Environmetrics

WILEY

RESEARCH ARTICLE OPEN ACCESS

Stratified, Spatially Balanced Cluster Sampling for Cost-Efficient Environmental Surveys

Juha Heikkinen | Helena M. Henttonen | Matti Katila | Sakari Tuominen

Natural Resources Institute Finland (Luke), Helsinki, Finland

Correspondence: Juha Heikkinen (juha.heikkinen@luke.fi)

Received: 2 May 2025 | **Revised:** 2 May 2025 | **Accepted:** 9 May 2025

Funding: This study was supported by the Natural Resources Institute Finland (Luke).

Keywords: generalized random-tessellation stratified design | local pivotal method | national forest inventory | simulated sampling | spatial autocorrelation | systematic sampling

ABSTRACT

Large-scale environmental surveys relying on intensive fieldwork are expensive, but survey sampling methodology offers several options to improve their cost-efficiency. For example, sites selected for field assessments can be arranged in clusters to reduce the time spent moving between the sites, and auxiliary data can be utilized to stratify the survey region and sample less important strata less densely. Geographically balanced and well-spread sampling can yield further improvements since the target variables of environmental surveys tend to be spatially autocorrelated. A combination of these ideas was illustrated and evaluated in the context of a national forest inventory, and alternative methods of spatially balanced sampling were compared. The main findings were that (i) both the local pivotal method and the generalized random-tessellation stratified design guaranteed a clearly better spatial regularity than systematic sampling when applied to fragmented regions resulting from stratification and (ii) they also ensured better global balance in unstratified sampling. In our case study, where stratification and sample allocation were based on high-quality auxiliary data, stratified sampling was clearly more efficient than unstratified for the primary survey target parameter. However, our results also illustrate that highly nonproportional sample allocation can be dangerous in a multi-purpose survey.

1 | Introduction

Environmental monitoring often relies on a sample survey based on an area frame, where field measurements are conducted at a sample of locations (e.g., by means of circular sample plots) selected from the survey region (Gregoire and Valentine 2007; Vallée et al. 2015; Prentius and Grafström 2022). To avoid selection bias, the field-assessed locations should be chosen by some probability sampling design (Särndal et al. 2003). While simple random sampling is notoriously inefficient (e.g., Fattorini et al. 2015), cost-efficient alternatives include stratified sampling (Rennolls 1989; Roux et al. 2013), spatially balanced sampling (e.g., Grafström, Schnell, et al. 2017), and cluster sampling (e.g., Vallée et al. 2015).

Stratified sampling has become increasingly attractive with the development of ever more precise high-resolution remote sensing and accurate positioning, enabling the construction and utilization of highly informative thematic maps of auxiliary variables covering the survey region (Nyyssönen et al. 1968; Rennolls 1989; Tomppo et al. 2014). Such maps can be used either to ensure that different sub-regions (strata), delineated from the auxiliary maps are represented in the sample in the same proportions as in the target population or to invest relatively more sampling

effort on more variable strata (Neyman 1934). Even in the former case, stratification improves efficiency, if within-stratum variances of the target variables of the survey are small compared to the differences in their stratum means (Cochran 1977, section 5.1).

The efficiency of spatially balanced sampling stems from the first law of geography (Tobler 1970); a larger proportion of the variation in the target population (survey region) can be captured in the sample by spreading it more regularly. Although systematic sampling is arguably the most common method of aiming at spatial balance (Mostafa and Ahmad 2018), several more flexible alternatives have been developed recently (Stevens and Olsen 2004; Grafström et al. 2012; Grafström and Tillé 2013; Benedetti and Piersimoni 2017; Chauvet and Le Gleut 2021; Prentius 2023).

Cluster sampling can be considered as the opposite of spatially balanced sampling: Groups of nearby locations are included in the sample. The motivation is that the obtained cost reductions can sometimes compensate for the loss in statistical efficiency caused by the increased spatial correlation within the sample (e.g., Vallée et al. 2015). For instance, the sample plots of the Finnish National Forest Inventory (NFI)—as those of many other NFI's (Tomppo et al. 2010, table 2.3)—are arranged in clusters so that all plots of one cluster are within walking distance and measuring one cluster should amount to one day's work (Tomppo et al. 2011, section 2.1). This reduces the fieldwork time per plot and thereby allows for a larger sample with a given cost. Clustering of sample plots can also be advantageous when surveying clustered populations (Prentius and Grafström 2022), but this idea will not be pursued here.

Räty et al. (2020) compared systematic sampling and the local pivotal method (LPM; Grafström et al. 2012) as alternative methods for spreading the sample regularly over the survey region. In that case study, the two methods were equally efficient. It is not obvious, however, whether this result holds when sampling a stratified population. Stratification based on high-resolution maps can lead to highly fragmented sampling strata, which may reduce the spatial balance of systematic samples.

Alternative sampling designs for an environmental survey can be compared based on replicated sampling from a synthetic population, for which the values of the survey target variable are known at all population units (e.g., Kermorvant et al. 2020). Ideally, the population units should form a dense grid of points covering the survey region. The values of the target variable are often obtained by some form of spatial interpolation of earlier survey data from the same region. As noted by Kermorvant et al. (2020), one needs to be aware of the risks associated with this approach. In particular, a reconstructed spatial surface of the target variable tends to be smoother, that is, less variable than the true surface. This can easily lead to an underestimation of the required sample size, if its determination is based purely on sampling simulations. On the other hand, the comparison of sampling designs can still be reliable, if the true spatial variation is sufficiently well featured in the synthetic population. An independent validation is important, if we also need to evaluate the absolute—in addition to the relative—efficiency of the planned survey.

The primary aim of this study was to assess the performance of systematic sampling against more recent spatially balanced designs in the context of a stratified environmental survey. Since our focus was on revealing potential problems of the commonly applied systematic sampling rather than on exhaustive comparison, we considered only two easily accessible alternative methods, LPM (Grafström et al. 2012) and generalized random-tessellation stratified design (GRTS; Stevens and Olsen 2004), as well as simple random sampling as the baseline. A sampling simulation study was conducted, aiming at an efficient NFI design for northernmost Finland, and the anticipated sampling variances were compared to those estimated from actual NFI data.

To gain a deeper understanding of the reasons behind the different performances of spatially balanced designs, both local and global aspects of spatial balance were studied from simulated stratum-specific and unstratified samples. Yet another aim of this paper was to present the NFI sampling design developed for the challenging study region to a wider readership. Stratified, spatially balanced cluster sampling has been found cost-efficient in surveying the fragmented target population and could serve as a framework for planning other surveys in similar environments.

2 | Material and Methods

2.1 | National Forest Inventory (NFI) of Finland

In Finland, as well as in many other countries, NFI is the main provider of nationwide information on the state and development of forest resources and several other aspects of the forest environment (Tomppo et al. 2010). The main targets of the Finnish NFI include land use and its changes, the volume of growing stock (stems of living trees), and its increment (Korhonen et al. 2024). Trees are measured in "forested land" (Table 1), which contains 75% of the land area of Finland (Korhonen et al. 2024) including all sites that fulfill the FAO forest definition (FAO 2023). The growing stock assessed in Finnish NFI includes all living trees growing on forested land and exceeding the minimum height limit of 1.35 m. In the most recently completed campaign, the 13th NFI of Finland (NFI13; 2019–2023), 62,266 field sample plots on land were assessed and nearly one million trees were measured (Korhonen et al. 2024). The survey target parameters considered in this study were the area of forested land (briefly, "forested area") and the total volume of growing stock on forested land ("total volume").

2.2 | Study Region

This study was originally motivated by the need to design NFI13 sampling in the Northernmost part of Finland, the municipalities of Enontekiö, Inari, and Utsjoki, with a land area of 28,158 km^2 on January 1, 2022 (National Land Survey of Finland 2022). The study region is located in the northern boreal zone close to the boundary of arctic tundra (Figure 1, left). While the fieldwork of recent NFI campaigns has covered other parts of the country with a systematic sampling grid during 5 years, field assessments of northernmost Finland only take place every tenth year (every second campaign), and a sparser network of field sample plots is

TABLE 1 | Land-use classes used in Finnish NFI (Korhonen et al. 2024) and their relationship to the terms used in this study. In NFI fieldwork, the land use class of all field sample plots is determined on-site, unless land use other than forestry can be determined without doubt from the base map. In MS-NFI, forestry land is separated from other classes using digital map data, and classification within forestry land is based on the combination of NFI field data and satellite images (Appendix A).

NFI land use class	This study	
Productive forest land	Forestry land	Forested land
Poorly productive forest land		
Unproductive forestry land		Unforested land
Other forestry land		
Agricultural land	Other than forestry land	
Built-up land		
Roads		
Power supply lines		
Inland water	Not land	
Sea		

used there. Due to high latitude and relatively high elevations, the study region is sparsely forested (forested land covers less than 50% of the land area) and most of the growing stock is clustered in specific parts of the region (Figure 1, right).

2.3 | NFI Designs in the Study Region

Two-phase stratified cluster sampling has been applied in the NFI of the study region since 1970 (Poso and Kujala 1971; Mattila 1985). The first-phase sample has been a dense square grid of clusters of field plots, for which auxiliary data are extracted from forest resource maps based on earlier NFI measurements supplemented with other data sources (Appendix A). The auxiliary data have been utilized to divide the first-phase sample into strata, and the second-phase sample, the clusters to be measured in the field, has been a stratified sub-sample of the first-phase sample. In Appendix B, we explain why it is important to stratify at the cluster level rather than at the plot level.

In the study region, the clusters of NFI13 (measured in 2022) consisted of eight sample plots with a 450 m minimum distance between the plot centers (Figure 2, left). The first-phase sample of clusters was constructed by locating cluster reference points on a square grid with 250 m × 250 m spacings between neighboring reference points (Korhonen et al. 2024). Thus, the minimal square containing all sample plot centers of one cluster is geographically strongly overlapping with similar squares of many other first-phase clusters, but all plot centers included in the sample were distinct, as illustrated in Figure 2 (right) for the 750 m × 750 m sub-grid (3 × 3 thinning) used in this study (Section 2.4). All those clusters on the grid that included at least one plot within the land area of the study region were included in the first-phase sample (Appendix A).

