Spatial Statistics 64 (2024) 100865

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

Exploiting nearest-neighbour maps for estimating the variance of sample mean in equal-probability systematic sampling of spatial populations

Sara Franceschi [a],[*], Lorenzo Fattorini [a], Timothy G Gregoire [b]

[a] University of Siena, Department of Economics and Statistics, P.zza San Francesco 8, 53100, Siena, Italy
[b] Yale School of the Environment, 195 Prospect Street, New Haven, CT 06511, USA

ARTICLE INFO

Keywords:
Spatial interpolation
Pseudo-populations
Design-based consistency
Simulation study

ABSTRACT

Because of its ease of implementation, equal probability systematic sampling is of wide use in spatial surveys with sample mean that constitutes an unbiased estimator of population mean. A serious drawback, however, is that no unbiased estimator of the variance of the sample mean is available. As the search for an omnibus variance estimator able to provide reliable results under any spatial population has been lacking, we propose a design-consistent estimator that invariably converges to the true variance as the population and sample size increase. The proposal is based on the nearest-neighbour maps that are taken as pseudo-populations from which all the possible systematic samples can be enumerated. As nearest-neighbour maps are design-consistent under equal-probability systematic sampling and mild conditions, the variance of the sample mean achieved from all the possible systematic samples selected from the map is also a consistent estimator of the true variance. Through a simulation study based on artificial and real populations we show that our proposal generally outperforms the familiar estimators proposed in literature.

1. Introduction

When sampling spatial units, the selection of spatially balanced samples, i.e., samples in which units are well spread throughout the study region, is suitable. Spatially balanced samples are usually constructed by avoiding or reducing the selection of contiguous units. That can be done by a plethora of spatial schemes explicitly constructed for this purpose, that continue to appear in statistical journals. Di Biase et al. (2024, Section 4) provide an updated review of these schemes, pointing out their complex nature, and comparing them with very simple schemes such as tessellation stratified sampling and systematic sampling (Breidt 1995) that easily provide spatially balanced samples without losing precision with respect to more complicated schemes.

In particular, when spatial populations are constituted by grids of regular polygons and can be partitioned into regular blocks of contiguous polygons, spatial balance is straightforwardly ensured by equal-probability systematic sampling (EPSS), in which one polygon is randomly selected in one block and then repeated in the other blocks. Under EPSS, the population mean is simply estimated by the sample mean that in this case coincides with the Horvitz-Thompson (HT) estimator (see Section 2).

Owing to its simplicity, EPSS has constituted a sort of standard design of wide application especially in forest surveys (Tomppo et al. 2010), even if the inability of estimating the variance of the sample mean unbiasedly and precisely is a well-known drawback. The

[*] Corresponding author.
E-mail address: sara.franceschi@unisi.it (S. Franceschi).

https://doi.org/10.1016/j.spasta.2024.100865
Received 27 May 2024; Received in revised form 21 October 2024; Accepted 23 October 2024
Available online 24 October 2024

problem is a long-standing issue in both the statistical and forestry literature (Langsæter 1926; Hasel 1938; Osborne 1942) and a huge sequence of estimators have been proposed, but none of them uniformly outperform the others, with performance depending on the population characteristics (see Section 3).

As an alternative to search for an improbable omnibus estimator, in this paper we propose the construction of a design consistent variance estimator based on population maps achieved by the nearest-neighbour (NN) interpolation of sampled values (see Section 4). The idea arose from a paper by Opsomer et al. (2007) devoted to model-assisted estimators of forest attributes under the two-phase systematic sampling adopted in U.S by the Department of Agriculture Forest Service. To check the precision of these estimators avoiding the use of standard but potentially unreliable variance estimators, the authors construct a so-called synthetic population able to mimic the one from which they are sampling, draw all possible systematic samples from that population, and take the variance of the resulting estimates across these samples as the estimator of the true variance.

Notwithstanding the originality of the proposal, it is unclear in what sense synthetic populations would be able to mimic true populations. In design-based inference, populations suitable to mimic true populations are referred to as pseudo-populations and they are usually adopted in the bootstrap estimation of sampling variances by selecting independent samples from them by replicating the same sampling scheme adopted to select the original sample (Quatember 2016). In spatial sampling, Franceschi et al. (2022) investigate the role of pseudo-populations for mimicking the characteristics of real spatial populations such as spatial trend and relationships and similarities among neighbouring locations. The results indicate the superiority of pseudo-populations achieved by the NN interpolation of the sampled values with respect to other pseudo-populations adopted in literature (Conti et al. 2020). The reason is that under suitable sampling schemes including EPSS, the NN interpolator is design-consistent when the number of polygons partitioning the study region increases and their size decreases (Fattorini et al. 2022).

Obviously in the case of EPSS there is no need of bootstrap resampling, as the number of possible samples in EPSS is usually so small that samples can be completely enumerated. In practice, because NN pseudo-populations converge to true populations for sufficiently large populations and sample size, the variance of the sample mean across all the possible systematic samples converges to the true variance and as such should provide an effective estimator.

In Section 5 the performance of the proposed estimator is assessed and compared with those of traditional estimators considered in Section 3. The simulation study is performed on four surfaces, two of them derived from real studies and two of artificial nature. To check the design-based consistency, these surfaces are partitioned into populations of quadrats of decreasing size from which samples of increasing sizes are selected by EPSS. Concluding remarks are given in Section 6.

2. Preliminaries and notations

Let A be a study area partitioned into a finite population U of N regular polygons and let y_j be the amount of an interest variable Y for each $j \in U$. Suppose $N = Kn$, so that U can be split into n equally-shaped blocks of K adjacent polygons $U_1,\dots,U_n$. The spatial EPSS consists in randomly selecting one polygon in the first block and then selecting the polygons having the same position in the other blocks (Fig. 1). In this way, the EPSS scheme partitions U into K possible samples $s_1,\dots,s_K$ of size n, each of them selected with probability $1/K = n/N$. Accordingly, for each $j \in U$, the first-order inclusion probability is invariably equal to $\pi_j = n/N$, and for each $h \neq j \in U$, the second-order inclusion probability is equal to $\pi_{j,h} = n/N$ if j and h are in the same sample s_k and is equal to 0 otherwise.

