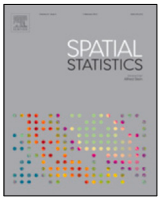
Spatial Statistics 41 (2021) 100477

Contents lists available at ScienceDirect

Spatial Statistics

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



A mixed sampling strategy for partially geo-referenced finite populations

Maria Michela Dickson [a,*], Flavio Santi [b], Emanuele Taufer [a], Giuseppe Espa [a]

[a] Department of Economics and Management, University of Trento, 38122, Trento, Italy
[b] Department of Economics, University of Verona, 37129, Verona, Italy

ARTICLE INFO

Article history:
Received 6 October 2019
Received in revised form 30 September 2020
Accepted 1 October 2020
Available online 9 October 2020

Keywords:
Spatial sampling
Design-based inference
Locational errors
Monte Carlo simulations

ABSTRACT

In the last few decades, sampling theory has been given a substantial boost by the growing availability of geo-referenced finite populations. Unfortunately, geo-referentiation is often incomplete or affected by locational errors for a portion of the units. Spatial sampling methods produce efficient estimates but suffer from consequences of flaws in geo-referentiation. This paper proposes a mixed sampling strategy for finite populations where a portion of the units is not correctly geo-referenced. The strategy exploits the available spatial information in the sampling design and adopts traditional sampling techniques for the remaining part of the population. Statistical properties of the strategy are explained and studied through Monte Carlo experiments on simulated and real data. An analysis of results in terms of efficiency and optimal sample composition is performed. The design-based nature of the proposed approach and its adaptability to several practical situations make it a general and easy-to-implement tool, which can outperform pure spatial sampling designs in terms of efficiency in estimation.

1. Introduction

Finite populations are distinguished as with- or without-frame: In the first case, the number of population units is known, and a complete list of units can be drawn up. In the second, population

* Corresponding author.
E-mail address: mariamichela.dickson@unitn.it (M.M. Dickson).

https://doi.org/10.1016/j.spasta.2020.100477
2211-6753/

size is unknown, and a list of units cannot be constructed. Dealing with the first situation, some information may be collected for all units of the population, and a sample is selected according to the chosen sampling design. In general, under the stipulated sampling design, the estimator properties are deduced from the distribution of all possible estimates. Considering that the higher the quality of information collected, the better the estimates for the phenomenon of interest, the sampling design is usually set up based on the statistical properties of adoptable estimators and on information collectible through the sampling procedure. This happens, for example, in the case of geo-referenced populations, which exhibit spatial autocorrelation among the units. Thus, in this type of situation, the sampling design and the estimation procedure of a feature can be conveniently defined to benefit from information about the locations of the units in a territory and either explicitly exploit the spatial autocorrelation or lessen its negative side effects, i.e. a decrease in precision, on the estimates. Finite populations can be efficiently sampled by exploiting spatial point-level information, as pointed out by the flourishing literature over the last two decades. Spatial sampling methods are mainly based on procedures that rely on the locations of the units (Hedayat et al., 1988; Hedayat and Stufken, 1998; Stevens and Olsen, 2004; Wright, 2008; Dickson and Tillé, 2016), or are based on distances among them (Grafström et al., 2012; Grafström, 2012; Grafström and Tillé, 2013). All methods have seen increasing success in literature and practice, due to the increasing availability of populations subjected to geo-referencing. As an example, although it is not feasible to map all units of a forest, recent research efforts have been directed over larger areas, with the consequent availability of geo-referenced wider areas, which make spatial sampling very useful in this field (Grafström et al., 2017). Or again, many national statistical institutes are building geo-referenced archives of businesses, to exploit spatial information of the units in economic studies (Cozzi and Filipponi, 2012).

Ideally, when a spatial sampling design for geo-referenced populations is adopted, the geographical coordinates of all units should be observed correctly. In actual fact, however, geo-referencing processes, both terrestrial and airborne scanning, often result in populations where a portion of the units is affected by non-negligible locational errors. Inaccuracy in the location of the units may be due to different reasons, such as geo-masking for privacy protection or technical problems in automated geo-coding. The latter occur when GPS trackers used in the surveyed area cannot establish a reliable connection to satellites at that time, e.g., because of a momentary technical malfunction, so that it is not possible to precisely detect locations, and some units are mislocated. In this common practical situation, other variables not concerning the position in space of the units are collected, to avoid a waste of time and money, and locations are imputed to one or some points of the sub-area interested by the problem. In some cases, a reasonable choice is to use geo-imputation algorithms that exploit known auxiliary variables to impute missing locations. However, these methods have limitations. The most notable are the assumptions that units similar for a given variable are certainly close, and that a given relationship among the target variable and covariates exists. For a discussion of geo-imputation techniques, see Henry and Boscoe (2008), Curriero et al. (2010) and Poloczek et al. (2014).

A less restrictive and preferable option is to assign the units whose locations cannot be correctly tracked to the centroid of the area to which they belong (Allshouse et al., 2010), to avoid introducing a new source of uncertainty with geo-imputation methods. However, even if recent improvements to reduce inaccuracy in geo-coding of units are given by the integration of different sources of information (e.g., in the environmental context, see Giannetti et al., 2018), errors about locations persist in affecting finite spatial populations. Therefore, locational errors affect the statistical properties of finite population estimators of the target variable total, introducing inefficiency in the estimation process if the mislocated units are treated as correctly geo-referenced ones (Dickson et al., 2018), or running into coverage errors if the affected units have null probabilities to be sampled.