Each first-phase cluster was assigned to one of five strata characterized in Table 2; the exact definitions of the strata are provided in Appendix A. The aim was to stratify clusters according to their predicted shares of forested plots and the expected volume of growing stock. While clusters of Strata 1 and 2 were expected to contain mainly productive forest plots, those of Stratum 45 were expected to be mainly unforested. The variance of forested area was expected to be smallest in Strata 1 (> 90% forested) and 45 (> 85% unforested) and the variance of total volume was expected to decrease with increasing stratum number: In productive forest land, both small and large volumes are possible, depending on the

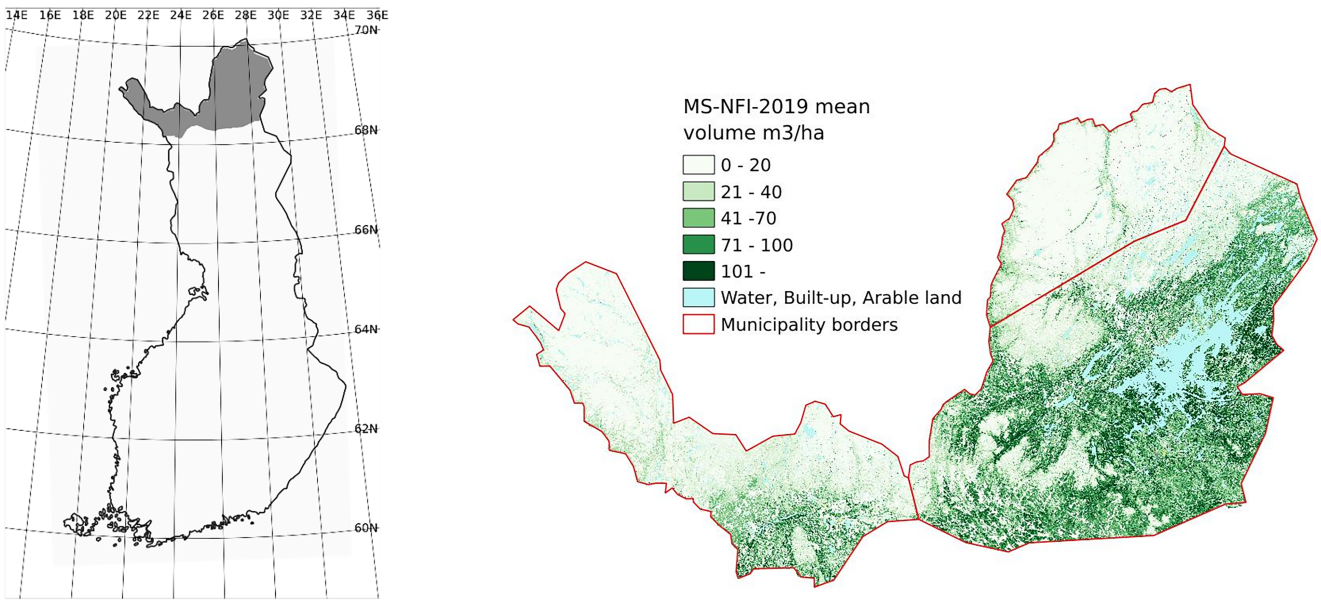

FIGURE 1 | Left: The location of the study area. Right: Map of the predicted mean volume of growing stock (m^3/ha); prediction method described in Appendix A. Digital map data: Contains data from the National Land Survey of Finland general map 1:4.5 M 06/2015 and municipal division 1:4.5 M 01/2018.

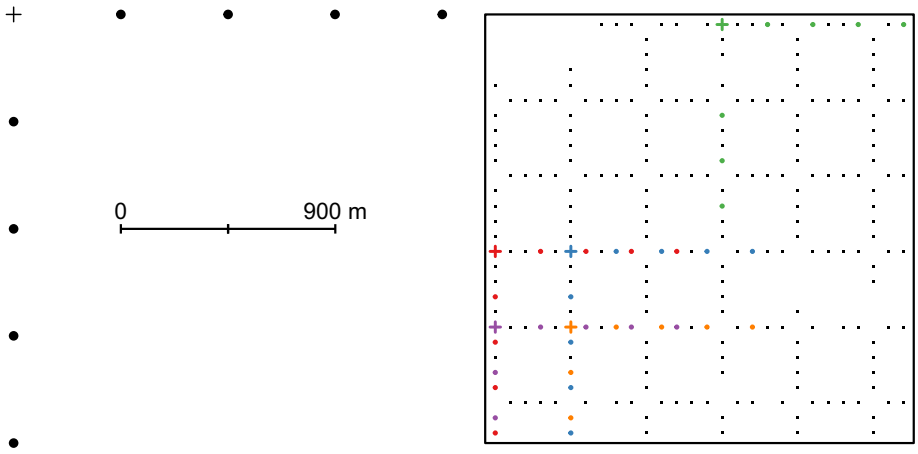

FIGURE 2 | Left: Relative locations of sample plot centers in one full cluster (dots) and the reference point of this cluster (cross). Right: Locations of all those sample plot centers in a 750 m × 750 m sub-grid (3 × 3 thinning) of the first-phase sample of NFI13 (250 m × 250 m grid) that are classified as land within a 4200 m × 4200 m square (the "missing" points are water). The plots are from 56 different clusters; plots of five clusters are shown as larger dots in different colors. Note that parts of the two southernmost colored clusters are outside the square.

TABLE 2 | Verbal description of strata, into which all NFI13 first-phase clusters of sample plots were assigned. For details of stratification, see Appendix A. The same stratification was applied both in the actual NFI13 sampling and in the sampling simulations reported in this paper.

Stratum	Description
0	Cluster is expected to contain plots with other land use than forestry.
1	Most of the plots are expected to be on productive forest land.
2	Approximately half of the plots are expected to be on productive forest land.
3	Most of the plots are expected to be on poorly productive forest land or on unproductive forestry land.
45	Most of the plots are expected to be on unproductive forestry land.

age of trees, but very large volumes are never achieved in poorly productive and unproductive land.

The special Stratum 0 was constructed to ensure the inclusion of other land-use plots in the sample. This was considered necessary since forestry land covers over 99% of the land area of Northernmost Finland (Korhonen et al. 2017, Appendix Table 1a, columns J and N/Ylä-Lappi), but monitoring of other land use is important for the greenhouse gas inventory, which uses NFI for this purpose. Between-cluster variance of volume predictions was actually greatest in this stratum (Table 3), which was not surprising, because it contains both forested plots and plots of other land use, where trees were not measured.

In each stratum, the second-phase (field) sample was a spatially systematic sub-sample of the first-phase sample (see Sections 2.4 and 2.5 for details). Based on the available fieldwork budget, the target size of the entire second-phase sample was set to $n^* = 195$ field plot clusters. The sub-sampling intervals varied between strata, aiming at Neyman allocation (for details, see Section 2.5 and Appendix C). The selected sample contained 196 clusters (Table 4).

2.4 | Sampling Simulations

The sampling simulations reported in this paper were targeted to evaluate recently developed methods of spatially balanced sampling as alternatives to the spatially systematic second-phase design applied in NFI13. We focused on second-phase sampling because—as explained and demonstrated in Appendix D—the impact of first-phase sampling is negligible, as long as the first-phase grid is sufficiently dense.

The synthetic population (Table 3), on which the evaluated sampling designs were simulated, consisted of the clusters whose reference points were located on a 750 m × 750 m square grid, which was a sub-grid (3 × 3 thinning) of the first-phase sample generated for NFI13 (Figure 3), stratified as explained in Section 2.3. One of the nine possible sub-grids was chosen arbitrarily (see Appendix D for simulations of all possible sub-grids). We have also conducted sampling simulations from the full first-phase grid but decided to report the—essentially identical—results for this smaller synthetic population in order to enable complete replicability in a reasonable computing time (all required data and computer codes are provided as Supporting Information on the journal's website).

The "true values" for each sample plot location p of all clusters c of the synthetic population were extracted from the same forest resource maps that were used to generate auxiliary data for NFI13 stratification (for details, see Appendix A). As a result, each plot location was classified as land within the study region, $u_p = 1$, or not, $u_p = 0$, and as forested, $f_p = 1$, or unforested, $f_p = 0$, and a prediction of mean volume of growing stock per hectare, v_p, was available for each forested plot ($f_p = 0$, if $u_p = 0$; $v_p = 0$, if $f_p = 0$). The required cluster-level values appearing in Equations (3-7) below were then the number of plots of cluster c classified

TABLE 3 | Characteristics and sampling intensities of the first-phase by stratum based on the 750 m x 750 m subgrid used as the synthetic population for sampling simulations (Section 2.4). A_h: Land area represented by stratum h (km^2), $A_{F,h}$: Forested area (km^2), $V_{F,h}$: Total volume (1000 m^3), $S_{\mathrm{R},h}$: Population "standard deviation" of volume between clusters (for exact definition, see Appendix C), and N_h: Number of clusters in stratum h. The sampling interval for systematic sampling, $k_h = \left[\pi_h^{-\frac{1}{2}}\right]$, was determined with Neyman allocation based on N_h and $S_{\mathrm{R},h}$ (Appendix C), and the median (med) over all possible systematic samples was used as the fixed sample size for the other methods. For unstratified sampling ("Stratum = all"), $\pi = 195/N$, where $N = 56{,}446$ is the total number of clusters in the synthetic population.

Stratum, h	A_h	$A_{F,h}$	$V_{F,h}$	$S_{\mathrm{R},h}$	N_h	π_h	k_h	Med
0	2592	1818	10,128	24.8	5386	0.0067	12	38
1	1947	1910	14,718	20.3	4441	0.0050	14	23
2	6175	5204	31,927	16.9	12,529	0.0047	15	56
3	8786	3994	13,476	12.4	17,199	0.0036	17	59
45	8658	1350	1676	3.3	16,891	0.0009	33	16
All	28,158	14,275	71,925		56,446	0.0035	17	196

TABLE 4 | Numbers of clusters, forested plots, and unforested plots by stratum in the NFI13 sample on the study region.

		Plots	
Stratum	Clusters	Forested	Unforested
0	29	143	45
1	27	165	3
2	67	398	59
3	62	194	248
45	11	14	66

as land, $n_c = \sum_{p\in c} u_p$, the number of forested plots among them, $m_c = \sum_{p\in c} f_p$, and the sum of the mean volume predictions of forested plots, $V_c = \sum_{p\in c} v_p$.

Both stratified and unstratified samples from the synthetic population U of clusters c were generated using the methods reviewed in Section 2.5. From each stratified sample s, composed by selecting one random sample s_h from each stratum U_h, forested area was estimated as

$$\hat{A}_{F,s} = \sum_h \hat{A}_{F,s,h} \tag{1}$$

and total volume as

$$\hat{V}_{F,s} = \sum_h \hat{V}_{F,s,h}, \tag{2}$$

where

$$\hat{A}_{F,s,h} = A_h \frac{\sum_{c\in s_h} m_c}{\sum_{c\in s_h} n_c} \tag{3}$$

and

$$\hat{V}_{F,s,h} = A_h \frac{\sum_{c\in s_h} V_c}{\sum_{c\in s_h} n_c} \tag{4}$$

are the stratum-specific estimates,

$$A_h = A \frac{\sum_{c\in U_h} n_c}{\sum_{c\in U} n_c} \tag{5}$$

is the land area represented by stratum h, and $A = 28{,}158$ km^2 is the known land area of the study region. The unstratified estimators from the unstratified samples s were simply

$$\hat{A}_{F,s} = A \frac{\sum_{c\in s} m_c}{\sum_{c\in s} n_c} \tag{6}$$

and

$$\hat{V}_{F,s} = A \frac{\sum_{c\in s} V_c}{\sum_{c\in s} n_c}. \tag{7}$$

Equations (3, 4, 6 and 7) are estimators of population ratios and, as such, slightly biased, but the bias becomes negligible in large samples (Cochran 1977, section 2.11). Furthermore, since all points on the land of the study region had equal inclusion probability density in first-phase sampling, (1), (2), (6), and (7) are also approximately unbiased estimators of the forest area and total volume in the study region, regardless of the first-phase variability in stratum areas A_h; the only source of bias is the ratio estimation (Cochran 1977, theorem 12.1).