Under EPSS, the HT estimator of the population mean

$$\overline{Y} = \frac{1}{N}\sum_{j\in U} y_j$$

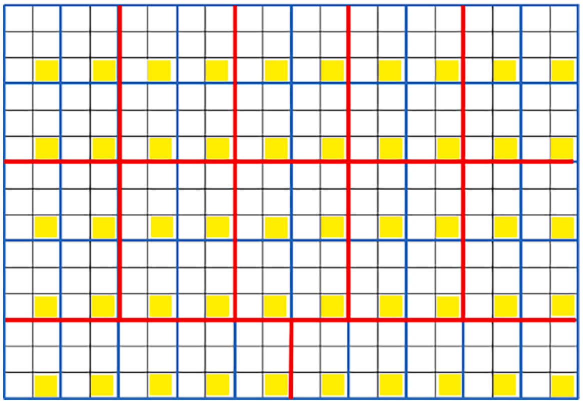

Fig. 1. Example of a population of $N = 300$ quadrats partitioned into $n = 50$ blocks of $K = 3x2 = 6$ contiguous quadrats. Blocks are identified by blue lines. One quadrat is randomly selected in the first block and systematically repeated in the others. Selected quadrats are highlighted in yellow. To treat the systematic sample as a stratified sample, 10 strata are constructed by joining $2x2 = 4$ adjacent blocks and 2 strata are constructed by joining 5 adjacent blocks in the lower part of the study region. Strata are identified by red lines.

coincides with the sample mean

$$\bar{y} = \frac{1}{n}\sum_{j\in S} y_j \tag{1}$$

where S denotes the sample randomly selected from the K possible samples. The HT estimator (1) is design unbiased with design-based variance

$$V_{SYS}(\bar{y}) = \frac{1}{K}\sum_{k=1}^{K}\left(\bar{y}_k - \overline{Y}\right)^2 \tag{2}$$

where $\bar{y}_k$ denotes the mean of the sample s_k.

Because the second-order inclusion probabilities are zero for any two polygons that do not belong to the same sample, the design is non-measurable, i.e. no unbiased estimator of (2) exists. The problem of estimating the variance of the sample mean in EPSS and in particular in EPSS of spatial populations is a long-standing and long-debated issue of sampling theory.

3. A review of variance estimation in equal-probability systematic spatial sampling

The standard approach for estimating the variance of sample mean in EPSS of both one-dimensional and spatial populations is obtained by assuming that the sample has been selected by simple random sampling. If simple random sampling is supposed to be with replacement (SRSWR), the variance estimator reduces to

$$\widehat{V}_{SRSWR} = \frac{v^2}{n} \tag{3}$$

where

$$v^2 = \frac{1}{n-1}\sum_{j\in S}\left(y_j - \bar{y}\right)^2$$

is the variance of the selected sample. The estimator (3) is also a member - achieved when $\pi_j = n/N$ - of a large class of variance estimators of type

$$\widehat{V} = \frac{1}{n(n-1)}\sum_{j\in S}\left(\widehat{\bar{y}}_j - \widehat{\overline{Y}}_{HT}\right)^2 \tag{4}$$

where each $\widehat{\bar{y}}_j = \left(ny_j\right)/\left(N\pi_j\right)$ is an unbiased estimator of the population mean and their average coincides with the HT estimator of the population mean. These estimators are applicable to any without replacement design as straightforward variance estimators that avoid the use of second-order inclusion probabilities. Though biased, these estimators have the appealing property to be invariably nonnegative and are traditionally considered as conservative. In his classic (reissued) textbook on variance estimation, Wolter (2007, section 2.4.5) has attempted to justify the conservative nature of variance estimators of type (4) under any single-stage without-replacement sampling design, but with no definitive result.

If simple random sampling is supposed to be without replacement (SRSWOR), the variance estimator is

$$\widehat{V}_{SRSWOR} = \frac{N-n}{N}\frac{v^2}{n} \tag{5}$$

which is invariably smaller than the SRSWR estimator (3). Opinions on these approaches are many and varied. Milne (1959), Cochran (1977), Ripley (1981) suggest that they are safe approaches both recommendable and acceptable in most cases, giving “a good idea of the true sampling variance” (Ripley 1981, p.27). On the other hand, other authors are more critical (e.g. Payandeh 1970, McRoberts et al. 2016). Actually, what can be said about the estimators (3) and (5) is that they are generally biased, but the magnitude and direction of the bias is obscure. In some cases, it can be shown that their bias tends to be positive, but this is not the case universally, especially when EPSS does not perform well because blocks do not provide an effective stratification of the population, i.e. when blocks fail to partition the population into homogeneous sets that greatly differ between them (Aronow and Samii 2013). The issue is also apparent from the plethora of simulation studies performed on one-dimensional populations (e.g. Wolter 1984) and spatial populations (e.g. Dunn and Harrison 1993, D’Orazio 2003, Strand 2017) where no definitive result about the conservative nature of (3) and (5) is reached.

A compelling result is achieved by Aronow and Samii (2013). The authors prove that the standard HT variance estimator

$$\widehat{V}_{HT} = \frac{1}{N^2}\left(\sum_{j\in S}\frac{1-\pi_j}{\pi_j^2}y_j^2 + 2\sum_{h>j\in S}\frac{\pi_{jh}-\pi_j\pi_h}{\pi_j\pi_h\pi_{jh}}y_j y_h\right)$$

that is design unbiased in presence of measurable design, turns out to be invariably conservative under non measurable designs when $y_j \geq 0$ for each $j \in U$. In the case of EPSS, inserting the expressions of the first and second order inclusion probabilities in the previous equation, the standard HT variance estimator reduces to

$$\widehat{V}_{HT} = \frac{N-n}{N}\overline{y}^2 \tag{6}$$

From Eq. (6) it is apparent the weak dependence of the HT estimator on the sample size n. This fact is likely to generate very large overestimations. Clearly, this is the price to paid for ensuring the conservative property of the estimator in presence of any population with non-negative y_js.