Exploiting the improvement in estimation due to spatial sampling methods, this paper proposes a sampling strategy that explicitly handles the problem of estimating the population total in partially geo-referenced populations. In particular, a sampling strategy that mixes spatial and non-spatial designs is proposed, to avoid coverage errors and biases resulting from the selection of mislocated units. The entire strategy is dealt with in a design-based perspective, so that no model assumption

is needed to manage the units affected by locational errors, either in the imputation phase or in the estimation procedure.

The paper is structured as follows: In Section 2, the methodological framework is defined and the proposed mixed sampling strategy presented. In Section 3, the statistical properties of the strategy are deepened. In Section 4, implementation issues of the proposed sampling strategy are discussed. In Section 5, the mixed sampling strategy is compared to spatial and non-spatial designs on simulated data with Monte Carlo simulations. An example on forestry data is also provided. In Section 6, the significance of the results of the work is deepened. In Section 7, the paper is concluded.

2. A mixed sampling strategy

Let $U = \{1, \ldots, k, \ldots, N\}$ be a finite population of N units. Given the probability space $(\Omega, \mathcal{A}, \mathbb{P})$, a measurable sampling design over U is a random set S defined as measurable mapping $S: (\Omega, \mathcal{A}) \to (\Omega_U, \mathcal{A}_U)$, where $(\Omega_U, \mathcal{A}_U)$ is a measure space over the set Ω_U of all admissible samples in U. Thus, a random sample $s \in \mathcal{A}_U$ is a realization of the sampling design S. If the probability of sample s to be selected is denoted with $p(s) \equiv \mathbb{P}[S = s]$, probability axioms guarantee that $\sum_{s\in\Omega_U} p(s) = 1$ (see e.g. Resnik, 1999).

If, for any unit $k \in U$, π_k denotes the probability that k is included in a sample drawn according to the sampling design S, then

$$\pi_k = \mathbb{P}[k \in S] = \sum_{s\in\{A\in\Omega_U : k\in A\}} p(s) = \sum_{s\in\Omega_U} \mathbb{1}_{\{k\in s\}} p(s). \tag{1}$$

It can be proved that inclusion probabilities π_k (for $k = 1, \ldots, N$) are known and are strictly positive for any $k \in U$. Thus, the population total

$$Y_U = \sum_{k\in U} y_k \tag{2}$$

of a variable of interest y can be consistently and unbiasedly estimated from sample s through the Horvitz–Thompson (H–T) estimator (Horvitz and Thompson, 1952):

$$\hat{Y}_U = \sum_{k\in s} \frac{y_k}{\pi_k}\,. \tag{3}$$

In the case of equal inclusion probabilities (e.g., in simple random sampling without replacement), the H–T estimator has the form $\hat{Y}_U = \frac{N}{n}\sum_{k\in s} y_k$.

When dealing with spatial populations, units of U lie in a two-dimensional space $X \subset \mathbb{R}^2$, so that their location is given by a pair of topographic coordinates (x_{1k}, x_{2k}), for any $k \in U$. Assume that a geo-referentiation process has been carried out for all N units of the population, and that the process failed for a number $N_M < N$ of units, which will be imputed to an incorrect location. The population units can then be gathered into two subsets U_G and U_M, which collect the N_G geo-referenced units and the N_M mislocated ones, respectively. The aim is to estimate the population total Y_U, and a sample is drawn from U for this purpose.

As U_G and U_M are disjoint sets ($U_G \cap U_M = \emptyset$), and the locational errors in a sub-area do not affect the correct geo-referencing procedure in others, the population total Y_U can be computed as the sum of the sub-population totals:

$$Y_U = Y_G + Y_M\,, \tag{4}$$

where $Y_G = \sum_{k\in U_G} y_k$ and $Y_M = \sum_{k\in U_M} y_k$.

According to decomposition (4), the two sub-populations U_G and U_M could be considered as two strata of U; thus, two different sampling strategies may be implemented for each sub-population. In particular, a spatial sampling design may be applied to U_G to exploit the geo-referential information available for units in U_G, whereas a non-spatial sampling scheme may be applied to U_M. The estimate of the population total is computed as the sum of the estimates of Y_G and Y_M may then be computed, and the population total can be estimated as

$$\hat{Y}_U = \hat{Y}_G + \hat{Y}_M\,. \tag{5}$$

The *mixed sampling strategy* (MSS) to which estimator (5) belongs permits the geo-referential information in U_G to be used, whereas it does not suffer from problems arising from geo-referential errors in U_M, because the location information is ignored for those units.

3. Statistical properties

3.1. Consistency and unbiasedness

The MSS estimator (5) is the sum of estimators $\hat{Y}_G$ and $\hat{Y}_M$. Thus, if they consistently estimate Y_G and Y_M, respectively, (5) will consistently estimate the population total Y_U, as a consequence of basic properties on convergence in probability (see e.g. van der Vaart, 1998).

Analogously, from the linearity of the MSS estimator, it follows that if $\hat{Y}_G$ and $\hat{Y}_M$ are unbiased estimators for sub-population totals Y_G and Y_M, then the properties of the expected value guarantee that (5) will be an unbiased estimator for population total Y_U.

Therefore, the consistency and unbiasedness of estimators $\hat{Y}_G$ and $\hat{Y}_M$ are *sufficient* conditions for the consistency and unbiasedness of the MSS estimator.