2.5 | Sampling Methods

The simulated sampling methods included simple random sampling (SRS), spatially systematic sampling (SYS; Dunn and Harrison 1993), generalized random-tessellation stratified design (GRTS; Stevens and Olsen 2004), and local pivotal method (LPM; Grafström et al. 2012). While SRS serves as a spatially unbalanced baseline and SYS is the current method of the Finnish NFI, GRTS, and LPM represent more recent alternatives for spatially balanced sampling. These two methods were selected for this study because their implementations in the R environment were the most easily accessible.

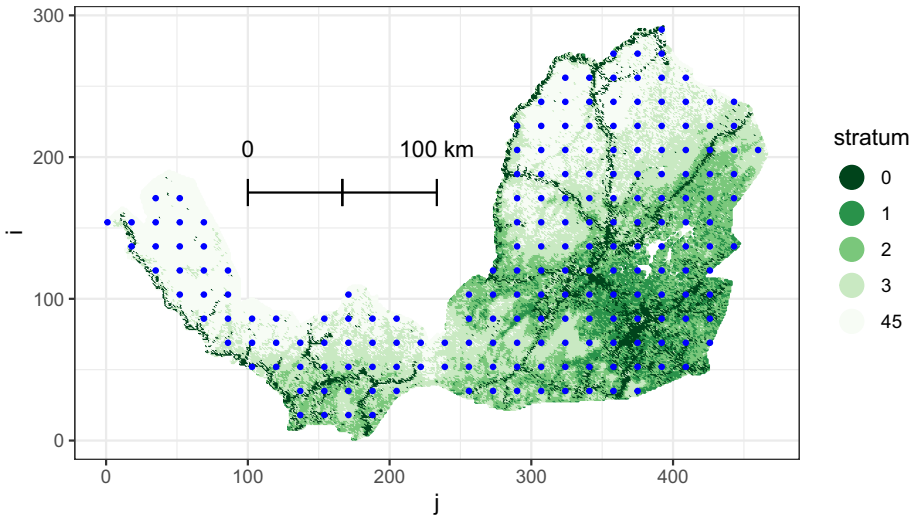

FIGURE 3 | The synthetic population used in sampling simulations (points in shades of green depicting locations of cluster reference points) and one unstratified systematic sample with sampling interval $k = 17$ ($n = 195$, blue dots).

To implement the spatially balanced designs, we attached to each element (cluster) of the synthetic population the grid indices (i, j), where $i \in \{1, 2, \ldots, N_i\}$ is the row number (from South to North), $j \in \{1, 2, \ldots, N_j\}$ the column number (from West to East), $N_i = 292$ the number of rows and $N_j = 467$ the number of columns of clusters in the synthetic population (Figure 3). This is equivalent to using the coordinates of the cluster reference points (Figure 2, left) because they were on a square grid and only relative distances matter in spatially balanced sampling.

A spatially systematic sample from the finite synthetic population is simply a sub-grid with a given sampling interval k. Given k and the starting point (i_0, j_0), randomly drawn from the set of $k \times k$ indices $\{(i, j) : i = 1, 2, \ldots, k, j = 1, 2, \ldots, k\}$, the unstratified SYS sample contains all clusters of the synthetic population—or its stratum—for which $(i, j) = (i_0 + ak, j_0 + bk)$ for some pairs (a, b) of integers $a, b = 1, 2, \ldots$ Thus, only k^2 distinct SYS samples are possible, and the sample size (number of clusters included in the sample) is necessarily random, that is, varies between different starting points depending on which clusters of the sub-grid belong to the synthetic population. Each of the k^2 samples is equally probable and each cluster of the synthetic population is included in one, and only one, of them, whereby the inclusion probability of each cluster is $\pi = k^{-2}$ and the expected sample size is $n^* = Nk^{-2}$, where $N = 56{,}446$ is the number of clusters in the synthetic population. The sampling interval was chosen so that the sample inclusion probabilities are as close as possible to a specified target inclusion probability, that is, $k = \left[\pi^{-\frac{1}{2}}\right]$, where $[x]$ stands for rounding to the nearest integer.

In the unstratified case, the target inclusion probability was determined as $\pi = \frac{n^*}{N}$, where $n^* = 195$ is the target sample size (see Section 2.3). For stratified SYS sampling, the stratum-specific target inclusion probabilities π_h were determined by Neyman allocation (Neyman 1934) based on the variability of the volume predictions between clusters of the same stratum (Table 3, Appendix C). The stratum-specific SYS samples with $k_h = \left[\pi_h^{-\frac{1}{2}}\right]$ were generated analogously to unstratified SYS samples with those clusters of the chosen sub-grid included in the sample, which belong to stratum h of the synthetic population.

The other methods included in this study allow for a fixed sample size. We considered that the fairest comparison is obtained by fixing their sample sizes n (unstratified sampling) and n_h (stratified sampling) to the median of the sizes of all possible SYS samples (Table 3).

An equal probability GRTS sample from a finite population $U = \{c_i \in R^2, i = 1, \ldots, N\}$ of geo-located elements is generated by mapping each element onto a unit-length segment $(x_i - 1, x_i] \subset (0, N] \subset R$ using restricted randomization that preserves spatial relationships as much as possible: $|x_i - x_j|$ is usually small, if $\|c_i - c_j\|$ is small. A systematic sample of n points y_k in $(0, N]$ is then selected (sampling interval N/n) and c_i is included in the GRTS sample from U, if $y_k \in (x_i - 1, x_i]$ for some $k = 1, \ldots, n$ (Stevens and Olsen 2004).

The basic idea of LPM is to create a strong negative correlation between the inclusion indicators of units that are close in distance so that they seldom appear together in a sample (Grafström et al. 2012). In practice, this is achieved by applying the pivotal method introduced by Deville and Tillé (1998) in such a way that at each step of the iterative method, sample inclusion or exclusion is decided for one or the other of a pair of nearby population units. Inclusion reduces the conditional inclusion probability of the "competitor," and exclusion increases it.

In comparisons reported by Grafström et al. (2012), LPM generally achieved somewhat better spatial balance than GRTS. This is not surprising, since LPM always searches the nearest neighbors for "competitions." In GRTS, on the other hand, some of the nearby points are mapped relatively far from each other; for example, addresses 022 and 200 in Figure 1 of Stevens and Olsen (2004).

All sampling simulations were implemented in the R environment (R Core Team 2022), SRS samples (without replacement)

were generated with the base function `sample`, LPM samples with the function `lpm1` of the package `BalancedSampling` (Grafström et al. 2022), and GRTS samples with the function `grts` of the package `spsurvey` (Dumelle et al. 2023). The cluster-level data and the R codes are provided as Supporting Information on the journal's website.

2.6 | Evaluation of Sampling Designs

To compare the sampling methods, we constructed all possible SYS samples from each stratum, as well as all possible unstratified SYS samples with the selected sampling intervals (k_h in Table 3), as well as 1000 random samples with the corresponding n or n_h's (med in Table 3; see Section 2.5) using the other methods (SRS, GRTS, and LPM); 1000 samples were found sufficient for the purposes of this paper based on the Monte Carlo confidence intervals reported in Tables 5 and 6. Each of the 1000 stratified SYS samples was generated by selecting randomly and independently one of the possible samples from each stratum; with the other methods, the t'th stratified sample, $t = 1, \ldots, 1000$, was the union of the t'th stratum-specific samples.

For $\theta \in \{A_F, V_F\}$, the design effect of design d in relation to a reference design d' was computed as the ratio of the empirical variances of estimates $\hat{\theta}_s$ over the samples s simulated with designs d and d':

$$\mathrm{DEFF}_{\hat{\theta}}(d, d') = \frac{\widehat{\mathrm{Var}}_d(\hat{\theta})}{\widehat{\mathrm{Var}}_{d'}(\hat{\theta})} \tag{8}$$

Since the variance of the sample mean is roughly inversely proportional to the sample size, the design effect can be interpreted as the ratio of anticipated sample sizes required for equally precise estimators of θ. For example, $\mathrm{DEFF}_{\hat{\theta}}(d, d') = 0.5$ means that the sample required with d' is twice as large as that required with d, or in other words, that the efficiency of d with respect to d' is 2. While the designs were compared to each other using the DEFF, the anticipated sampling uncertainty was reported in terms of the relative standard deviation (%)

$$\mathrm{rsd}_d(\hat{\theta}) = 100\,\frac{\sqrt{\widehat{\mathrm{Var}}_d(\hat{\theta})}}{\theta}, \tag{9}$$

where θ is the true value of the target parameter computed as in Equations (1), (2), (6), or (7) but with $s_h = U_h$ or $s = U$. For estimation of Monte Carlo confidence intervals of $\mathrm{rsd}_d(\hat{\theta})$, see Appendix E.

The relative standard errors anticipated by sampling simulations were also contrasted with the sampling errors estimated from the actual NFI13 field data (Korhonen et al. 2024). The layout of the NFI13 field sample plots in the study region is very similar to one of the simulated stratified SYS samples. The difference is — as discussed in Sections 2.3 and 2.4 — that a denser first-phase sample was used in the NFI13 design, leading to a slightly different allocation of the second-phase clusters between strata (Table 3, "med" vs. Table 4, "Clusters"). The stratified estimates of forested area and total volume were computed from NFI13 data in the same way as explained in Section 2.4 for the simulated samples. Uncertainty due to sampling was assessed using a variation of the approximate variance estimator proposed by Grafström and Schelin (2014), which is generally conservative (Räty et al. 2020). A detailed definition of the estimator is given in Appendix F.

To find reasons for the differences in the efficiencies of alternative sampling methods, we also computed the values of two measures of spatial balance from the simulated samples. Stevens and Olsen (2004) proposed to use the variance of the sizes of Voronoi polygons associated with the sample points as a measure of spatial regularity (local spatial balance) of sample s with respect the sample population U. In our finite population setting, this measure is defined as

$$\mathrm{SR}(s, U) = \frac{1}{n}\sum_{c \in s}\left(\frac{nN_c}{N} - 1\right)^2, \tag{10}$$

TABLE 5 | Means and relative standard deviations (Rsd, Equation 9) of stratified and unstratified estimates of the total volume and forested area over the whole study region. The difference between the mean and the corresponding synthetic population parameter (Table 3, "Stratum = all") is the anticipated bias and Rsd is the anticipated relative standard error. Monte Carlo 95% confidence intervals (inside parentheses) were computed as explained in Appendix E. Note that the Monte Carlo error associated with unstratified systematic sampling is 0, since exhaustive simulation was possible.

	Total volume		Forested area	
Sampling design	**Mean**	**Rsd, %**	**Mean**	**Rsd, %**
Unstratified				
SRS	71,952	7.601 (7.243, 7.942)	14,270	5.087 (4.846, 5.317)
SYS	71,918	4.440	14,274	2.785
LPM	71,915	4.230 (4.045, 4.407)	14,273	2.893 (2.749, 3.029)
GRTS	71,835	4.716 (4.514, 4.910)	14,251	3.196 (3.059, 3.327)
Stratified				
SRS	71,961	3.460 (3.307, 3.606)	14,269	4.176 (3.987, 4.357)
SYS	72,006	3.218 (3.083, 3.348)	14,282	3.569 (3.413, 3.718)
LPM	71,887	2.977 (2.850, 3.099)	14,258	3.239 (3.095, 3.376)
GRTS	71,932	2.949 (2.810, 3.081)	14,268	3.473 (3.322, 3.617)

TABLE 6 | Spatial regularity and global balance of unstratified SRS, SYS, and LPM samples covering the whole study region and of the stratum-specific samples. For the significance of differences in SR and computation of Monte Carlo 95% confidence intervals of GB (inside parentheses), see Appendix E. Note that the Monte Carlo error associated with systematic sampling is 0 since exhaustive simulation was possible.