Besides these naïve variance estimators, an alternative method of long standing in sampling literature is to treat the systematic sample as if it was a stratified random sample. Yates (1981, section 7.18) suggests constructing artificial non-overlapping strata by joining 4 adjacent blocks in a 2 × 2 pattern, in such a way that spatial variability is considered in both directions with 4 sample observations per strata. Some strata constituted by a different number of blocks are allowed when n is not a multiple of 4 or a 2 × 2 collapse is not possible (Fig. 1). In analogy with Strand (2017), denote by L the number of strata artificially constructed by joining adjacent blocks, by N_l the number of polygons within the stratum l, by S_l the set of polygons in the systematic sample S belonging to the stratum l and by n_l the size of S_l with $n_l \geq 2$ to allow variance estimation within strata. Then, from the standard formula of stratified sampling the variance estimator is

$$\widehat{V}_{STR} = \sum_{l=1}^{L} w_l^2 \frac{N_l - n_l}{N_l}\frac{v_l^2}{n_l} \tag{7}$$

where $w_l = N_l/N$ is the l-stratum weight and

$$v_l^2 = \frac{1}{n_l - 1}\sum_{j\in S_l}\left(y_j - \overline{y}_l\right)^2$$

is the estimator of variance within the stratum l, with

$$\overline{y}_l = \frac{1}{n_l}\sum_{j\in S_l} y_j$$

If n is a multiple of 4 and the stratification achieved joining 2 × 2 adjacent blocks can be exactly performed, then $L = n/4$, $N_l = 4K$, $n_l = 4$, $w_l = 4/n$ for each stratum l and the estimator (7) reduces to

$$\widehat{V}_{STR4} = \frac{N-n}{N}\frac{4}{n^2}\sum_{l=1}^{n/4} v_l^2 \tag{8}$$

In the case of one-dimensional populations, Wolter (2007, section 8.2) suggests two corrections of $\widehat{V}_{SRSWOR}$ based on the serial correlation between sample observations. In particular, the first correction is achieved by multiplying $\widehat{V}_{SRSWOR}$ by half of the Durbin-Watson statistics. D'Orazio (2003) propose extending this correction to spatial populations multiplying $\widehat{V}_{SRSWOR}$ by the Geary's spatial autocorrelation index (Cliff and Ord 1973, p.8). The resulting estimator is

$$\widehat{V}_{GEARY} = \frac{N-n}{2DNn}\sum_{j\in S}\sum_{h\neq j\in S}\delta_{j,h}\left(y_j - y_h\right)^2 \tag{9}$$

where

$$D = \sum_{j\in S}\sum_{h\neq j\in S}\delta_{j,h}$$

and $\delta_{j,h} = 1$ if the sampled polygons j and h belong to adjacent block having at least a side or a vertex in common, and $\delta_{j,h} = 0$ otherwise. The second correction, originally proposed by Cochran (1946) for one-dimensional populations, has been extended to spatial population by D'Orazio (2003) - and subsequently by McGarvey et al. (2016) and Strand (2017) - substituting the serial correlation index by the Moran's spatial autocorrelation statistic (Cliff and Ord 1973, p.8). The resulting estimator is

$$\widehat{V}_{MORAN} = \begin{cases} \dfrac{N-n}{N}\dfrac{v^2}{n}\left[1 + \dfrac{2}{\ln I} + \dfrac{2}{\left(I^{-1}-1\right)}\right] & \text{if } I > 0 \\ \dfrac{N-n}{N}\dfrac{v^2}{n} & \text{otherwise} \end{cases} \tag{10}$$

where

$$I = \frac{n}{(n-1)v^2 D} \sum_{j \in S} \sum_{h \neq j \in S} \delta_{j,h} \left(y_j - \overline{y}\right)(y_h - \overline{y})$$

is the Moran's statistics and the $\delta_{j,h}$s are as in Eq. (9).

In his comparison of variance estimators of sample mean under spatial EPSS - the most updated at least to our knowledge - Strand (2017) considers the model-based approach by Brus and Saby (2016) in which the semi-variance concept is exploited to rewrite the variance of the sample mean, also providing an estimator for it. We avoid considering this method because it necessitates estimating the semi-variogram also for lags much smaller than the distances among sample observations.

4. A variance estimator based on nearest-neighbour spatial interpolation

Denote by $a_1, \ldots, a_N$ the N regular polygons of size b partitioning A with centroids $c_1, \ldots, c_N$, and suppose the existence of a Riemann integrable function $d(p)$ from A onto $[0, L]$ giving the density of Y at any location $p \in A$, so that the amount of Y within a_j is given by $y_j = \int_{a_j} d(p)\lambda(dp)$ while $d_j = y_j / b$ is the density in the j-th polygon.

Even if an assumption about the population values may appear unsuitable in a design-based framework, where no assumption is necessary, Fattorini et al. (2018) point out that the Riemann integrability assumption seems adequate in most environmental investigations. In real world, the density of a variable, even if high, never attains infinity. In real world, there are parts of the study region in which density changes smoothly throughout space, well approaching the theoretical condition of continuity. Even when density changes abruptly, that usually occurs along borders delineating sudden variations in the characteristic of the study region. Thus, these borders may be realistically approximated by curves on the study area, well approaching the theoretical condition of discontinuity over a region of zero measure. This situation is typical in forest mapping (forest/nonforest) where $d(p)$ is 1 within forest areas and 0 otherwise, thus jumping from 0 to 1 along the forest edges, which may be realistically considered regions of zero measure.

The estimation of y_j for each polygon $j \in U$ is essential in many environmental surveys as it enables the mapping of the spatial distribution of Y. Mapping of finite populations of polygons is usually approached by model-dependent techniques, where the y_js are assumed as outcomes of spatial autoregressive models and connections in the polygon lattice define neighbours used to model spatial dependency (e.g. Cressie 1993, chapter 6). More recently, mapping has been performed in a design-based framework adopting a class of inverse distance weighting (IDW) interpolators (Fattorini et al. 2018) or the NN interpolator (Fattorini et al. 2022). In this study we focus on the use of NN interpolator that, besides its large use in environmental investigations (e.g., Li and Heap 2008), it avoids the choice of the smoothing parameter that is necessary when using IDW interpolators. Because the polygon size b is always known, for finite sample it is theoretically equivalent to interpolate the y_js or the d_js. However densities are more suitable for working in an asymptotic scenario in which the polygon size approaches zero so that the y_js also approach zero.

Quoting from Fattorini et al. (2022), the NN interpolator of d_j is given by

$$\widehat{d}_j = I_j d_j + \frac{1 - I_j}{card(H_j)} \sum_{i \in H_j} d_i \tag{11}$$

where I_j is the sample indicator variable equal to 1 if $j \in S$ and equal to 0 otherwise, and $H_j = \left\{ i : \ \| c_i - c_j \| = min_{h \neq j \in S} \| c_h - c_j \| \right\}$ is the set of sampled polygons nearest to polygon j. For finite samples, no properties are available because (11) has intractable design-based expectation and variance. However, some appealing results can be achieved in a design-based asymptotic scenario.