3.2. Variance and variance estimation

The variance of the H–T estimator (3) for fixed-size samples is given by $Var(\hat{Y}_U) = -\frac{1}{2}\sum_{k,i\in U}(\pi_{ki} - \pi_k\pi_i)(\frac{y_k}{\pi_k} - \frac{y_i}{\pi_i})^2$, where π_{ki} is the probability that a pair of units k and i is selected in the same sample, also known as second-order inclusion probability. If the sampling designs defined over U_G and U_M are independent, then $\hat{Y}_G$ and $\hat{Y}_M$ are independent H–T estimators of partial totals Y_G and Y_M, and the covariance $\text{cov}(\hat{Y}_G, \hat{Y}_M)$ equals zero. Thus, the variance of the sum of the two independent H–T estimators can be derived as the sum of the variance of the H–T estimators, such as $Var(\hat{Y}_U) = Var(\hat{Y}_G) + Var(\hat{Y}_M)$.

Under the proposed mixed sampling strategy, the expression of variance estimation depends on sampling methods that are implemented on U_G and U_M, respectively. To select units in U_M, any non-spatial sampling method may be used, for example, simple random sampling without replacement (SRSWOR). To derive variance estimation on this partition of the population, the well-known Sen–Yates–Grundy variance estimator (Särndal et al., 1992) is a feasible option. On the other side, in the area where the geo-referentiation process has been successful, any spatial sampling method can be applied. A vast literature is devoted to this topic, especially in the last few years, also due to the increase in spatial information availability (Stevens and Olsen, 2004; Grafström, 2012; Grafström et al., 2012; Grafström and Tillé, 2013). Selection procedures are mainly based on the exclusion of sampling neighboring units, to obtain samples spread over the area under study, that is spatially balanced samples. Avoiding nearby units appearing in the same sample, spatial sampling methods may produce some null second-order inclusion probabilities for units close together, for example, for populations clustered in small groups. This makes the use of the Sen–Yates–Grundy variance estimator unfeasible or by using approximations, able only to produce biased estimates. Other alternatives, such as the Hajék–Rosén estimator (Hajék, 1981; Rosén, 1997a,b) or the local mean variance estimator (Stevens and Olsen, 2003), usable according to specific sampling designs, do not constitute optimal solutions for estimating variance for spatial sampling. Nevertheless, the Hansen–Hurwitz variance estimator, which does not involve second-order inclusion probabilities, tends to produce overestimates of the sampling variance (Wolter, 2007). This problem was reviewed by Grafström et al. (2012). It is clear that finding a design-unbiased variance estimator of $\hat{Y}_U$ is challenging and depends strongly on the spatial structure of the population under analysis.

4. Some empirical remarks

To better understand the advantages of using the proposed mixed sampling strategy instead of other methods, some issues deserve special attention.

The implementation of the MSS requires some a priori choices to be made. First, the sampling design S should be set up over U, and designs S_G and S_M should be independent. Thus, both may

be chosen regardless of the other, according to the characteristics of the respective target sub-populations and the available information about them (such as geographic coordinates). As the MSS estimator $\hat{Y}_U$ inherits its statistical properties from estimators $\hat{Y}_G$ and $\hat{Y}_M$, a sound choice of sampling designs S_G and S_M will positively affect the statistical efficiency of $\hat{Y}_U$. Second, the sample size n and the share $\psi \in [0, 1]$ of units in the sample drawn from U_G should be fixed, so that ψn and $(1-\psi)n$ are the number of sample observations drawn from U_G and U_M, respectively. Generally, the value of sample size n is driven by practical constraints, including time, costs and feasibility of a survey (see e.g. Groves, 2004). However, the choice about the sample composition in terms of (relative) sub-sample sizes (that is, the value of ψ) is inherently statistical, and should be made a priori and according to some criterion. The share ψ has to be set to minimize the mean squared error of $\hat{Y}_U$. Unfortunately, a general indication for defining ψn and $(1-\psi)n$ does not exist, because it is closely related to the population and sampling designs adopted over U_G and U_M. However, the choice may be guided by the intention of respecting in the sample the proportion of units present in the two sub-areas of the population, that is, $\psi = \xi$, where $\xi = N_G/N$ is the portion of correctly geo-referenced units in U, or by another choice if it is preferable or required.

Another aspect that needs further discussion concerns the treatment of mislocated units and the preferability of MSS compared to other sampling designs. Three alternative approaches may be considered in handling locational errors, such as removing units that lack their correct locations from the target population, imputing the location of these units to a unique arbitrary location within the shadow sub-area and randomly imputing these locations over the shadow sub-area from which they come. All three options allow any spatial sampling design to be applied, as they always produce a population where all units are geo-referenced. However, they have different implications. In the first case, some units are removed from the population, and therefore, they cannot be sampled, leading to biased estimates of the population total (4). The implications of the second approach essentially depend on how the spatial sampling algorithm used manages units that stay in on the same location. A large variety of methods for sampling spatial finite populations have been presented in the literature (see De Gruijter et al., 2006, and Wang et al., 2012, for detailed reviews), ranging from traditional methods, such as simple random sampling or systematic sampling to sample points or plots, to methods that explicitly consider the position in the space of the units and distances among them in the selection process, increasing the efficiency in the estimation. The latter start from the idea that nearby units show similar values in the target variable. Thus, distances among units may be exploited to avoid selecting neighbors in the same sample. The spread of the sample can be achieved in different ways, as demonstrated by a wide range of contributions (see Benedetti et al., 2017, for a review). We focus on several recent articles that present local pivotal methods (LPM, Grafström et al., 2012, spatially correlated Poisson sampling (SCPS, Grafström, 2012) and local cube (LC, Grafström and Tillé, 2013, due to their remarkable efficiency and very fast implementation. When some units of a population are mislocated and lie in the same location, for example, the centroid of the shadow area, all these methods can be applied on the full population by disregarding this fact because computing of the distance among the units is always possible. Nevertheless, the distance is not correctly computed among the units in the shadow sub-area, and among the units in U_M and the border units in U_G, causing the impossibility to select a well-spread sample onto the study area. In fact, the implementation of a spatial sampling method, for example, the LPM, on a partially geo-referenced population leads to a spatially spread sub-sample in U_G and to a simple random sub-sample in U_M, with the loss in efficiency that is entailed if compared with the application of the spatial sampling method on a full correctly geo-referenced population. A similar argument can be made when mislocated units are randomly imputed to two or more locations in the shadow sub-area: A spatial sampling method may be applied on the whole population, but the quality of results will be affected by the incorrect locations.