		Spatial regularity, SR				Global balance, GB
Stratum	Design	Mean	Median	Min	Max	
Unstratified	SRS	0.326	0.320	0.198	0.580	9.59 (9.22, 9.94)
	SYS	0.029	0.029	0.018	0.037	2.18
	LPM	0.062	0.062	0.045	0.086	0.98 (0.94, 1.01)
	GRTS	0.125	0.125	0.084	0.192	1.56 (1.51, 1.62)
0	SRS	0.488	0.463	0.159	1.630	19.70 (19.02, 20.36)
	SYS	0.351	0.328	0.154	0.945	18.89
	LPM	0.112	0.109	0.050	0.201	4.28 (4.14, 4.42)
	GRTS	0.193	0.187	0.077	0.454	6.45 (6.19, 6.70)
1	SRS	0.381	0.360	0.090	1.187	12.89 (12.42, 13.34)
	SYS	0.293	0.271	0.083	0.893	12.77
	LPM	0.118	0.115	0.036	0.277	5.74 (5.56, 5.92)
	GRTS	0.183	0.176	0.049	0.531	6.82 (6.58, 7.05)
2	SRS	0.386	0.369	0.176	0.977	12.92 (12.41, 13.41)
	SYS	0.225	0.213	0.101	0.602	11.26
	LPM	0.092	0.090	0.043	0.177	2.54 (2.46, 2.62)
	GRTS	0.160	0.155	0.071	0.297	3.47 (3.35, 3.58)
3	SRS	0.381	0.363	0.146	0.970	16.00 (15.32, 16.64)
	SYS	0.202	0.196	0.098	0.459	11.95
	LPM	0.097	0.096	0.052	0.174	3.03 (2.93, 3.13)
	GRTS	0.175	0.171	0.082	0.367	4.50 (4.35, 4.65)
45	SRS	0.465	0.403	0.080	2.999	38.60 (37.11, 40.04)
	SYS	0.178	0.161	0.026	0.649	21.13
	LPM	0.114	0.110	0.033	0.283	10.05 (9.77, 10.33)
	GRTS	0.196	0.187	0.052	0.499	13.31 (12.85, 13.76)

where N is the number of clusters in U (the entire synthetic population or one of its strata), n is the number of clusters in sample s, and N_c is the number of those clusters in U that are closer to c than to any other cluster in s. Spatial regularity of each simulated stratum-specific and unstratified sample was computed using function `sb` of the `BalancedSampling` package (Grafström et al. 2022) and median values of spatial regularity were compared between designs using the Mann–Whitney U test (Appendix E).

Inspired by the general idea of balanced sampling (Deville and Tillé 2004), we also constructed a simple measure of "global spatial balance" of a design: the root mean squared difference of the means of cluster indices (coordinates) between the sample and the sampled population. Based on the T simulated samples from design d, global balance of d with respect to the sampled population U was estimated as

$$\mathrm{GB}(d,U)=\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\bar{i}_t-\bar{I}\right)^2+\left(\bar{j}_t-\bar{J}\right)^2}, \tag{11}$$

where $\bar{i}_t$ and $\bar{j}_t$ are the means of the indices of clusters included in the t'th sample and $\bar{I}$ and $\bar{J}$ are the population means of the indices; for confidence intervals of $\mathrm{GB}(d,U)$, see Appendix E. Apparently, this measure of spatial balance has not been explicitly proposed earlier, but the doubly balanced spatial sampling method of Grafström and Tillé (2013) essentially aims at simultaneous optimization of SR and GB.

For both measures, smaller values indicate better balance. Spatial regularity is optimized, when all n sample points represent equal portion $1/n$ of the population, "representing" a population element meaning being closer to it than any of the other sample points. If all $N_i \times N_j$ clusters of the first-phase grid would belong to the target population, then a centered SYS sample would yield the optimal regularity since the sizes of all Voronoi polygons would be equal. However, other designs may result in samples with better spatial regularity than that of the SYS samples, when the target population is irregular and/or fragmented. Global balance, on the other hand, is optimized when the sample means of coordinates are always equal to their population means. This is important when the survey target variable has a strong, monotonic spatial trend as a function of coordinates. In the extreme, where the target variable is a deterministic linear function of the coordinates, GB = 0 would result in zero sampling error.

3 | Results

The anticipated bias of all stratified and unstratified estimators was less than 0.2% (compare columns "Mean" of Table 5 to row "Stratum = all" of Table 3), that is, less than 10% of their anticipated standard errors (column "Rsd" of Table 5). When estimating the total volume, the anticipated design effect of stratification (combined with Neyman allocation) ranged from 0.21 (SRS) to 0.53 (SYS; Table 5). This means that an unstratified SYS sample would need almost twice the size of a stratified sample to yield equal precision; with SRS, almost a five-fold sample would be required. On the contrary, stratification improved the efficiency of the estimators of the forested area only when combined with SRS. This was expected, since the allocation of the samples into strata was optimized for the total volume and the strata with the greatest sampling intensities had, by design, the greatest proportions of forested land; for optimal estimation of the forested area, the less forested strata should have been sampled more intensively.

Designs aiming at spatial balance were always much better than SRS; the design effect ranged from 0.30 (forested area, unstratified SYS) to 0.87 (total volume, stratified SYS). With stratification, LPM led to more efficient estimators than SYS both for the total volume (DEFF = 0.86) and for the forested area (DEFF = 0.82). In other words, replacing SYS with LPM is expected to yield 14% savings in measurement costs, if a given minimum precision is required both for the forested area and for the total volume. This is obviously due to the better spatial regularity and global balance of within-stratum LPM (Table 6, Figure 4); LPM outperformed SYS consistently across all strata (Table 7).

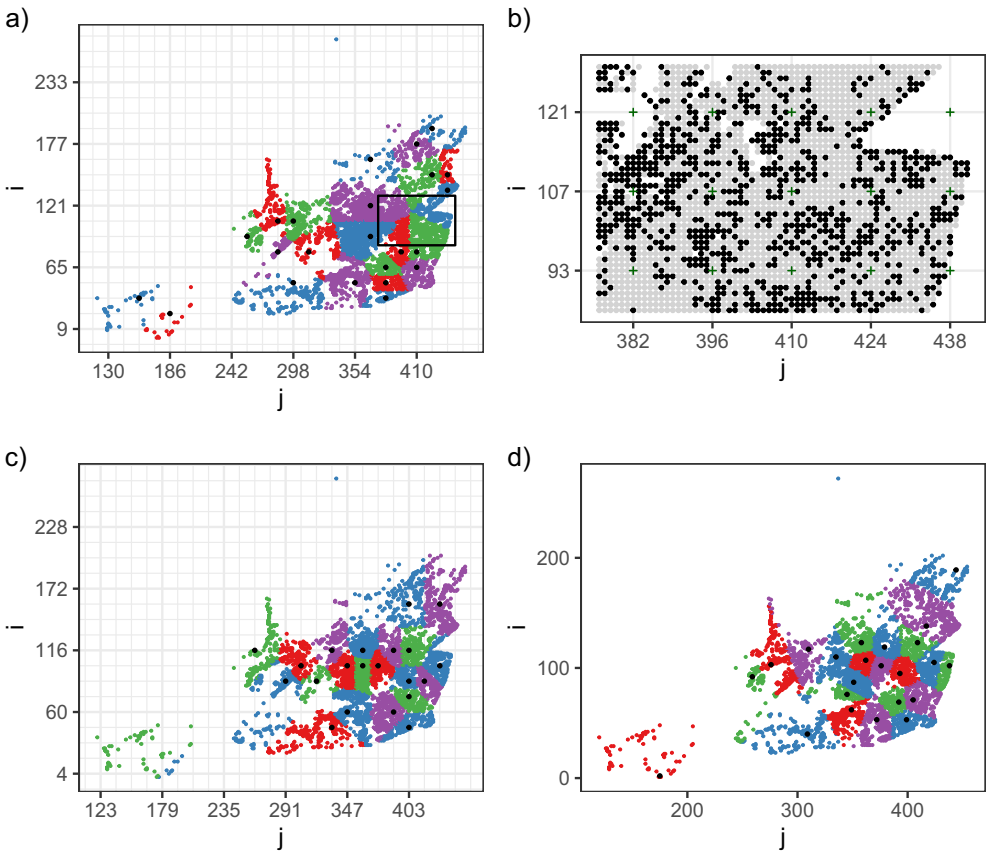

FIGURE 4 | Samples from stratum "1." (a) The systematic sample with the poorest spatial regularity (0.893). (b) A part of the square grid from which the sample of panel (a) was selected (the green crosses). The black dots depict all clusters of the synthetic population that belong to stratum "1," the gray ones those belonging to other clusters; within the rectangle, none of the clusters in the sampling grid belongs to stratum "1." (c) The systematic sample with the best spatial regularity (0.083). (d) The LPM sample with the median spatial regularity (0.118). In panels (a), (c), and (d), the colored dots show all clusters of the synthetic population that belong to stratum "1" and clusters in Voronoi polygons associated with different sample points (used to quantify spatial regularity) are shown in different colors.

A somewhat surprising result was that LPM led to a smaller sampling variance in volume estimates than SYS also in the unstratified setting. Since all unstratified systematic samples had better spatial regularity than any of the LPM samples, this is evidently due to the poorer global balance of SYS (Table 6), combined with the strong trend in volume across the region (Figure 5).

The field assessments of NFI13 revealed that cluster-level stratification was successful. On one hand, most of the clusters (120 out of 196) were full, that is, eight plots on land within the study region, and only 21 clusters had four plots or less. On the other hand, only a few unforested plots were found in stratum 1 and, apart from the special stratum 0, the proportion of unforested plots grew consistently with the stratum number (Table 4).

The stratum-specific NFI13 estimates of forested area and total volume were quite similar to their population means in the first-phase sample (Table 8 vs. Table 3), which indicates that the auxiliary data obtained before the measurements were strongly correlated with the current state assessed in the field. The relative sampling error of the total volume estimated from the NFI13 field data (3.16%, Table 8) was within the Monte Carlo confidence intervals of the anticipated sampling error (Table 5; stratified SYS).

The sampling error of the forested area estimate was significantly smaller than expected. This stems largely from the much smaller estimated than anticipated sampling variance in stratum 45 (Table 8 vs. Table 7). It should be noted, however, that the variance estimates for stratum 45 are based on as few as 11 clusters (Table 4).

TABLE 7 | Anticipated relative standard errors of stratum-specific estimators of the total volume and forest area, % of the true values.