To this purpose, suppose a sequence $\{U_k\}$ of partitions of A in which each partition U_k is constituted of N_k regular polygons of size b_k with $N_k \rightarrow \infty$ and $b_k \rightarrow 0$. In practice, A is partitioned into an increasing number of regular polygons of decreasing size, so that it is possible to suppose a sequence of designs selecting samples of increasing size $n_k \rightarrow \infty$. In this framework, under an EPSS that selects a constant fraction p of polygons from each U_k, Fattorini et al. (2022) prove the uniform design-based consistency of the NN interpolator (11), i.e.

$$\operatorname*{plim}_{k \to \infty} \max_{j \in U_k} \left| \widehat{d}_{j(k)} - d_{j(k)} \right| = 0$$

where $d_{j(k)}$ is the density in the j-th polygon of the k-th population and $\widehat{d}_{j(k)}$ is the corresponding NN interpolator of type (11). In practice, for population and sample size sufficiently large, under EPSS the map achieved using (11) approaches the true map. Moreover, since the interpolator of type (11) is bounded by L, the uniform design-based consistency ensures the design-based asymptotic unbiasedness.

Besides its practical importance, population mapping has also a theoretical relevance as the resulting map $\left\{ \widehat{y}_j, \ j \in U \right\}$ - where $\widehat{y}_j = b\widehat{d}_j$ are the interpolated values achieved by rescaling the interpolated densities - provides a pseudo-population from which the variance of EPSS can be simply estimated by mimicking Eq. (2). The resulting estimator turns out to be

$$\widehat{V}_{NN} = \frac{b^2}{K} \sum_{k=1}^{K} (\widehat{\overline{d}}_k - \widehat{\overline{D}})^2 \tag{12}$$

where

$$\widehat{\bar{d}}_k = \frac{1}{n}\sum_{j \in s_k} \widehat{d}_j$$

is the mean of the interpolated densities within the possible sample s_k, and

$$\widehat{\bar{D}} = \frac{1}{N}\sum_{j \in U} \widehat{d}_j$$

is the mean of the interpolated densities in the whole population. Because for population and sample size sufficiently large the $\widehat{d}_j$s converge to d_js in probability, the ratio $\widehat{V}_{NN}/V_{SYS}(\bar{y})$ does not depend on b and approaches 1 in probability. In this sense $\widehat{V}_{NN}$ is a consistent estimator of $V_{SYS}(\bar{y})$.

5. Simulation study

A 27 ha section of Harvard Forest (Petersham, Massachusetts, USA) is considered. Data are available for download at https://harvardforest1.fas.harvard.edu/exist/apps/datasets/showData.html?id=hf253. Spatial coordinates and above ground biomass (AGB) of the living trees censused in 2014 within the 27 ha portion are considered, the density of AGB per ha is derived for each 30 m square area partitioning the portion, and an artificial continuous surface is achieved by means of ordinary kriging interpolation across a grid of 540 × 720 points. In addition, we have considered two artificial surfaces over a study area of 27 ha to generate values of the density of AGB per ha. For each point $\boldsymbol{p} = [p_1, p_2]$, the function

$$f(\mathbf{p}) = 1 + p_1 + p_2 \tag{13}$$

has produced a surface with a linear spatial trend, while the function

$$f(\mathbf{p}) = 3p_1^2 + 6p_2^2 \tag{14}$$

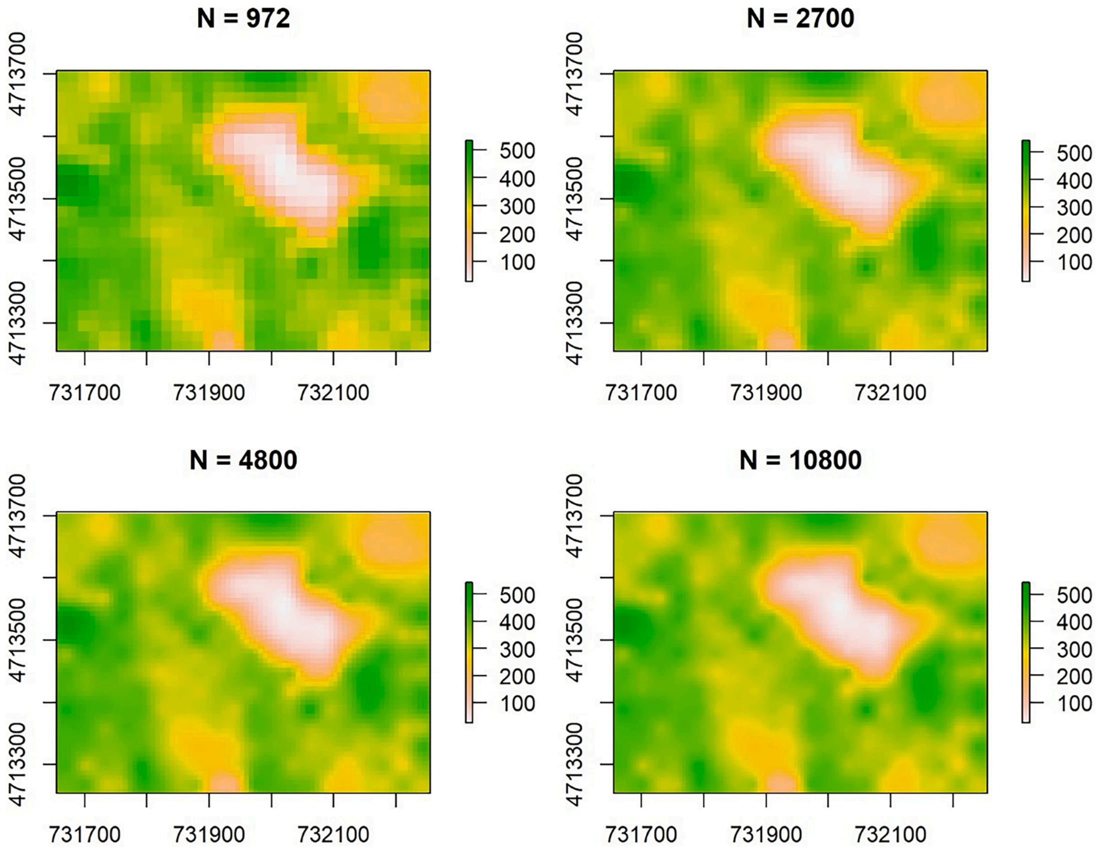

Fig. 2. Maps of AGB densities (m^3 x ha) for the Harvard Forest populations of $N = 972,\ 2700,\ 4800,\ 10,800$ squares, with average density of 322.28 m^3 per ha.

has produced a surface with a quadratic spatial trend. The two functions have then been rescaled to ensure a maximum AGB density of 500 and a minimum of 20. Finally, a real surface representing forest is considered in a region of 23,400 ha located in Tuscany (Central Italy). In this area, forest and non-forest categories were recorded in the year 2000 in correspondence with the 25 m square pixels partitioning the region. Data are available for download at https://forest.jrc.ec.europa.eu/en/past-activities/forest-mapping/#Downloadforestmaps.