5. Simulation experiments

In this section, the results of three Monte Carlo experiments are illustrated and discussed where the MSS is compared to the SRSWOR and the LPM. The use of local pivotal method is illustrative,

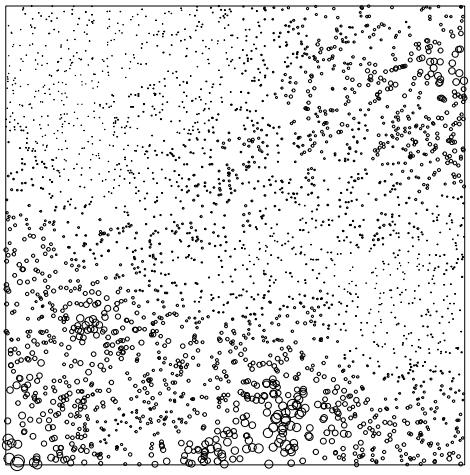


Fig. 1. Population of $N = 3000$ units in a unit square $X = [0, 1]^2$ used in the Monte Carlo simulation. The size of the points is proportional to the values y_k of the variable of interest y.

and it was chosen for its better performances in estimation and time of execution compared to similar selection methods.

The first experiment aims at investigating the use of the MSS when $\psi = \xi$, and the similarities highlighted among the LPM and the MSS in this case. The second Monte Carlo simulation considers examples where $\psi = \xi$ is not the optimal choice, and it is shown that in such situations the MSS outperforms the LPM. The third experiment is a comparison of the methods on a real environmental dataset.

To conduct the first two experiments, a population of $N = 3000$ units was generated over a unit square (thus $X = [0, 1]^2 \subset \mathbb{R}^2$) according to a uniform Poisson process (see e.g. Diggle, 2014). For each unit $k \in U$, the values of y_k are generated by a log-Gaussian random field with $\mu = 0$, $\sigma^2 = 1$, and correlation function $\rho(d) = \mathrm{e}^{-\frac{d}{\phi}}$, where $\phi = 1$, and d is the Euclidean distance between units (for further details about Gaussian random fields, see e.g. Diggle and Ribeiro, 2007, Ch. 3). The resulting population is shown in Fig. 1. Note that evident spatial clusters result in bottom-left and top-right sub-areas where the values y_k are large, while in the central zone along the NW-SE direction the values y_k are relatively small.

In the population, a shadow sub-area $X_M \subset X$ has been delimited according to several criteria. As Fig. 2 shows, shadow sub-areas are circles with two different centers and various radii. The centers of the coordinates $c_1 = (0.1, 0.1)$ and $c_2 = (0.7, 0.7)$ localize the shadow sub-areas where the variable of interest y exceeds the population mean and where y is approximately in line with it, respectively. The radii of the circles are set to include in the shadow sub-area 10%, 20% and 40% of the population units (300, 600 and 1200 units, respectively). These sizes are purely illustrative because the size of the shadow sub-areas may be of any size. Units belonging to the shadow sub-areas have been collapsed all in the same point, such as the centroid of the belonging sub-area (Fig. 2).

The performances of the methods are evaluated in all the experiments in terms of the relative root mean squared error, defined as $rRMSE = \frac{\sqrt{\sum_{i=1}^{R}(\hat{Y}_{iU} - Y_U)^2/R}}{Y_U}$ where $R = 10\,000$ Monte Carlo replications of the estimators.

5.1. MSS when $\psi = \xi$

In this section, the performances of the MSS design when $\psi = \xi$ are compared with Monte Carlo simulations with those of the SRSWOR and the LPM. This possibility occurs when a wise choice is

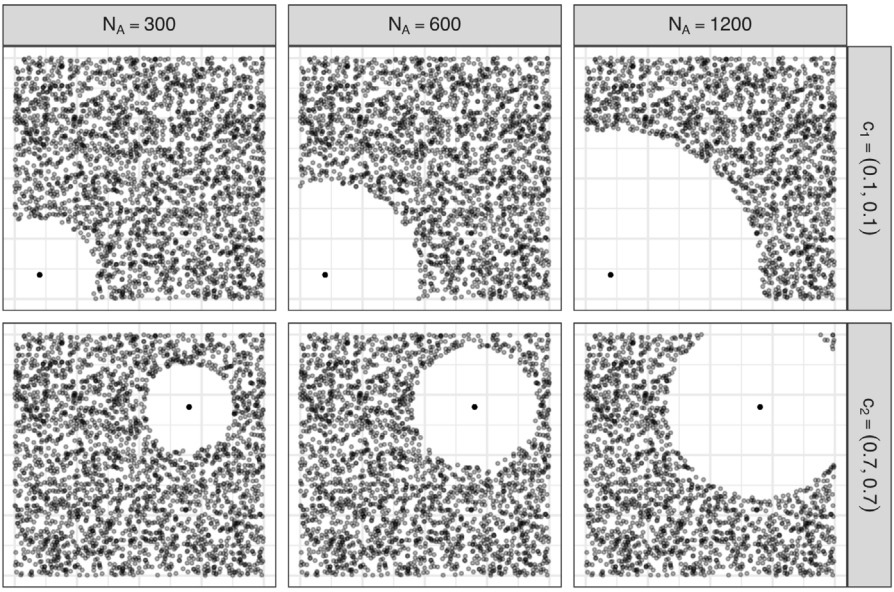


Fig. 2. Population represented in Fig. 1 where the six shadow sub-areas are illustrated. Conditions $N_M = 300; 600; 1200$ correspond to 10%, 20% and 40%, respectively, of the population size.