	Total volume			Forested area		
Stratum	**SYS**	**LPM**	**GRTS**	**SYS**	**LPM**	**GRTS**
0	8.74	7.17	7.87	5.76	4.76	5.00
1	5.35	5.00	5.04	1.14	1.10	1.11
2	4.12	3.79	4.05	1.97	1.97	1.86
3	9.36	8.58	8.97	6.37	6.28	6.16
45	38.99	34.05	38.39	30.90	27.47	30.88

4 | Discussion

The case study reported in this paper highlights the potential of stratified cluster sampling for cost-efficient environmental surveys. Clusters of points or plots to be assessed in the field can be arranged so that the work is efficient. In our application, it meant that all plots of one cluster are within walking distance and the number of plots is such that the assessment of one cluster should amount to 1 day's work. With cluster-level stratification, intact clusters can be maintained, although strata are sampled with different intensities to focus the survey effort on the most important parts of the target region. In this case, we assumed that measurement costs are directly proportional to the number of clusters, which can be argued based on the relatively high cost of accessing the cluster in comparison to the actual measurements. If an estimate of the relative costs of measuring clusters in different strata were available, then they could be easily taken into account in the allocation (see, e.g., Tomppo et al. 2014).

In our sampling simulations, stratification worked well because we had relatively good auxiliary data—essentially

TABLE 8 | Stratum-specific and whole-region (stratified) estimates of total volume and forested area based on NFI13 results together with the estimated sampling error (SE), % of the estimate.

	Total volume		Forested area	
Stratum	**1000 m^3**	**SE, %**	**km^2**	**SE, %**
0	12,538	9.93	1941	4.44
1	16,318	7.03	1940	0.95
2	37,879	3.80	5394	1.23
3	14,939	9.02	3879	5.81
45	1378	24.78	1506	15.69
Whole region	83,052	3.16	14,659	2.35

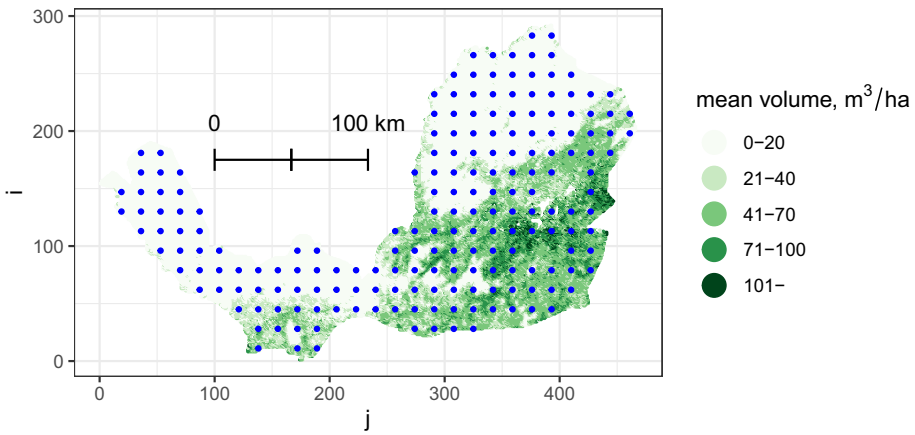

FIGURE 5 | An unstratified systematic sample with poor global balance yielding a 10% underestimation of the total volume. The sample mean of the j-coordinate is 282.0, while the population mean is 286.4 and the means of LPM samples range between 284.4 and 289.0.

high-resolution thematic maps of the main target variables of the survey based on a recent similar survey. For example, stratified systematic sampling was 90% more efficient than unstratified systematic sampling in the estimation of the total volume of the synthetic population. The results computed from the actual—stratified systematic—NFI sample measured in the field confirmed that stratification was successful also in the real-life application: Sampling variance estimated from the field data (3.16^2) was 51% of that anticipated for unstratified SYS (4.44^2).

An alternative to stratification would be to spread the sample evenly in a higher-dimensional space spanned by both the coordinates and other auxiliary variables (Grafström, Schnell, et al. 2017; Grafström, Zhao, et al. 2017). This approach has shown great potential to improve efficiency, but the associated variance estimates have sometimes been overly optimistic (Grafström, Schnell, et al. 2017; Räty et al. 2020). On the contrary, variance estimation for stratified sampling is robust, as long as the number of strata is moderate and the sample size sufficient in each stratum. In our case, variance estimation in stratum "45" turned out to be somewhat unreliable, but that stratum had a minor contribution to the total volume.

While highly nonproportional allocation (highly variable sample inclusion probabilities) can be extremely efficient in the estimation of that survey target for which it has been optimized, it can also be deleterious for other targets that are not strongly correlated with the primary target. In our case, nonproportional allocation combined with spatially balanced designs (SYS, LPM, and GRTS) led to poorer precision in the estimation of the forested area than that of their unstratified versions. In the present application, this was considered acceptable since we know that area estimation can be markedly improved by plot-level post-stratification (e.g., Haakana et al. 2020).

It is well known that spatially balanced sampling designs can be highly efficient in surveys of spatially autocorrelated target variables. In our case study, unstratified estimation of the total volume provided an extreme example of this: In comparison to SRS, the efficiencies of spatially balanced designs ranged from 2.6 (GRTS) to 3.2 (LPM), meaning that the sample size requirement is reduced to one third.

LPM turned out to be more efficient than SYS in volume estimation, whether stratified or unstratified, as well as in stratified estimation of forested areas. The obvious reason for the success of stratified LPM in comparison to stratified SYS is that it results in better spatial balance within the highly fragmented strata: Systematic sampling yields perfect spatial regularity over a rectangle containing the target population while LPM balances the sample with respect to the actual finite population.

But in the case of unstratified estimation of the total volume, we saw that spatial regularity alone is not enough: SYS samples were clearly more regular than LPM samples, but the latter still performed better. The difference in global spatial balance was identified as a plausible explanation; SYS can lead to a relatively large sampling error when the grid is not centered in relation to the survey region. Thus, LPM can be recommended as a safer alternative also in the unstratified case. This seems to contradict our earlier result in Räty et al. (2020), but the number of clusters considered in that study was much greater and the spatial trend in the target population was less evident.

The impact of first-phase sampling on anticipated variances was found negligible even with a much sparser grid than that used in NFI13 (Appendix D). However, a large first-phase sample is important so that the stratum weights A_h are stable. With the 750 m grid, their anticipated standard deviations were 1.5% or less. The actual first-phase sample of NFI13 (a 250 m grid) included a nine-fold number of clusters meaning that the standard deviation of stratum weights would be less than 0.5%. This can be considered negligible considering the much greater anticipated standard errors deviations due to second-phase sampling. In our application, dense first-phase sampling led to strongly overlapping clusters, but spatially regular second-phase sampling guarantees that the field plot clusters rarely or, in SYS, never overlap.

Although sampling simulation is a great tool to compare designs, it has an inevitable problem: The true population is never available; otherwise, the whole survey would be unnecessary. Any reasonable reconstruction tends to be a smoother version of the true population, whether produced with geostatistical interpolation as discussed by Kermorvant et al. (2020) or with nonparametric methods like k-NN as in this study. In our case, we assumed that the optical satellite images used in k-NN interpolation allow for a sufficiently faithful reconstruction of the spatial variation in our target variables so that at least the comparisons between different ways to aim for spatial balance lie on a strong basis.

We used the same underlying forest resource maps in stratification, in Neyman allocation, and to determine the "true values." Thus, both the stratum assignments and sample allocation between strata are somewhat "too good," although the impacts of this triple use of the same data were somewhat alleviated by using them in different resolutions (cluster-level stratification versus plot-level extraction of true values). However, our exercise was not that unrealistic, because land-use classes do not change very quickly and good prior information about within-stratum variances in tree volume is usually available from earlier inventories. For a more realistic anticipation of sampling uncertainty, one could use an older map for stratification and allocation.

In any case, sampling simulations can never confidently anticipate the absolute level of sampling uncertainty in a real-life inventory. It is thus crucial to validate—and re-scale, if necessary—the anticipations based on an independent validation with real field data, such as illustrated here.

Supporting Information for this article Appendix A–F containing further details of the methods and Supporting Information containing the data and R codes is available online at the journal's website.

Acknowledgments

We wish to thank Dr Kari Korhonen for his helpful comments on the first version of this manuscript. We are also grateful to an anonymous referee for two exceptionally thorough and constructive reviews. Digital

map data (Figure 1) National Land Survey of Finland, reproduced under license MML/VIR/−MYY/328/08. Open access publishing facilitated by Luonnonvarakeskus, as part of the Wiley - FinELib agreement.

Ethics Statement

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data and R codes used in the study will be made available online at the journal's website upon eventual publication. Comprehensive results of NFI13, a small part of which was used for Table 8, have been published in Korhonen et al. (2024). Data and R codes used in the study are included as Supporting Information Data S1.

References

Axelsson, P. 2000. "DEM Generation From Laser Scanner Data Using Adaptive TIN Models." *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 33: 110–117.

Benedetti, R., and F. Piersimoni. 2017. "A Spatially Balanced Design With Probability Function Proportional to the Within Sample Distance." *Biometrical Journal* 59: 1067–1084. https://doi.org/10.1002/bimj.201600194.

Chauvet, G., and R. Le Gleut. 2021. "Inference Under Pivotal Sampling: Properties, Variance Estimation, and Application to Tesselation for Spatial Sampling." *Scandinavian Journal of Statistics* 48: 108–131. https://doi.org/10.1111/sjos.12441.

Cochran, W. G. 1977. *Sampling Techniques*. 3rd ed. John Wiley & Sons.

Deville, J.-C., and Y. Tillé. 1998. "Unequal Probability Sampling Without Replacement Through a Splitting Method." *Biometrika* 85: 89–101. https://doi.org/10.1093/biomet/85.1.89.

Deville, J.-C., and Y. Tillé. 2004. "Efficient Balanced Sampling: The Cube Method." *Biometrika* 91: 893–912. https://doi.org/10.1093/biomet/91.4.893.

Dumelle, M., T. Kincaid, A. R. Olsen, and M. Weber. 2023. "Spsurvey: Spatial Sampling Design and Analysis in R." *Journal of Statistical Software* 105: 1–29. https://doi.org/10.18637/jss.v105.i03.

Dunn, R., and A. R. Harrison. 1993. "Two-Dimensional Systematic Sampling of Land Use." *Applied Statistics* 42: 585–601. https://doi.org/10.2307/2986177.

FAO. 2023. *Terms and Definitions: FRA 2025*. Forest Resources Assessment Working Paper 194. https://www.fao.org/3/cc4691en/cc4691en.pdf.

Fattorini, L., P. Corona, G. Chirici, and M. Pagliarella. 2015. "Design-Based Strategies for Sampling Spatial Units From Regular Grids With Applications to Forest Surveys, Land Use, and Land Cover Estimation." *Environmetrics* 26: 216–228. https://doi.org/10.1002/env.2332.

Grafström, A., J. Lisic, and W. Prentius. 2022. "BalancedSampling: Balanced and Spatially Balanced Sampling," R Package Version 1.6.3. https://cran.r-project.org/package=BalancedSampling.

Grafström, A., N. L. P. Lundström, and L. Schelin. 2012. "Spatially Balanced Sampling Through the Pivotal Method." *Biometrics* 68: 514–520. https://doi.org/10.1111/j.1541-0420.2011.01699.x.

Grafström, A., and L. Schelin. 2014. "How to Select Representative Samples." *Scandinavian Journal of Statistics* 41: 277–290. https://doi.org/10.1111/sjos.12016.

Grafström, A., S. Schnell, S. Saarela, S. P. Hubbell, and R. Condit. 2017. "The Continuous Population Approach to Forest Inventories and Use of Information in the Design: Continuous Population Approach." *Environmetrics* 28: e2480. https://doi.org/10.1002/env.2480.