From the four surfaces, four populations of $N = 972{,}2700{,}4800$ and 10,800 squares are constructed by partitioning the study regions into $27 \times 36,\ 45 \times 60,\ 60 \times 80,\ 90 \times 120$ squares. For the first three surfaces, the AGB densities within squares are taken as population values so that the four population means simply coincide with the average AGB per ha on the whole study area that does not change with N. Similarly, in the case of forest-non forest surface the population values are the fractions of forest extents within quadrats expressed as percentages, so that the four populations means coincide with the percentage of forest extent on the whole area (see Figs. 2–5).

Sampling from the resulting populations is performed by selecting $n = 81{,}225{,}400$ and 900 squares by means of EPSS, for a constant sampling fraction of 1/12. In practice, the populations are partitioned into blocks of 3×4 contiguous squares and one square is randomly selected in the first block and then repeated in the other blocks.

For each population, the design-based variance of the sample mean (V) is computed by means of Eq. (2) and is compared with the variance that would be achieved under SRSWOR (V_0) by the ratio DE=V_0/V to evidence the design effect, i.e. the improvement in precision achieved by EPSS. Moreover, for each population and each of the 12 possible samples, the variance estimates of the sample mean are computed by the estimators (3), (5), (6), (7), (9), (10), and the expectations of these estimators are computed as the mean of the 12 resulting estimates. Finally, for each population and each possible sample the estimated map of AGB densities or forest proportions is obtained from (11) and the variance estimate of the sample mean based on the NN interpolation is computed by means of (12). Once again, the expectation of the NN-based variance estimator is computed as the mean of the 12 resulting estimates.

The ratios between expectations of the variance estimators and the actual design-based variances are reported in Table 1 for each population and each estimator, with the exception of the HT estimator (6) whose ratios are not reported because the estimator turns out to be excessively conservative in all cases.

The results of Table 1 show a similarity between the DE values in the fourth columns of the tables and the ratios in the sixth columns

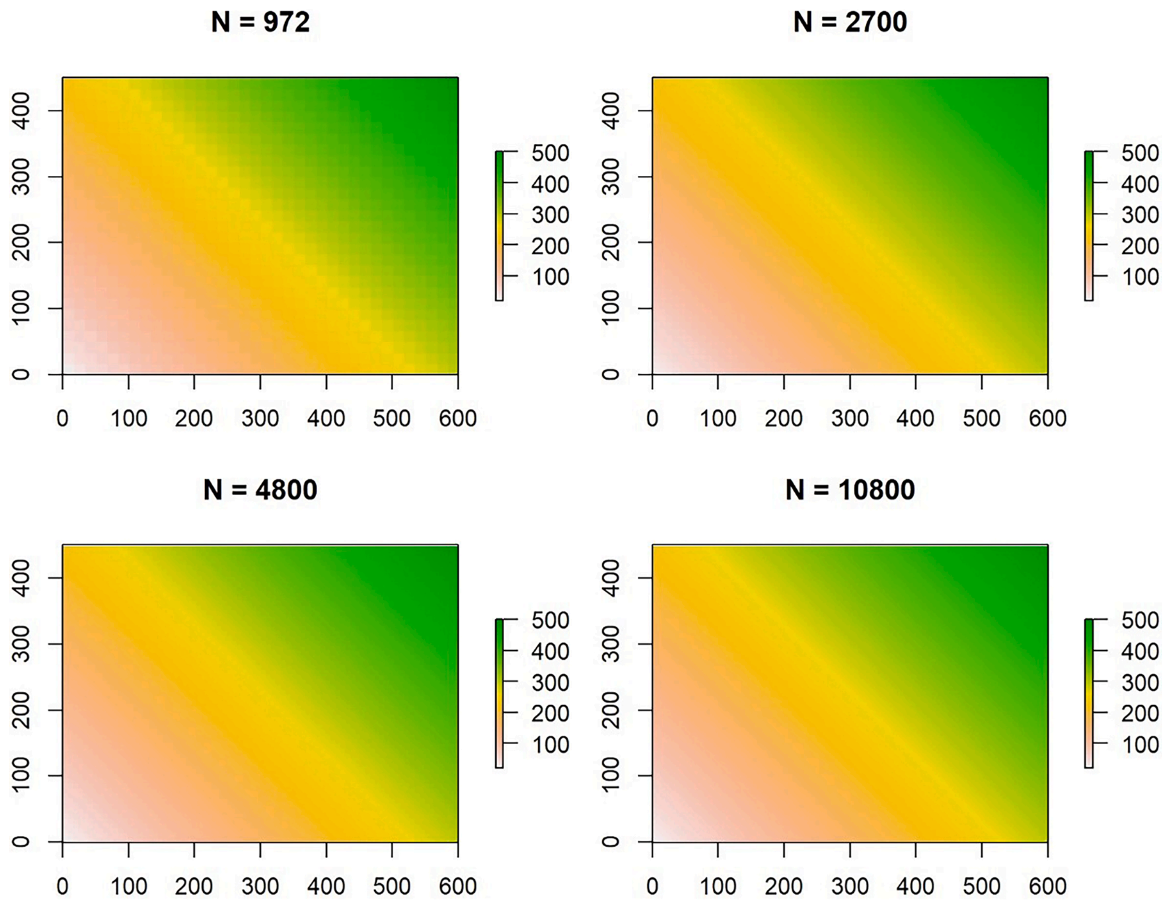

Fig. 3. Maps of AGB densities (m^3 x ha) for the populations of $N = 972,\ 2700,\ 4800,\ 10{,}800$ squares generated from the surface with a linear spatial trend, with average density of 260 m^3 per ha.