to proportionally represent the two strata of the population, that is, the geo-referenced sub-area and the mislocated one. The MSS was based on an LPM design over the correctly geo-referenced sub-population U_G and the SRSWOR over the mislocated sub-population U_M, whereas the sample composition parameter ψ was equal to $\xi = N_G/N$. The experiment was carried out on for sample sizes equal to 150, 300 and 600, which correspond to the sampling fractions of 5%, 10% and 20%, respectively.

Simulation results are reported in Table 1. Results confirm the analysis previously stated, as the MSS and the LPM are equivalent sampling designs in terms of efficiency if the MSS is based on samples where the number of correctly geo-referenced and mislocated units proportionally reproduces the composition of the population (that is, $\psi = \xi$). The rRMSE values are essentially the same despite the sample size, the center of the shadow area and the number of non-geo-referenced units.

The proposed experiment investigates the case in which only an area of the population is shaded. Actually, it could frequently happen that the geo-coding process is imperfect over some areas of the population. In these situations, however, selecting spatial samples is possible, and the considerations above may be extended. In fact, the MSS can be easily applied with more than one shadow area if $\psi = \xi$, because it is always possible to select units proportional to the size of each area. Furthermore, the LPM applied on the whole incorrectly geo-referenced population may experience a loss in estimation efficiency, as this approach currently faces a greater source of "disturbance" in properly computed distance among the units. In addition, with this method the proportional selection of units from all areas may not be respected, due to rounding problems.

5.2. MSS when $\psi \neq \xi$

The second experiment analyzes how the rRMSE of the MSS changes with ψ. In many practical cases, the size of n needs to be set differently than a proportional representation of the sub-areas of

Table 1
Relative root mean squared error (rRMSE) of the estimators of the population total Y_U under different mis-geo-referentiation schemes for 10 000 Monte Carlo replications. The centers of the coordinates $c_1 = (0.1, 0.1)$ localize the shadow sub-areas where the variable of interest y exceeds the population mean.

Shadow zone		Design	Sample size (n)		
Center	N_M/N		150	300	600
Clean data		SRSWOR	0.0689	0.0420	0.0279
Clean data		LPM	0.0263	0.0127	0.0073
c_1	10%	LPM	0.0301	0.0168	0.0104
c_1	10%	MSS	0.0301	0.0166	0.0103
c_1	20%	LPM	0.0338	0.0195	0.0126
c_1	20%	MSS	0.0334	0.0196	0.0122
c_1	40%	LPM	0.0526	0.0315	0.0209
c_1	40%	MSS	0.0518	0.0316	0.0208
c_2	10%	LPM	0.0270	0.0135	0.0079
c_2	10%	MSS	0.0268	0.0135	0.0078
c_2	20%	LPM	0.0339	0.0195	0.0126
c_2	20%	MSS	0.0338	0.0195	0.0124
c_2	40%	LPM	0.0353	0.0201	0.0128
c_2	40%	MSS	0.0353	0.0200	0.0124

the population. This happens, for example, when some zones are difficult to reach due to the shape of the territory. Then, it may be decided to adequately set ψ according to the practical fact.

The settings of the simulation experiment are the same adopted in the previous section except the following:

1. Only the shadow sub-area centered in $c_1 = (0.1, 0.1)$ and including $N_M = 600$ units (20% of the population) was considered.
2. The MSS estimator was computed for $\psi \in \{0.05, 0.10, 0.15, \ldots, 0.85, 0.90, 0.95\}$.
3. Heteroscedastic random fields were considered. In addition to the homoscedastic case (referred to as *case* 1) where $\sigma_k = 1$ for any unit $k \in U$, a random field where $\sigma_k = 1 + (c_{x_1k}^2 + c_{x_2k}^2)^{-1}$ (referred to as *case* 2) and $\sigma_k = 1 + (c_{x_1k}^2 + c_{x_2k}^2)$ (referred to as *case* 3) are considered, having been denoted with (c_{x_1k}, c_{x_2k}) the Cartesian coordinates of unit $k \in U$.

The forms of heteroscedasticity specified in the latter point have different implications in terms of the variance of the variable of interest among the correctly geo-referenced and mislocated units. In particular, case 1 implies that the variance of y is the same over U_G and U_M; case 2 implies that the variance of y is larger over U_M than over U_G, whereas case 3 implies that the variance of y is larger over U_G than over U_M.

The results of the simulations are summarized graphically in Fig. 3. Tables of complete results are reported in supplementary material, Appendix A.

Some considerations must be drawn. First, as verified in the previous simulation experiment, the MSS estimator performed as well as the LPM when $\psi = \xi$ (vertical dotted line in Fig. 3). However, Fig. 3 shows that if ψ is properly set, the MSS may exhibit a lower rRMSE than the LPM. Second, the maximum gain in terms of the rRMSE that can be obtained with the optimal choice of sample composition parameter ψ is bounded by the difference in the rRMSE between the LPM estimator and the LPM estimator based on correctly located data (called the *Clear LPM*). Formally:

$$rRMSE(\hat{Y}^{(\mathrm{MSS})}_{\psi^*}) - rRMSE(\hat{Y}^{(\mathrm{ClearLPM})}) \leq rRMSE(\hat{Y}^{(\mathrm{LPM})}) - rRMSE(\hat{Y}^{(\mathrm{ClearLPM})}). \tag{6}$$

Third, the upper bound (6) is affected by the sample size and by the presence and form of heteroscedasticity. As Fig. 3 shows, bound (6) decreases as the sample size gets larger and when the variance of y_k for mislocated units of U_M gets smaller compared to that of geo-referenced units U_G.