Grafström, A., and Y. Tillé. 2013. "Doubly Balanced Spatial Sampling With Spreading and Restitution of Auxiliary Totals." *Environmetrics* 24: 120–131. https://doi.org/10.1002/env.2194.

Grafström, A., X. Zhao, M. Nylander, and H. Petersson. 2017b. "A New Sampling Strategy for Forest Inventories Applied to the Temporary Clusters of the Swedish National Forest Inventory." *Canadian Journal of Forest Research* 47: 1161–1167. https://doi.org/10.1139/cjfr-2017-0095.

Gregoire, T. G., and H. T. Valentine. 2007. *Sampling Strategies for Natural Resources and the Environment*. Chapman and Hall/CRC. https://doi.org/10.1201/9780203498880.

Haakana, H., J. Heikkinen, M. Katila, and A. Kangas. 2020. "Precision of Exogenous Post-Stratification in Small-Area Estimation Based on a Continuous National Forest Inventory." *Canadian Journal of Forest Research* 50: 359–370. https://doi.org/10.1139/cjfr-2019-0139.

Kermorvant, C., S. Coube, F. D'amico, N. Bru, and N. Caill-Milly. 2020. "Sequential Process to Choose Efficient Sampling Design Based on Partial Prior Information Data and Simulations." *Spatial Statistics* 38: 100439. https://doi.org/10.1016/j.spasta.2020.100439.

Korhonen, K. T., A. Ihalainen, A. Ahola, et al. 2017. *Suomen Metsät 2009–2013 Ja Niiden Kehitys 1921–2013 [In Finnish]*. Natural Resources Institute Finland. http://urn.fi/URN:ISBN:978-952-326-467-0.

Korhonen, K. T., M. Räty, H. Haakana, et al. 2024. "Forests of Finland 2019–2023 and Their Development 1921–2023." *Silva Fennica* 58: 24045. https://doi.org/10.14214/sf.24045.

Mäkisara, K., M. Katila, and J. Peräsaari. 2022. *The Multi-Source National Forest Inventory of Finland—Methods and Results 2017 and 2019*. Natural Resources and Bioeconomy Studies 90/2022. Natural Resources Institute Finland. http://urn.fi/URN:ISBN:978-952-380-538-5.

Mattila, E. 1985. *The Combined Use of Systematic Field and Photo Samples in a Large-Scale Forest Inventory in North Finland*. Vol. 131. Communicationes Instituti Forestalis Fenniae. http://urn.fi/URN:ISBN:951-40-0702-6.

Mostafa, S. A., and I. A. Ahmad. 2018. "Recent Developments in Systematic Sampling: A Review." *Journal of Statistical Theory and Practice* 12: 290–310. https://doi.org/10.1080/15598608.2017.1353456.

National Land Survey of Finland. 2022. "*Area of Finland by Municipality in January 1, 2022. National Land Survey of Finland*." [In Finnish and Swedish]. https://www.maanmittauslaitos.fi/sites/maanmittauslaitos.fi/files/attachments/2022/01/Vuoden_2022_pinta-alatilasto_kunnat_maakunnat.pdf.

Neyman, J. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97: 558–625. https://doi.org/10.2307/2342192.

Nyyssönen, A., S. Poso, and C. Keil. 1968. "The Use of Aerial Photographs in the Estimation of Some Forest Characteristics." *Acta Forestalia Fennica* 82: 1–35. https://doi.org/10.14214/aff.7177.

Poso, S., and M. Kujala. 1971. "Groupwise Sampling Based on Photo and Field Plots in Forest Inventory of Inari, Utsjoki, and Enontekiö [in Finnish With English Summary]." *Folia Forestalia* 132: 1–40. http://urn.fi/URN:NBN:fi-metla-201207191799.

Prentius, W. 2023. "Locally Correlated Poisson Sampling." *Environmetrics* 35, no. 2: e2832. https://doi.org/10.1002/env.2832.

Prentius, W., and A. Grafström. 2022. "Two-Phase Adaptive Cluster Sampling With Circular Field Plots." *Environmetrics* 33: e2729. https://doi.org/10.1002/env.2729.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org.

Räty, M., M. Kuronen, M. Myllymäki, A. Kangas, K. Mäkisara, and J. Heikkinen. 2020. "Comparison of the Local Pivotal Method and Systematic Sampling for National Forest Inventories." *Forest Ecosystems* 7: 54. https://doi.org/10.1186/s40663-020-00266-9.

Rennolls, K. 1989. "Design of the Census of Woodlands and Trees 1979–82," Forestry Commission, Edinburgh, Occasional paper 18. https://www.forestresearch.gov.uk/publications/archive-design-of-the-census-of-woodland-and-trees-1979-82/.

Roux, E., P. Gaborit, C. A. Romaña, R. Girod, N. Dessay, and I. Dusfour. 2013. "Objective Sampling Design in a Highly Heterogeneous Landscape—Characterizing Environmental Determinants of Malaria Vector Distribution in French Guiana, in the Amazonian Region." *BMC Ecology* 13: 45. https://doi.org/10.1186/1472-6785-13-45.

Särndal, C.-E., B. Swensson, and J. H. Wretman. 2003. *Model Assisted Survey Sampling*. Springer.

Stevens, D. L., and A. R. Olsen. 2004. "Spatially Balanced Sampling of Natural Resources." *Journal of the American Statistical Association* 99: 262–278. https://doi.org/10.1198/016214504000000250.

Tobler, W. R. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* 46: 234–240. https://doi.org/10.2307/143141.

Tomppo, E., T. Gschwantner, M. Lawrence, and R. E. McRoberts, eds. 2010. *National Forest Inventories: Pathways for Common Reporting*. Springer Netherlands. https://doi.org/10.1007/978-90-481-3233-1.

Tomppo, E., M. Haakana, M. Katila, and J. Peräsaari. 2008. *Multi-Source National Forest Inventory*. Springer Netherlands. https://doi.org/10.1007/978-1-4020-8713-4.

Tomppo, E., and M. Halme. 2004. "Using Coarse Scale Forest Variables as Ancillary Information and Weighting of Variables in k-NN Estimation: A Genetic Algorithm Approach." *Remote Sensing of Environment* 92: 1–20. https://doi.org/10.1016/j.rse.2004.04.003.

Tomppo, E., J. Heikkinen, H. M. Henttonen, et al. 2011. *Designing and Conducting a Forest Inventory—Case: 9th National Forest Inventory of Finland*. Springer. https://doi.org/10.1007/978-94-007-1652-0.

Tomppo, E., R. Malimbwi, M. Katila, et al. 2014. "A Sampling Design for a Large Area Forest Inventory: Case Tanzania." *Canadian Journal of Forest Research* 44: 931–948. https://doi.org/10.1139/cjfr-2013-0490.

Vallée, A., B. Ferland-Raymond, L. Rivest, and Y. Tillé. 2015. "Incorporating Spatial and Operational Constraints in the Sampling Designs for Forest Inventories." *Environmetrics* 26: 557–570. https://doi.org/10.1002/env.2366.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **DATA S1** Supporting Information.

Appendix A

Stratification of NFI13 First-Phase Sample and Determination of the "True Values" for the Synthetic Population

Sources of Auxiliary Data

Both stratification (Section 2.3) and determination of the true values (Section 2.4) were mainly based on the results of the multi-source NFI of the year 2019 (MS-NFI-2019; Mäkisara et al. 2022). MS-NFI produces thematic maps of forest resources (e.g., Figure 1, right) by combining NFI field plot data with optical satellite images and digital maps. Forestry land is separated from other land use classes based on the National Land Survey (NLS) of Finland's topographic database, and the values of such target variables as land use class within forestry land (Table 1) and mean volume of growing stock, m^3/ha, are predicted for 16 m × 16 m pixels classified as forestry land. A nonparametric k-nearest neighbor (k-NN) method is used in the prediction (Tomppo and Halme 2004; Tomppo et al. 2008).

Since MS-NFI had been found quite unreliable in distinguishing between poorly productive and unproductive forest land in some parts of the study region, the canopy height model of Finland (CHM) was used as supplementary data in stratification. CHM is based on airborne laser scanning (ALS) data acquired by NLS in 2008–2019 as a part of the first round of the national laser scanning program (https://www.maanmittauslaitos.fi/en/maps-and-spatial-data/expert-users/product-descriptions/laser-scanning-data). The main objective of the program was to generate an accurate digital terrain elevation model (DEM), but the project has also provided useful data for forest inventory. The flight altitude during scanning operations has been approximately 1800–2500 m, which results in a nominal laser point density of approximately 0.5 points/m^2. A DEM has been produced using the ALS points classified as ground returns by the classification routine of Axelsson (2000). The final product is an elevation map with a 2 m spatial resolution. Based on the DEM and original ALS data, CHM has been derived from the vegetation returns (laser echoes) of the ALS data by calculating the difference between the z-coordinate values of vegetation point returns and the DEM. The resulting CHM is a raster format data set with 1 m spatial resolution (https://www.metsakeskus.fi/fi/avoin-metsa-ja-luontotieto/metsatietoaineistot/metsavaratiedot; in Finnish). The ALS-based CHM is available as an open-access data set in the Suomen Metsäkeskus data server (https://metsakeskus.maps.arcgis.com/apps/MapSeries/index.html?appid=af8f213287464fb9b0639b5d8095c644). For this study, CHM was aggregated to 10 m × 10 m pixels by the third quartile value of all contributing 1 m × 1 m pixels to obtain a local estimate of average tree height.

Stratification

In NFI13, each first-phase cluster c was assigned to one of five strata based on locations $z_p \epsilon R^2$ of its eight sample plot centers (Figure 2, left). First, each z_p was shifted to the nearest crossing of MS-NFI-2019 pixel boundaries (Figure A1) and plot $p \epsilon c$ was classified as land within study region, $u_p = 1$, if any of the surrounding 2 × 2 MS-NFI-2019 pixels was classified as land and located within the study region. For each land plot, $p \in c; u_p = 1$, the MS-NFI-2019 predictions of NFI land use class (Table 1), were extracted from windows W_p of 4 × 4 pixels around the shifted location z_p as illustrated in Figure A1, and the prediction of mean height, h_p, was computed as an average of CHM height predictions over those windows of 5 × 5 pixels whose central pixel contained the original location. In small parts of the region, where CHM data were missing, MS-NFI-2019 predictions of the basal area median tree height of a stand were used.