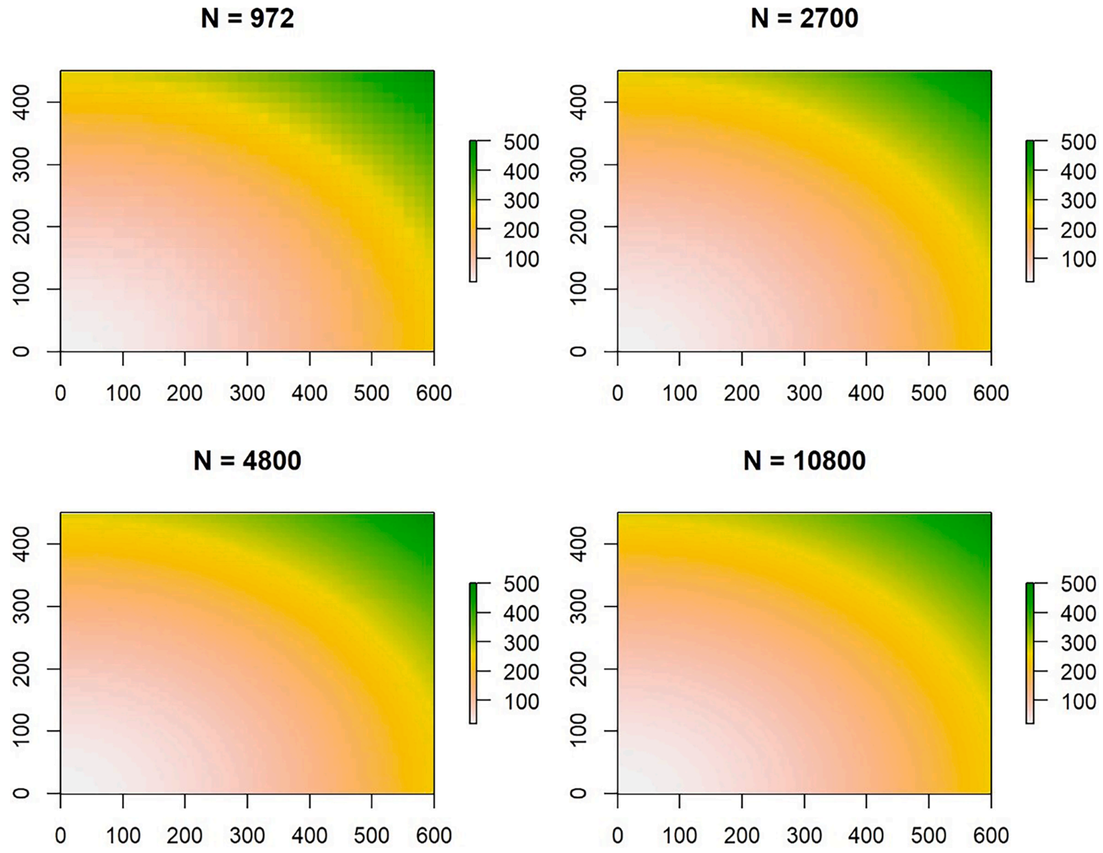

Fig. 4. Maps of AGB densities (m^3 x ha) for the populations of $N = 972,\ 2700,\ 4800,\ 10,800$ squares generated from the surface with a quadratic spatial trend, with average density of 180.26 m^3 per ha.

regarding the performance of the estimator (5). These similarities are due to the fact that, as proven in Appendix A of the Supplementary Material file, the DE values differ from the ratios quantifying the performance of estimator (5) by a quantity that vanishes for sufficiently large samples. Therefore, for large n these ratios are equivalent to DE indexes.

From the results of Table 1 it is apparent that the performance of the variance estimators (3), (5), (7), (9) and (10) greatly depend on the population characteristics. The naïve estimators (3) and (5) are too conservative, especially in the case of Harvard Forest populations but also in the case of forest extent populations. On the other hand, these estimators perform exceptionally well in the case of trended populations arising from the surfaces (13) and (14) with values of the ratios quantifying their performance that (owing to the approximation to the first two decimal digits) turn out to be invariably equal to 1. These results are quite surprising because owing to the similarity of these ratio with DEs, they evidence the inability of EPSS to provide improvements with respect to SRSWOR under linear and quadratic trends in both the spatial coordinates. These results are indeed in contrast with the literature on EPSS that, at least for one-dimensional populations, invariably stresses the EPSS effectiveness under trends (e.g., Särndal et al. 1992, example 3.4.2). On the other hand, linear trends in both the spatial coordinates entail some periodicities in the spatial populations of polygons that, as it is well recognized in literature, are likely to deteriorate the performance of EPSS (e.g., Wu and Thompson 2020, section 2.3). The failure of EPSS under linear trends is analytically demonstrated in Appendix B of the Supplementary Material file, where it is proven that the ratios quantifying the performance of estimator (5) do not depend on sample sizes and are near to 1. These results approximately hold also for estimator (3) that for large population and sample sizes is equivalent to (5) and also for quadratic trends of type (14) that generate populations similar to those achieved under linear trends.

Finally, regarding the estimators (7), (9) and (10), they provide severe under evaluations of the actual variances in the case of trended populations while providing conflicting results in the case of Harvard Forest and forest extent populations. In conclusion, at least for the considered populations, the variance estimator (12) appears to be the most stable performing satisfactorily in all situations with ratios that are the nearest to 1, thereby confirming the theoretical finding of the previous section.

6. Conclusions

Decades of studies on the estimation of the variance of the sample mean in EPSS have shown the weakness of searching for an omnibus estimator able to provide reliable variance estimates in any situations. As a matter of fact, the performance of any proposed