Fourth, in case 1 (homoscedasticity) and 2 (the variance over U_M larger than the variance over U_G) the rRMSE curve is relatively flat around its minimum, suggesting that if $N_M^{-1} \sum_{k \in U_M} \sigma_k^2 \geq N_G^{-1} \sum_{k \in U_G} \sigma_k^2$, parameter ψ should be considerably smaller than ξ. However, the performances

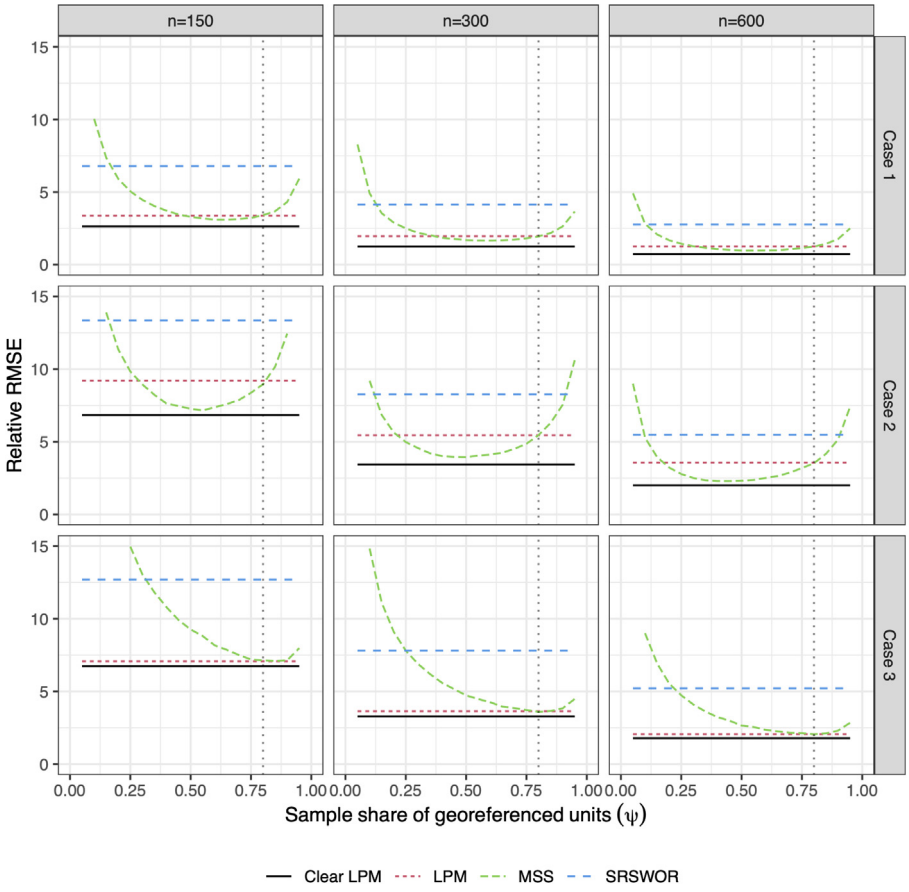


Fig. 3. Relative root mean squared error (rRMSE) of the LPM design based on correctly geo-referenced data (*clear LPM*, drawn as a black solid line), LPM (red dotted line), MSS (green dashed line) and SRSWOR (blue dashed line). The rRMSE (y-axis) is shown as a function of ψ (x-axis). Values were multiplied by 100. As all designs except the MSS do not depend on ψ, their rRMSE is depicted as an horizontal line. The dotted vertical line (gray) shows the value of $\xi = 0.8$ (thus, $N_M = 600$). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of the MSS estimator are not sensitive to small changes in the value of ψ. The sensitivity of the rRMSE of the MSS estimator is larger in case 3, where the optimal sample composition requires that $\psi^* > \xi$. The choice of ψ leads to a moderate reduction in the rRMSE of the MSS estimator, as previously noted.

5.3. An example on real data

To give an example on real data, Monte Carlo simulations are used for comparing the estimation methods on the `longleaf` dataset of the R package `spatstat` (Baddeley and Turner, 2005) collected by Platt et al. (1988). The dataset includes information concerning 584 Longleaf pine (*Pinus palustris*) trees settled in a forest portion of 40 000 square meters in southern Georgia (USA), which although a not very big area is useful for conducting and showing a practical experiment. For each tree, the location and the bole basal area at breast height (1.30 m, expressed in cm) are available. The diameters of the trees range from 2 cm to 75.9 cm, with the mean equal to 26.84 cm and the median

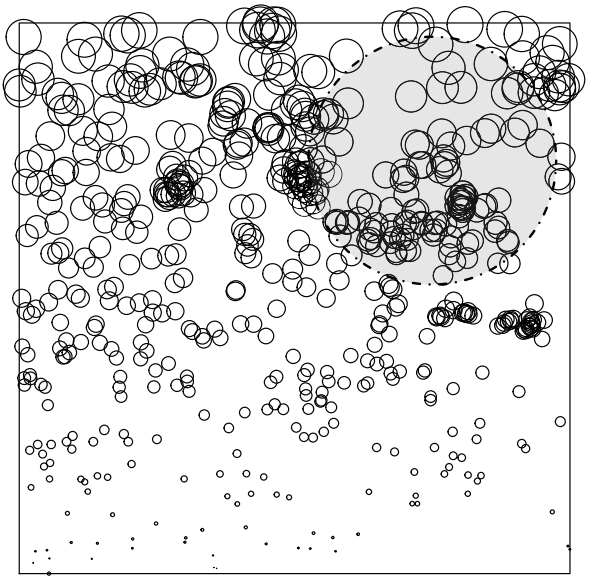


Fig. 4. Population of 584 Longleaf pine trees in an area of 200 × 200 m (`longleaf` dataset). The circles are centered on the pines' coordinates, and their sizes are proportional to the pines' diameters. The gray circle delimits the shadow sub-area.