For MS-NFI-2019 pixel i included in one of the extracted windows W_p, let,

$$l_i = \begin{cases} 0, \text{if } i \text{ is outside the study region or classified as water,} \\ 1, \text{if } i \text{ is inside the study region and classified as productive forest land,} \\ 2, \text{if } i \text{ is inside the study region and classified as poorly productive forest land,} \\ 3, \text{if } i \text{ is inside the study region and classified as unproductive forestry land,} \\ 4, \text{if } i \text{ is inside the study region and classified as other forestry land,} \\ 5, \text{otherwise : other than forestry land.} \end{cases}$$

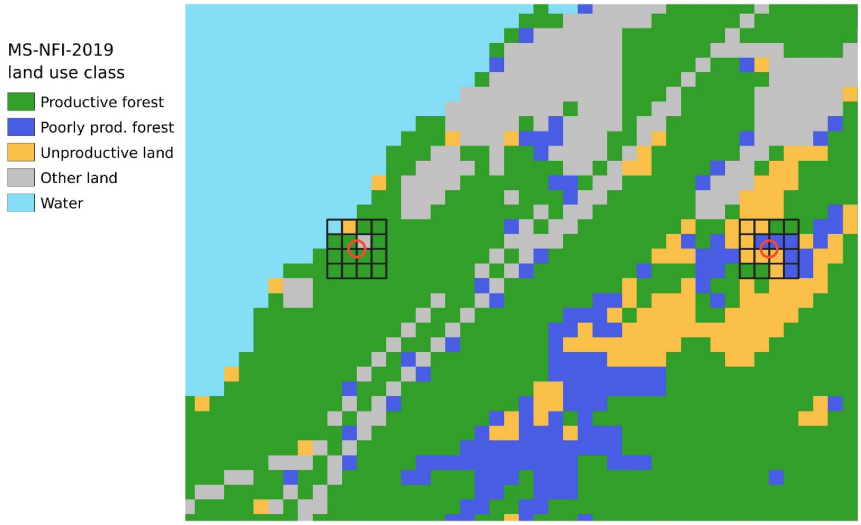

FIGURE A1 | Two (shifted) sample plot locations (red circles) of one first-phase cluster c, MS-NFI-2019 classifications of nearby 16 m × 16 m pixels, and windows of 4 × 4 pixels were used in stratification. Due to the one pixel representing other land use than forestry, this cluster would be assigned to Stratum 0. For demonstration, let us assume that both of the original plot locations were to the North-East direction from the shifted location and that cluster c contained only these two land plots within the study region, then $n_c = 2$ (number of plots classified as land) and $m_c = 1$ (number of plots on forested land = union of productive and poorly productive forest).

TABLE A1 | Stratification rules for Strata 1, 2, 3, and 45.

		Expected proportion of productive forest land, P_{c1}			
		≤ 1%[a]	> 1%, ≤ 25%	> 25%, ≤ 85%	> 85%
Expected proportion of unproductive forest land, P_{c3}	< 10%	**Stratum 3**	**Stratum 3**	**Stratum 2**	**Stratum 1**
	≥ 10%, < 40%	**Stratum 3**	**Stratum 3**	**Stratum 2**	**Stratum 2**
	≥ 40%, ≤ 85%	**Stratum 3**	**Stratum 3**	**Stratum 3**	**Stratum 3**
	> 85%	**Stratum 45**	**Stratum 3**	**Stratum 3**	**Stratum 3**

[a]No pixels classified as productive forest land.

If any of the classifications l_i, $i \in \bigcup_{p \in c} W_p$, indicated other than forestry land, $l_i = 5$, then cluster c was assigned in Stratum 0. Otherwise, the CHM-adjusted MS-NFI-2019 predictions

$$l\prime_i = \begin{cases} 3, \text{if } l_i = 2, i \in W_p, \text{and } h_p < 5\text{m} \\ l_i, \text{otherwise,} \end{cases}$$

were used to compute the expected proportions of productive forest land plots and unproductive forestry land plots among the land plots of cluster c:

$$P_{c,l} = \frac{\sum_{p \in c} \sum_{i \in W_p} 1_{\{l\}}(l_i)}{\sum_{p \in c} \sum_{i \in W_p} 1_{\{1,2,3,4\}}(l_i)}, l = 1, 3,$$

where $1_A(a)$ is an indicator function that gets value 1 when element a is included in set A and value 0 otherwise, and the cluster was then assigned in one of the other strata according to Table A1.

True Values

While stratification was based on 64 m × 64 m windows around the sample plot locations as described above, the true values for the synthetic population (Section 2.4) were directly extracted from those MS-NFI-2019 pixels that contained the original locations z_p. The area covered by one pixel, 16 m × 16 m, is approximately equal to the area covered by one NFI13 field sample plot, which includes trees up to 9 m from the center point (Korhonen et al. 2024, section 2.2.2).

Appendix B

Why Cluster-Level Stratification?

The two-phase sampling with cluster-level stratification, as applied by Finnish NFI in the study region (Appendix A), may appear overly complex. A simpler alternative would be to stratify the study region based on the pixel-level values of the MS-NFI-2019 land-use map. Figure B1 illustrates one option for doing that, assuming that we want to (i) sample areas that are expected to be more forested more densely than areas expected to be less forested, (ii) insist on cluster sampling for cost-efficiency, and (iii) use spatially systematic sampling (Section 2.5) in the location of field-plot clusters to achieve a spatially balanced sample.

The idea in this alternative approach would be that one systematic (one-phase) sample would be selected including plots that are expected to be forested and another, sparser one including plots that are expected to be unforested. This could be done by placing one randomly positioned square grid of cluster reference points over the study region and including in the resulting sample of clusters only those plots that are classified as forested and then placing another, sparser grid, positioned randomly and independently of the first grid and including only plots that are classified as unforested. This would result in a stratified sample fulfilling the requirements (i)–(iii). However, since the strata tend to be highly fragmented, this would lead to a large proportion of "incomplete" clusters, as illustrated in Figure B1, and it would be difficult to arrange one cluster to be 1 day's work. Thereby the advantage of clustering would have been largely lost.

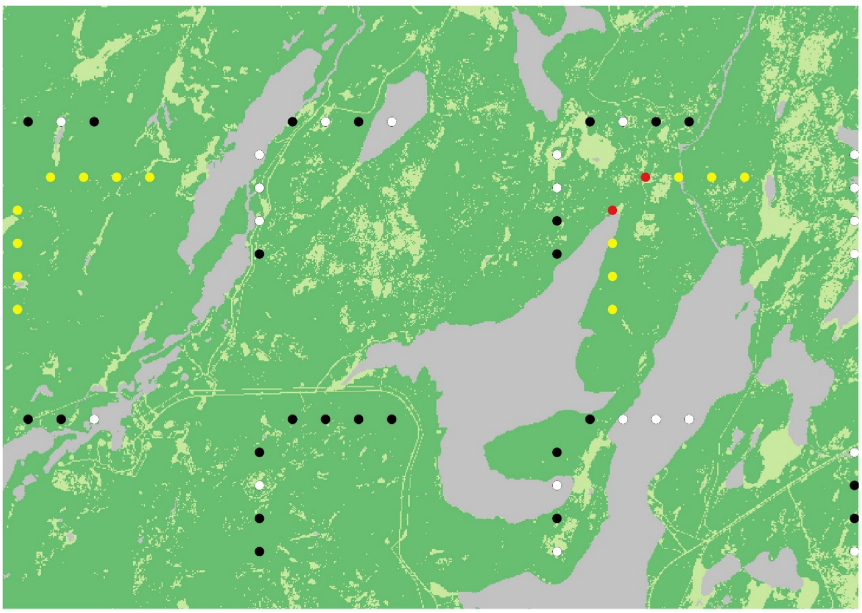

FIGURE B1 | Aggregated land use class predictions from MS-NFI-2019 (dark green: Forested land, light green: Other lands; gray: Water) and an example of a systematic cluster sample based on plot-level stratification: In order to yield twice as high inclusion probability of forested than other land plots, plots classified as forested land (black points) would be measured from the denser square grid of clusters (of which eight clusters are intersecting the figure region), and plots classified as other lands (red points) would be measured from the sparser grid (two clusters intersecting the Figure region).

Appendix C

Allocation of Simulated Samples Into the Strata

In conventional stratified sampling, a random sample of fixed-size n_h is independently selected from each stratum h of the target population. The design-unbiased estimator of the population total of survey target variable y is

$$\hat{Y} = \sum_h N_h \bar{y}_h, \tag{C1}$$

where N_h is the size of stratum h and $\bar{y}_h$ is the mean in the sample from that stratum. Given the stratum-specific population variances, S_h^2, $h = 1, \ldots, H$, of y, Neyman (1934) derived a rule for optimal allocation of a fixed total sample size n between the strata: With the condition that the stratum-specific sample sizes n_h must satisfy the restriction $\sum_h n_h = n$, sampling variance $\hat{Y}$ is minimized when

$$\frac{n_h}{n} = \frac{N_h S_h}{\sum_{h'} N_{h'} S_{h'}} \tag{C2}$$

This result relies on the formula

$$\mathrm{Var}(\bar{y}_h) = \frac{S_h^2}{n_h}. \tag{C3}$$

For our application, we developed a modification of (Equation C2) to the case, where stratum-specific estimators are ratio estimators and the target is to derive optimal stratum-specific inclusion probabilities $\pi_h = E(n_h)/N_h$ for systematic sampling (with random sample sizes; cf. Section 2.5) so that

$$\sum_h \pi_h N_h = n^*. \tag{C4}$$

In our stratified estimator of total volume (Equations 2 and 4 in the main text)

$$\hat{V}_{F,s} = \sum_h \hat{V}_{F,s,h} = \sum_h A_h \bar{v}_h, \tag{C5}$$

the estimators

$$\bar{v}_h = \frac{\sum_{c \in s_h} V_c}{\sum_{c \in s_h} n_c} \tag{C6}$$

of stratum-specific mean volumes

$$\bar{V}_h = \frac{\sum_{c \in U_h} V_c}{\sum_{c \in U_h} n_c} \tag{C7}$$

are ratios of two random variables with approximate variances (Särndal et al. 2003 section 5.6)

$$\mathrm{Var}(\bar{v}_h) \approx \frac{S_{\mathrm{R},h}^2}{n_h}, \tag{C8}$$

where

$$S_{\mathrm{R},h}^2 = \frac{\frac{1}{N_h}\sum_{c \in U_h}\left(V_c - \bar{V}_h n_c\right)^2}{\left(\frac{1}{N_h}\sum_{c \in U_h} n_c\right)^2}, \tag{C9}$$

and N_h is the total number of clusters in stratum h. The modification of (Equation C2) to our case is then

$$\frac{E(n_h)}{n^*} = \frac{A_h S_{\mathrm{R},h}}{\sum_{h'} A_{h'} S_{\mathrm{R},h'}}, \tag{C10}$$

leading to

$$\pi_h = \frac{n^*}{N_h} \frac{A_h S_{\mathrm{R},h}}{\sum_{h'} A_{h'} S_{\mathrm{R},h'}}. \tag{C11}$$

As explained in Section 2.5, this rule was used to determine the target inclusion probabilities π_h for stratified SYS sampling and to select the sampling intervals k_h accordingly. Since we were sampling from a finite

population, the target $E(n_h) = k_h^{-2} N_h = \pi_h N_h$ could be achieved only approximately. For that reason, we did not choose the (fixed) sample sizes of the other methods directly with the Neyman rule but tried to make them as comparable as possible to SYS, by setting their sample sizes equal to the median number of clusters over all possible SYS samples.

Appendix D

On the Role and Impact of First-Phase Sampling

Eight further replications of our synthetic population (Section 2.4), 750 m grids, were available as the other similar sub-grids of the NFI13 first-phase grid (250 m; Section 2.3). They were utilized to assess the impact of first-phase sampling through ratios

$$\frac{\widehat{\mathrm{Var}}\left(\hat{\theta}_2 - \hat{\theta}_1\right)}{\widehat{\mathrm{Var}}\left(\hat{\theta}_2\right)}, \tag{D1}$$

where $\widehat{\mathrm{Var}}\left(\hat{\theta}_2 - \hat{\theta}_1\right)$ is the variance of the second-phase estimates around the estimate from the first-phase sample from which they were drawn—that is, replications of what we reported in the main text—and $\widehat{\mathrm{Var}}\left(\hat{\theta}_2\right)$ is the overall variance of the second-phase estimates including the variability between first-phase samples. Both of the variances in (Equation D1) were estimated from the combination of simulated second-phase samples from all sub-grids. We also studied the impact of decreasing the density of the first-phase grid by using all 36 grids with 1500 m spacing and all 81 grids with 2250 m spacing as "first-phase samples."

