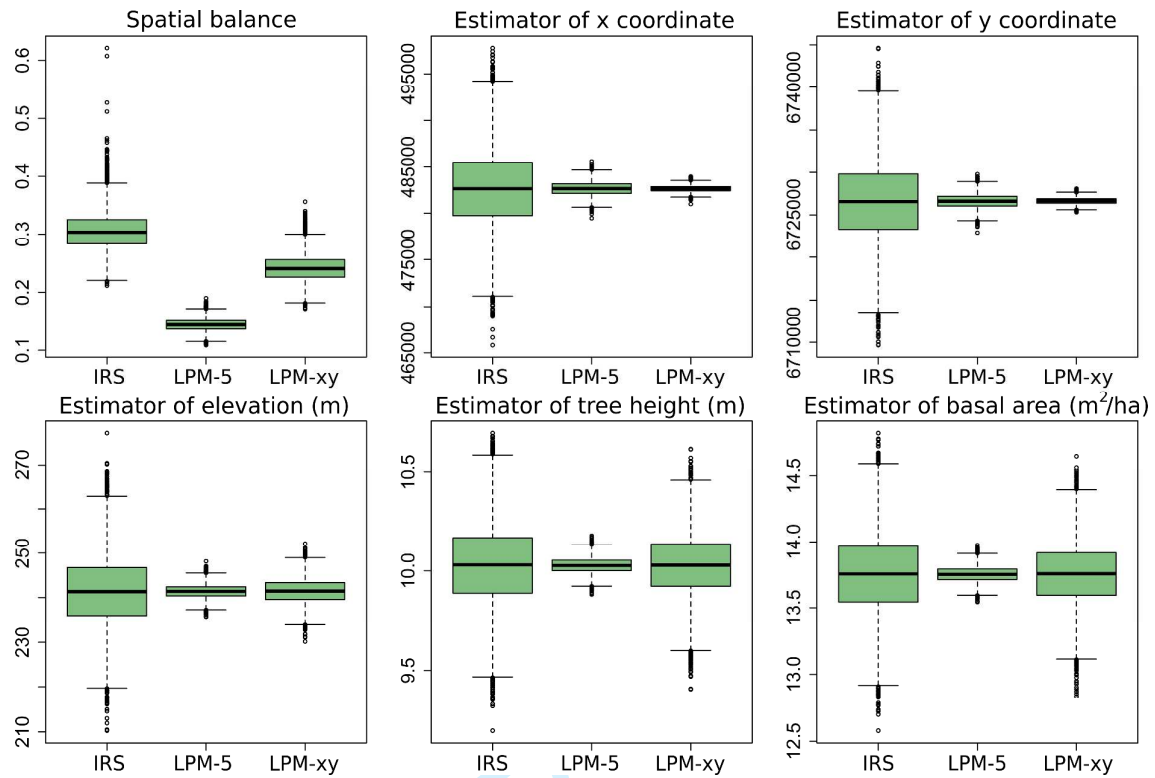


491

492 **Figure 5.** Illustration of how we derive the plot total of auxiliaries for a 7 m radius plot. Each cell  
493 receives a weight proportional to the area of its intersection with the plot, which correlates to the  
494 intensity of the color in the figure. (a) An example for the tree height and the basal area, which are  
495 available on a 12.5m x 12.5m grid. (b) An example for elevation, which is available on a 2m x 2m  
496 grid.

497

498



499

500

501 **Figure 6.** Box plots of spatial balance and estimators for the five auxiliary variables. LPM-5, the local  
 502 pivotal method with all five auxiliary variables; LPM-xy, the local pivotal method with only xy-  
 503 coordinates; IRS, independent random sampling.