Table 2
Relative root mean squared error (rRMSE) of the population total estimators for different sample sizes under various sampling designs: simple random sampling without replacement (SRSWOR), local pivotal method (LPM), local pivotal method on original correctly-geo-referenced data (Clear LPM) and mixed sampling strategy (MSS) for various values of ψ. The Monte Carlo simulations use the `longleaf` dataset and are based on 10 000 replications.

Design	ψ	Sample size (n)				
		29	58	88	117	175
SRSWOR	–	0.1236	0.0851	0.0666	0.0559	0.0435
Clear LPM	–	0.0909	0.0594	0.0441	0.0349	0.0264
LPM	–	0.1120	0.0682	0.0525	0.0425	0.0322
MSS	0.70	0.1010	0.0638	0.0472	0.0395	0.0289
MSS	0.75	0.0983	0.0627	0.0471	0.0386	0.0283
MSS	0.80	0.0971	0.0623	0.0470	0.0378	0.0289
MSS	0.85	0.1010	0.0632	0.0490	0.0395	0.0298
MSS	0.90	0.1030	0.0668	0.0526	0.0429	0.0337

equal to 26.15 cm. As adult trees are conventionally defined as those trees with a diameter greater than or equal to 30 cm, the dataset contains 313 young trees and 271 adult trees. The population shows spatial correlation among the units, as shown in Fig. 4.

The gray circle represents a shadow sub-area that includes 20% of population units; thus, $N_M = 117$, and $\xi = 0.8$. Sampling designs were applied for various sampling fractions: 5% (29 observations), 10% (58 observations), 15% (88 observations), 20% (117 observations) and 30% (175 observations). The simulation is based on 10000 replications of a random experiment where SRSWOR, LPM and MSS designs are applied to the population to estimate the total bole basal area. The MSS design was applied with $\psi = 0.7; 0.75; 0.80; 0.85; 0.9$. Results of Monte Carlo simulations are reported in Table 2.

As expected, in this case the SRSWOR is also the most inefficient design, whereas the difference in the rRMSE between the Clear LPM and LPM designs is small (see (6)). When $\psi = \xi = 0.8$, so that the share of the units in the sample and the portion of correct geo-referenced units in the

population are equal, the MSS performs better than the LPM, although they should be equivalent, as previously stated. In the present case, the improvement is probably slight because of the border effect, which tends to emerge more frequently in populations of moderate size and is more evident with different populations.

6. Discussion

Having to deal with finite populations partially geo-referenced, in which a portion of any size of the units is mislocated, is a challenge for researchers and practitioners in many fields of research. When a sample has to be selected from a population, it cannot be overlooked that this is affected by locational errors. This is because the introduction of sources of errors should be avoided when spatial sampling methods are used, which constitute a valuable tool for selecting samples from spatial finite populations. To avoid losing all the positive effects in terms of efficiency and representativeness of spreading a sample over the territory, namely, to use spatial sampling, this study tries to give an answer: Exploit spatial point-level information, when it is available, and resort to traditional random sampling methods, when it is not, without excluding any unit from the possible selection.

The proposed MSS is able to handle the cited issues without resorting to models or strong assumptions. Populations affected by locational errors may be viewed as stratified populations, on which is possible to apply different methods on different sub-areas. The explained statistical properties and the presented examples demonstrate the strong adaptability of the method to many practical situations. Specifically, the flexibility in choosing the sample size in each sub-population, correctly geo-referenced and mislocated, is particularly relevant in those cases where the sampling happens forcibly in a given area, for example, due to a particular shape of the territory that makes it the only choice, or to a prescribed and not very modifiable number of units. Moreover, regarding what is said about the sampling methods to mix, the proposed strategy is simply extendable to several other spatial sampling designs based on the distance among the units, for the correctly geo-referenced sub-area, and to several traditional sampling methods, for the mislocated one. As previously stated, multiple shadow areas may occur in practice, and the mixed sampling strategy reveals its strengths once again, due to the possibility of properly setting the share of units to be sampled in each area, which is an avenue not available with spatial algorithms alone.

7. Conclusions

In this paper, a mixed sampling strategy is proposed for partially geo-referenced populations. The use of a spatial sampling design for the geo-referenced sub-population and a non-spatial design on the incorrectly geo-referenced sub-population makes the overall sampling strategy consistent with the available information on population units. In addition, it makes the strategy flexible, in terms of sub-population sampling fractions and in terms of sub-population sampling strategies. The properties of the proposed estimator are ascribable to the properties of the well-known H–T estimator, as sub-populations are independent and thus, estimators of sub-population totals.