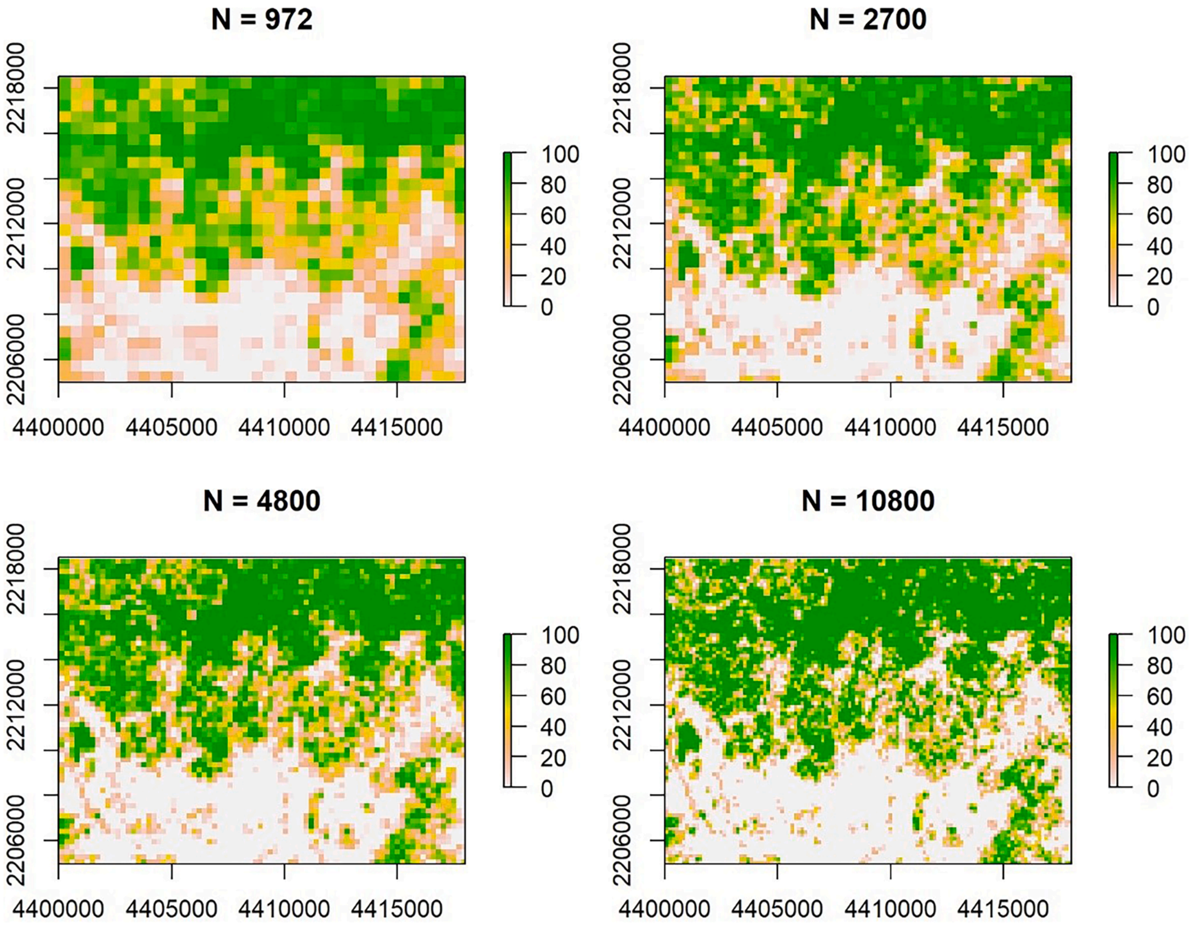

Fig. 5. Maps of percentages of forest extents for the populations of $N = 972,\ 2700,\ 4800,\ 10,800$ squares generated from the forest/non forest surface in a region of 24,300 ha located in Tuscany (Central Italy), with an average percentage of 47%.

estimator crucially depends on the characteristic of the surveyed population. In this framework, it seems suitable to search for a consistent estimator that invariably converges in probability to the true variance as the population and sample size increase.

In the case of spatial populations, a straightforward solution is to use the NN maps as pseudo-populations and to take the variances of the sample mean computed from them as estimates of the actual variances. As NN maps are consistent, in the sense that they converge to the true maps under an asymptotic scenario in which the study area is partitioned into an increasing number of regular polygons of decreasing size in such a way that population and sample size can approach infinity, the variances computed from these maps are also consistent.

Fattorini et al. (2022) show how the NN interpolation is the borderline case of IDW interpolation as the smoothing parameter approaches infinity. Both the interpolations ensure consistency so that we can use also IDW maps as pseudo-populations to achieve consistent variance estimators. We have preferred NN interpolation as it avoids the arbitrary choice of smoothing parameter or its computationally intensive data driven selection (Fattorini et al. 2023) and because the NN interpolated values have the same support of the survey variable, even when the support is discrete. This feature has practical relevance as it allows the application of NN interpolation for constructing maps of dichotomous 0–1 variables (e.g. forest/non forest maps).

Finally, it is worth noting that NN and IDW interpolations can be viewed as a model-assisted interpolation based on the very simple model assumption that neighbouring locations tend to be more similar than locations far apart. Even when this situation does not hold, consistency continues to hold because it is ensured in a design-based scenario by the use of EPSS. Thus, the unique crucial assumption is the Riemann integrability of the density function of the survey variable throughout the study region, that, as argued in Section 3, is likely to hold in most real situations

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.spasta.2024.100865.

Table 1
Variance (V) and design effect (DE) of EPSS in the four populations of quadrats derived from the Harvard forest surface (Table 1a), linear spatial trend surface (Table 1b), quadratic spatial trend surface (Table 1c) and percentages of forest extent derived from the forest-non forest surface (Table 1d) together with the ratios between the expectations of the variance estimators of sample mean achieved from the estimators (3), (5), (7), (9), (10), (12), denoted by *SRSWR*, *SRSWOR*, *STR*, *GEARY*, *MORAN*, *NN* respectively, and the actual variances of the third column.

Table 1a. Average density: 322.28 m^3 x ha

N	*n*	V	DE	*SRSWR*	*SRSWOR*	*STR*	*GEARY*	*MORAN*	*NN*
				(3)	(5)	(7)	(9)	(10)	(12)
972	81	3.47	28.03	30.91	28.33	13.01	16.15	3.24	2.37
2700	225	0.92	38.48	42.14	38.63	8.60	11.10	1.87	2.00
4800	400	0.49	40.72	44.52	40.81	6.75	7.54	1.19	1.86
10,800	900	0.26	33.76	36.86	33.79	2.66	3.16	0.46	1.37

Table 1b. Average density: 260 m^3 x ha

N	*n*	V	DE	*SRSWR*	*SRSWOR*	*STR*	*GEARY*	*MORAN*	*NN*
				(3)	(5)	(7)	(9)	(10)	(12)
972	81	111.61	1.00	1.09	1.00	0.04	0.05	0.03	1.32
2700	225	40.18	1.00	1.09	1.00	0.02	0.02	0.02	1.40
4800	400	22.60	1.00	1.09	1.00	0.01	0.01	0.01	1.41
10,800	900	10.05	1.00	1.09	1.00	0.004	0.005	0.01	1.51

Table 1c. Average density: 180.26 m^3 x ha

N	*n*	V	DE	*SRSWR*	*SRSWOR*	*STR*	*GEARY*	*MORAN*	*NN*
				(3)	(5)	(7)	(9)	(10)	(12)
972	81	108.68	1.07	1.17	1.07	0.05	0.07	0.04	1.40
2700	225	39.11	1.07	1.17	1.07	0.02	0.03	0.02	1.49
4800	400	22.00	1.07	1.17	1.07	0.01	0.01	0.02	1.51
10,800	900	9.78	1.07	1.17	1.07	0.01	0.01	0.01	1.62

Table 1d Average percentage: 47%

N	*n*	V	DE	*SRSWR*	*SRSWOR*	*STR*	*GEARY*	*MORAN*	*NN*
				(3)	(5)	(7)	(9)	(10)	(12)
972	81	4.95	2.88	3.17	2.91	1.24	1.22	0.28	1.66
2700	225	1.06	5.53	6.05	5.54	2.25	2.36	0.51	2.61
4800	400	1.27	2.81	3.07	2.81	1.17	1.20	0.26	1.27
10,800	900	0.31	5.66	6.18	5.66	2.41	2.44	0.53	2.07

References

Aronow, P.M., Samii, C., 2013. Conservative variance estimation for sampling designs with zero pairwise inclusion probabilities. Surv. Methodol. 39, 231–241.