Standard deviations of stratum weights A_h between the first-phase samples were all less than 1.5% with 750 m spacing (Table D1). With 1500 m spacing, they approximately doubled, and with 2250 m spacing, tripled. In each case, their share of the total sampling variance was less than 2% (Table D2).

In applications like ours, where the values for the first-phase sample are extracted from existing high-resolution images, a very large sample can be used. As long as the sample size is large enough, different first-phase samples lead to practically identical results (as demonstrated here), so that the impact of first-phase sampling is negligible. In our case, it was simply a technique to convert the continuous population of all points on the land area of the study region into a finite population of clusters. Further discussion of this technique can be found in Grafström et al. (2017).

TABLE D2 | Ratios of variance due to second-phase sampling to total sampling variance (Equation 3) in estimates computed from 1000 simulated systematic two-phase samples using first-phase samples with different densities.

1st phase grid Spacing, m	Forested area		Total volume	
	Stratified	Unstratified	Stratified	Unstratified
750	0.9996	0.9978	0.9999	0.9994
1500	0.9947	0.9944	0.9891	0.9966
2250	0.9899	0.9836	0.9821	0.9923

Appendix E

Statistical Significance of Differences Between Sampling Designs

When $\mathrm{Var}_d(\hat{\theta})$ is estimated through the simulation of random samples (Section 2.6) and the resulting estimates $\hat{\theta}_t$, $t = 1, \ldots, T$, there is a Monte Carlo error associated with it. Considering that $\widehat{\mathrm{Var}}_d(\hat{\theta})$ is the mean of T random variables $\delta_t = \left(\hat{\theta}_t - \hat{\mathrm{E}}_d\hat{\theta}\right)^2$, where $\hat{\mathrm{E}}_d\hat{\theta}$ is the mean of $\hat{\theta}_t$'s, the Monte Carlo variance of $\widehat{\mathrm{Var}}_d(\hat{\theta})$ was estimated by $\mathrm{Var}_{\mathrm{MC}}\left[\widehat{\mathrm{Var}}_d(\hat{\theta})\right] = \frac{\sigma^2(\delta_t)}{T}$, where $\sigma^2(\delta_t)$ is the empirical variance of δ_t over the simulated samples. The 95% Monte Carlo confidence interval for $\widehat{\mathrm{Var}}_d(\hat{\theta})$ were constructed as $\widehat{\mathrm{Var}}_d(\hat{\theta}) \pm u\left[\widehat{\mathrm{Var}}_d(\hat{\theta})\right]$, where

$$u\left[\widehat{\mathrm{Var}}_d(\hat{\theta})\right] = 1.96 \times \left\{\mathrm{Var}_{\mathrm{MC}}\left[\widehat{\mathrm{Var}}_d(\hat{\theta})\right]\right\}^{\frac{1}{2}},$$

TABLE D1 | Means and standard deviations (SD) of the estimates of land area represented by stratum h (km^2), A_h, forested area (km^2), $A_{F,h}$, and total volume (1000 m^3), $V_{F,h}$, from the first-phase samples with different densities (grid spacing).

		A_h		$A_{F,h}$		$V_{F,h}$	
h	Grid, m	Mean	SD	Mean	SD	Mean	SD
0	750	2551.7734	27.3718	1789.1839	19.2609	9953.5671	107.6693
	1500	25517878	62.9247	1789.1949	45.4795	9953.6303	255.2889
	2250	2551.7762	81.0407	1789.1864	60.4966	9953.5827	358.2412
1	750	1975.3044	21.8758	1938.0213	22.3642	14,930.8113	178,5696
	1500	1975.3082	42.5993	1938.0242	42.3916	14,930.8351	322.5735
	2250	1975.3058	70.0659	1938.0229	69.1416	14,930.8394	522.8026
2	750	6193.2946	32.6858	5211.0541	27.4108	31,899.9495	182.5491
	1500	6193.2919	653504	5211,0502	53,7325	31,899,9541	329,0130
	2250	6193.3018	102.7444	5211.0592	86.2404	31,899.9302	551.1941
3	750	8834.3900	42.0737	4038.3122	25.1096	13,514.3625	80.4899
	1500	8834.3700	80.7013	4038.3018	44.8766	13,514.3316	180.2160
	2250	8834.3785	121.7381	4038.3084	66.0076	13,514.3214	285.3939
45	750	8603.1076	31.3897	1339.0678	11.1353	1664.9115	24.0589
	1500	8603.1121	59.4978	1339.0663	24.5764	1664.9065	40.8436
	2250	8603.1077	90.1519	1339.0691	34.7693	1664.9204	50.7331
All	750	28,157.8700		14,315.6393	18.5171	71,963.6020	84.5208
	1500	28,157.8700		14,315.6374	33.5687	71,963.6576	215.6484
	2250	28,157.8700		14,315.6460	57.2253	71,963.5942	321.9401

and that for $\mathrm{rsd}_d(\widehat{\theta}) = 100 \times \left[\widehat{\mathrm{Var}}_d(\widehat{\theta})\right]^{1/2}/\theta$ (Equation 9 in the main text) as

$$\mathrm{rsd}_d(\widehat{\theta}) \pm 100 \times \frac{u\left[\widehat{\mathrm{Var}}_d(\widehat{\theta})\right]^{\frac{1}{2}}}{\theta}.$$

Only those DEFF's were reported in the results text, for which the corresponding intervals did not overlap or had minor overlap in comparison to their length.

Similarly, $\mathrm{GB}(d,U)^2$ (Equation 11 in the main text) is the mean of random variables $d_t = \left(\bar{i}_t - \bar{I}\right)^2 + \left(\bar{j}_t - \bar{J}\right)^2$, its Monte Carlo variance was estimated by $\mathrm{Var}_{\mathrm{MC}}\left[\mathrm{GB}(d,U)^2\right] = \frac{\sigma^2(d_t)}{T}$, where $\sigma^2(d_t)$ is the empirical variance of d_t over the simulated samples, and its Monte Carlo confidence interval was constructed as $\mathrm{GB}(d,U)^2 \pm u\left[\mathrm{GB}(d,U)^2\right]$, where

$$u\left[\mathrm{GB}(d,U)^2\right] = 1.96 \times \left\{\mathrm{Var}_{\mathrm{MC}}\left[\mathrm{GB}(d,U)^2\right]\right\}^{1/2}.$$

Finally, the confidence interval of $\mathrm{GB}(d,U)$ was constructed as

$$\mathrm{GB}(d,U)^2 \pm \left\{u\left[\mathrm{GB}(d,U)^2\right]\right\}^{1/2}.$$

Spatial regularity (Equation 10) is different since it is a sample-specific rather than a design-specific measure. Thus, we have T simulated values $\mathrm{SR}(s_t,U)$ from each design. Statistical significance in differences of median values of $\mathrm{SR}(s_t,U)$ between the designs was assessed using the Mann–Whitney U test (for details, see the R code for Table 6 included in the Data S1). Exhaustive results are not reported here, but all differences in median $\mathrm{SR}(s,U)$ between SYS and LPM (Table 6) are highly significant.

Appendix F

Estimation of Sampling Uncertainty From the Field Sample

For the Horvitz–Thompson estimator

$$\widehat{Y} = \frac{N}{n}\sum_{c\in s} y_c$$

of the population total $Y = \sum_{c=1}^{N} y_c$ based on a spatially balanced equal probability sample s of n clusters c, the variance estimator proposed by Grafström and Schelin (2014) is

$$\widehat{\mathrm{Var}}_{\mathrm{SB}}(\widehat{Y}) = \left(\frac{N}{n}\right)^2 \sum_{c\in s} \frac{n_c^*}{n_c^* - 1}\left(y_c - \frac{1}{n_c^*}\sum_{c'\in s_c^*} y_{c'}\right)^2, \tag{F1}$$

where the subset s_c^* of n_c^* clusters include cluster c and its nearest neighboring clusters. For this study, we needed a slight modification, since our stratum-specific estimators (Equation 1) are ratios of two random quantities, that is, have the general form

$$\widehat{\theta} = \frac{\sum_{c\in s} y_c}{\sum_{c\in s} x_c},$$

where both y_c and x_c are random. The conventional variance estimator for such a ratio is (e.g., Cochran 1977, section 2.11)

$$\widehat{\mathrm{Var}}_{\mathrm{SRS}}(\widehat{\theta}) = \frac{\sum_{c\in s} z_c^2}{\left(\sum_{c\in s} x_c\right)^2},$$

where $z_c = y_c - \widehat{\theta}x_c$, and the analogous modification of the estimator (Equation F1) is

$$\widehat{\mathrm{Var}}_{\mathrm{SB}}(\widehat{\theta}) = \frac{1}{\left(\sum_{c\in s} x_c\right)^2} \sum_{c\in s} \frac{n_c^*}{n_c^* - 1}\left(z_c - \frac{1}{n_c^*}\sum_{c'\in s_c^*} z_{c'}\right)^2. \tag{F2}$$

Variances of stratum-specific estimators from NFI13 field data (section 2.7) were estimated using Equation (F2). The neighborhoods s_c^* consisted of cluster c and its Voronoi-neighbors (Figure F1). Variances of the stratified estimators of population totals were estimated as sums of stratum-specific estimators.

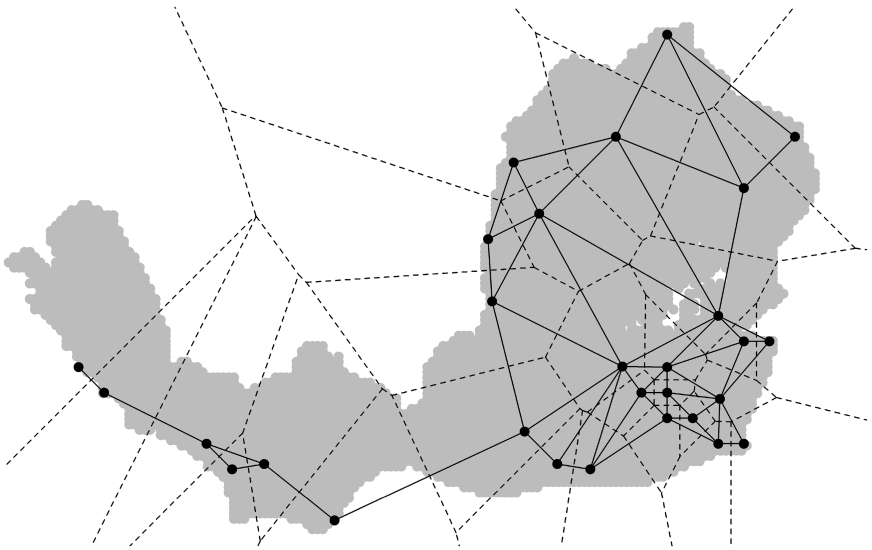

FIGURE F1 | Locations of NFI13 field clusters belonging to stratum 0 (dots), their Voronoi tiles (dashed lines), and Voronoi-neighborhood graph (solid lines).