The Monte Carlo simulations carried out on artificial and real data clearly showed that the proposed mixed sampling strategy outperforms the exclusive use of spatial sampling designs if the composition parameter ψ is properly set. When $\psi = \xi$, and SRSWOR is adopted for incorrectly geo-referenced sub-population, the performances of the mixed sampling strategy are similar to those of the LPM adopted on the whole population imperfectly geo-referenced, with the advantage of setting the portion of units to select in each stratum of the population. In addition, the MSS permits the composition of the sample ψ to be set independently of the population composition ξ. A general rule for setting parameter ψ close to the optimum ψ^* is not available but may be decided from time to time based on specific requirements of the survey and the population characteristics. Improvements in the performances of the MSS estimator could potentially be attained by exploiting available auxiliary information in the selection step, to use unequal probability sampling designs over both sub-areas where units lie.

References

Allshouse, W.B., Fitch, M.K., Hampton, K.H., Gesink, D.C., Doherty, I.A., Leone, P.A., Serre, M.L., Miller, W.C., 2010. Geomasking sensitive health data and privacy protection: an evaluation using an e911 database. Geocarto Int. 25 (6), 443–452.

Baddeley, A., Turner, R., 2005. Spatstat: An R package for analyzing spatial point patterns. J. Stat. Softw. 12 (6), 1–42.

Benedetti, R., Piersimoni, F., Postiglione, P., 2017. Spatially balanced sampling: a review and a reappraisal. Internat. Statist. Rev. 85 (3), 439–454.

Cozzi, M., Filipponi, D., 2012. The new geospatial business register of local units: potentiality and application areas. In: 3rd Meeting of the Wiesbaden Group on Business Registers-International Roundtable on Business Survey Frames, Washington, DC, pp. 17–20.

Curriero, F.C., Kulldorff, M., Boscoe, F.P., Klassen, A.C., 2010. Using imputation to provide location information for nongeocoded addresses. PLoS One 5 (2), e8998.

De Gruijter, J., Brus, D.J., Bierkens, M.F., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer Science & Business Media.

Dickson, M.M., Giuliani, D., Espa, G., Bee, M., Taufer, E., Santi, F., 2018. Design-based estimation in environmental surveys with positional errors. Environ. Ecol. Stat. 25 (1), 155–169.

Dickson, M.M., Tillé, Y., 2016. Ordered spatial sampling by means of the traveling salesman problem. Comput. Statist. 31 (4), 1359–1372.

Diggle, P.J., 2014. Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, third ed. CRC Press.

Diggle, P., Ribeiro, Jr., P., 2007. Model Based Geostatistics. springer, New York.

Giannetti, F., Puletti, N., Quatrini, V., Travaglini, D., Bottalico, F., Corona, P., Chirici, G., 2018. Integrating terrestrial and airborne laser scanning for the assessment of single-tree attributes in mediterranean forest stands. Eur. J. Remote Sens. 51 (1), 795–807.

Grafström, A., 2012. Spatially correlated Poisson sampling. J. Statist. Plann. Inference 142 (1), 139–147.

Grafström, A., Lundström, N.L., Schelin, L., 2012. Spatially balanced sampling through the pivotal method. Biometrics 68 (2), 514–520.

Grafström, A., Tillé, Y., 2013. Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. Environmetrics 24 (2), 120–131.

Grafström, A., Zhao, X., Nylander, M., Petersson, H., 2017. A new sampling strategy for forest inventories applied to the temporary clusters of the swedish national forest inventory. Can. J. Forest Res. 47 (9), 1161–1167.

Groves, R.M., 2004. Survey Errors and Survey Costs, Vol. 536. John Wiley & Sons.

Hajék, J., 1981. Sampling from a Finite Population. Marcel Dekker.

Hedayat, A., Rao, C., Stufken, J., 1988. Sampling plans excluding contiguous units. J. Statist. Plann. Inference 19 (2), 159–170.

Hedayat, A., Stufken, J., 1998. Sampling designs to control selection probabilities of contiguous units. J. Statist. Plann. Inference 72 (1), 333–345.

Henry, K.A., Boscoe, F.P., 2008. Estimating the accuracy of geographical imputation. Int. J. Health Geogr. 7 (1), 3.

Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. J. Amer. Statist. Assoc. 47 (260), 663–685.

Platt, W.J., Evans, G.W., Rathbun, S.L., 1988. The population dynamics of a long-lived conifer (*Pinus palustris*). Amer. Nat. 131 (4), 491–525.

Poloczek, J., Treiber, N.A., Kramer, O., 2014. Knn regression as geo-imputation method for spatio-temporal wind data. In: International Joint Conference SOCO14-CISIS14-ICEUTE14. Springer, pp. 185–193.

Resnik, S.I., 1999. A Probability Path. Birkhäuser.

Rosén, B., 1997a. Asymptotic theory for order sampling. J. Statist. Plann. Inference 62 (2), 135–158.

Rosén, B., 1997b. On sampling with probability proportional to size. J. Statist. Plann. Inference 62 (2), 159–191.

Särndal, C., Swensson, B., Wretman, J., 1992. Model Assisted Survey Sampling. springer, New York.

Stevens, Jr., D.L., Olsen, A.R., 2003. Variance estimation for spatially balanced samples of environmental resources. Environmetrics 14 (6), 593–610.

Stevens, Jr., D.L., Olsen, A.R., 2004. Spatially balanced sampling of natural resources. J. Amer. Statist. Assoc. 99 (465), 262–278.

van der Vaart, A.W., 1998. Asymptotic Statistics. Cambridge University Press.

Wang, J.-F., Stein, A., Gao, B.-B., Ge, Y., 2012. A review of spatial sampling. Spat. Statist. 2, 1–14.

Wolter, K., 2007. Introduction to Variance Estimation. Springer Science & Business Media.

Wright, J.H., 2008. Two-dimensional balanced sampling plans excluding adjacent units. J. Statist. Plann. Inference 138 (1), 145–153.