Breidt, F.J., 1995. Markov chain designs for one-per-stratum sampling. Surv. Methodol. 21, 63–70.

Brus, D.J., Saby, N.P.A., 2016. Approximating the variance of estimated means for systematic random sampling, illustrated with data of the French soil monitoring network. Geoderma 279, 77–86. https://doi.org/10.1016/j.geoderma.2016.05.016.

Cliff, A.D., Ord, J.K., 1973. Spatial autocorrelation. Pion, London.

Cochran, W.G., 1946. Relative accuracy of systematic and random samples for a certain class of populations. Ann. Math. Stat. 17, 164–177. https://doi.org/10.1214/aoms/1177730978.

Cochran, W.G., 1977. Sampling Techniques, 3rd edn. Wiley, New York.

Conti, P.L., Marella, D., Mecatti, F., Andreis, F., 2020. A unified principled framework for resampling based on pseudo-populations: asymptotic theory. Bernoulli 26, 1044–1069. https://doi.org/10.3150/19-BEJ1138.

Cressie, N., 1993. Statistics for spatial data. Wiley, New York.

D'Orazio, M., 2003. Estimating the variance of the sample mean in two-dimensional systematic sampling. J. Agric. Biol. Environ. Stat. 8, 280–295. https://doi.org/10.1198/1085711032174.

Di Biase, R.M., Marcheselli, M., Pisani, C., 2024. Achieving spatial balance in environmental surveys under constant inclusion probabilities or density functions. Environmetrics, e2869. https://doi.org/10.1002/env.2869.

Dunn, R., Harrison, A.R., 1993. Two-dimensional systematic sampling of land use. J. R. Stat. Soc. C: Appl. Stat. 42, 585–601. https://doi.org/10.2307/2986177.

Fattorini, L., Marcheselli, M., Pratelli, L., 2018. Design-based maps for finite populations of spatial units. J. Am. Stat. Assoc. 113, 686–697. https://doi.org/10.1080/01621459.2016.1278174.

Fattorini, L., Marcheselli, M., Pisani, C., Pratelli, L., 2022. Design-based properties of the nearest neighbor spatial interpolator and its bootstrap mean squared error estimator. Biometrics 78, 1454–1463. https://doi.org/10.1111/biom.13505.

Fattorini, L., Franceschi, S., Marcheselli, M., Pisani, C., Pratelli, L., 2023. Design-based spatial interpolation with data driven selection of the smoothing parameter. Environ. Ecol. Stat. 30, 103–129. https://doi.org/10.1007/s10651-023-00555-w.

Franceschi, S., Di Biase, R.M., Marcelli, A., Fattorini, L., 2022. Some empirical results on nearest-neighbour pseudo-populations for resampling from spatial populations. Stats (Basel) 5, 385–400. https://doi.org/10.3390/stats5020022.

Hasel, A.A., 1938. Sampling error in timber surveys. J. Agric. Res. 57, 713–736.

Langsæter, A., 1926. Omberegning av middelfeilen ved regelmessige linjetakseringer (About calculation of standard error for systematic strip survey). Meddelelser fra Det Norske Skog-forsøksvesen 2, 5–43.

Li, J., Heap, A.D., 2008. A review of spatial interpolation methods for environmental scientists. Geoscience Australia, Canberra. Record 2008/23.

McGarvey, R., Burch, P., Matthews, J.M., 2016. Precision of systematic and random sampling in partitioned populations: habitat patches and aggregation organisms. Ecol. Appl. 26, 233–248. https://doi.org/10.1890/14-1973.

McRoberts, R.E., Vibrans, A.C., Sannier, C., Nasset, E., Hansen, M.C., Walters, B.F., Lingner, D.V., 2016. Methods for evaluating the utilities of local and global maps for increasing the precision of estimates of subtropical forest area. Can. J. For. Res. 46, 924–932. https://doi.org/10.1139/cjfr-2016-0064.

Milne, A., 1959. The centric systematic area-sample treated as random sample. Biometrics 15, 270–297. https://doi.org/10.2307/2527674.

Opsomer, J.D., Breidt, F.G., Moisen, G.G., Kauermann, G., 2007. Model-assisted estimation of forest resources with generalized additive models. J. Am. Stat. Assoc. 102, 400–409. https://doi.org/10.1198/016214506000001491.

Osborne, J.G., 1942. Sampling errors of systematic and random surveys of cover type areas. J. Am. Stat. Assoc. 37, 256–264.

Payandeh, B., 1970. Relative efficiency of two-dimensional systematic sampling. For. Sci. 16, 271–276. https://doi.org/10.1093/forestscience/16.3.271.

Quatember, A., 2016. Pseudo-populations. a basic concept in statistical surveys. Springer, Berlin.

Ripley, B.D., 1981. Spatial Statistics. Wiley, London.

Särndal, C.E., Swensson, B., Wretman, J., 1992. Model assisted survey sampling. Springer, New York.

Strand, G.H., 2017. A study of variance estimation methods for systematic spatial sampling. Spat. Stat. 21, 226–240. https://doi.org/10.1016/j.spasta.2017.06.008.

Tomppo, L.M., Gschwantner, R.E., McRoberts, R.E., 2010. National forest inventories: pathways for common reporting. Springer, Heidelberg.

Wolter, K.M., 1984. An investigation of some estimators of variance for systematic sampling. J. Am. Stat. Assoc. 79, 781–790. https://doi.org/10.1080/01621459.1984.10477095.

Wolter, K.M., 2007. Introduction to variance estimation, 2nd edn. Springer-Verlag, New York.

Yates, F., 1981. Sampling methods for censuses and surveys, 4th edn. Griffin, London.

Wu, C., Thompson, M.E., 2020. Sampling Theory and Practice. Springer, Cham (Switzerland).