View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Stirling Online Research Repository

https://doi.org/10.1007/s10651-017-0396-9

Noname manuscript No.
(will be inserted by the editor)

A proposal for a two-step sampling design to oversample units responding to prescribed characteristics

Federico Andreis · Marco Bonetti

Received: date / Accepted: date

Abstract We introduce a novel method to extract a sample from a finite population where units with desired characteristics are over-represented. The approach is both sequential and adaptive and allows, via suitable compositions of predictive and objective functions, to target specific subsets of the population. We consider the problem of estimation and conjecture the validity of a modified Horvitz-Thompson estimator capable to account for the imbalance induced by the targeting procedure. After discussing how to apply the method to the sampling of geographically distributed units, we investigate its potential via simulations.

Keywords Adaptive sampling · π-ps designs · Sequential methods · Hot deck imputation · Oversampling · Spatial sampling

1 Introduction

Efficient and reliable monitoring systems are essential components of the management of wildlife and agricultural ecosystems, as pointed out by many authors (see, for example, [11] and references therein). Within this framework, survey sampling is usually the tool of choice, and proposals on how to efficiently implement surveys for environmental monitoring purposes abound in the literature. In particular, during recent years, the topic of adaptive sampling has drawn the attention of the research community and has been studied

Federico Andreis
Dondena Centre for Research on Social Dynamics and Public Policy, Bocconi University, Milano, Italy
E-mail: federico.andreis@unibocconi.it

Marco Bonetti
Dondena Centre for Research on Social Dynamics and Public Policy, Bocconi University, Milano, Italy

and applied in a variety of instances in the environmental field ([29], [22], [26], [9]). Adaptive sampling is characterized by the fact that the procedure used to select units depends on evidence collected during the survey itself; a notable feature of adaptive designs is that they are typically well suited to surveying populations where the variable of interest has a highly skewed distribution, also in a geographical sense ([28], [27], [20]); for example, when dealing with a dichotomous survey variable such as the presence or absence of a rare and geographically clustered characteristic, adaptive sampling strategies have proved to be reliable in over-representing units that have the trait of interest, while retaining the possibility of drawing valid inference [1]. Over-representation of units responding to prescribed characteristics may be a highly desirable feature, especially when resources are limited: in the environmental setting, budget constraints usually limit the possible survey effort ([11]), hence emphasising the need for an efficient use of the available resources. Relevant examples include expensive field studies needed to identify local maxima such as sources of pollution, areas where soil erosion is reaching critical levels, or where individuals of rare species are observed: the ability to obtain samples where the sought-after characteristics are more likely to appear, is then very important. Existing adaptive designs, such as the commonly employed Adaptive Cluster Sampling (ACS, [30]) and the more recent proposals of Adaptive Geostatistical Designs (AGD, [8], [20]), are successful in providing the desired over-representation and guarantee valid sample-based inference; however, they suffer from important drawbacks, mainly related to their producing variable sample sizes (ACS, see [1] for a discussion, and [5], [13], [14] for some recent proposals addressing the issue), to their difficulty of implementation (AGD), and to their inability to provide control on the magnitude of the over-representation (both). An extensive comparison of all of these approaches, while outside the scope of this work, could provide additional important information.
In this paper, we outline a novel design-based method for selecting samples with fixed size from finite populations that aims at achieving two main goals: i) valid estimation of population parameters, and ii) controlled over-representation of units that possess prescribed characteristics. Our approach differs from the existing literature in that it constitutes a simple, fixed size and non-parametric alternative to current design- (e.g., ACS) and model-based (e.g., AGD) methods, and offers greater flexibility in terms of defining and controlling the objectives and magnitude of the over-representation than other available options. Despite the (desirable) conceptual simplicity of our method, resort to numerical methods is necessary when it comes to the estimation task, due to the dual sequential and adaptive nature of the approach: for this reason, some discussion is devoted to the development of a suitable algorithm that allows the estimation of the design's inclusion probabilities, in the spirit of [10].

The paper is structured as follows: in Section 2 we describe our proposal and outline the algorithm for its implementation. Section 3 contains details on choices for the design setup and on how to inform and control the feature of oversampling of units with prescribed characteristics; the potential of the method for applications to spatial sampling problems is also discussed. In Section 4 we address the problem of estimation. In Section 5 we present empirical evidence supporting the potential of our proposal, also in a spatial setting, and conclude the paper with some final remarks in Section 6.

2 The design

Consider a finite population U of size N, and let the units in U be indexed by $k = 1, ..., N$; denote by y_k and x_k the values associated to the k-th unit of, respectively, a non-negative survey variable Y and a positive variable X positively correlated with Y, usually labelled as the auxiliary information. Let Y be real, and X possibly vector-valued. X is known for all subjects, while Y is unknown. The object of interest is a population parameter θ that can be expressed as a function of the distribution of Y, such as the population total or mean; besides the estimation task, the targeting of specific subsets of the population is also of interest. Let $n = n_1 + n_2$ be a prescribed fixed sample size. Our proposal is a two-step sampling design in which a sample $\mathbf{s}_1$ of size n_1 is initially collected via Simple Random Sampling (SRS) from U; the sample evidence is then used to define the new variable

$$\widetilde{Y}_k = \mathbb{I}_{\mathbf{s}_1}(k) Y_k + [1 - \mathbb{I}_{\mathbf{s}_1}(k)]\, \widehat{Y}_k, \quad k \in U \tag{1}$$

where $\mathbb{I}_A(k)$ is the indicator function that takes on value 1 if k is in set A, 0 otherwise, and $\widehat{Y}_k$ is an estimator of the survey variable value for out-of-sample units as a function of $y_k, k \in \mathbf{s}_1$ and $x_k, k \in U$. At the second step, an additional set $\mathbf{s}_2$ of size n_2 is collected from $U \setminus \mathbf{s}_1$ by means of a fixed-size π-ps design using $\widetilde{Y}$ as auxiliary variable. The final sample is $\mathbf{s} = \mathbf{s}_1 \cup \mathbf{s}_2$, of size n. Without loss of generality, in this paper we consider the Pareto design (see, for example, [25]) for the second step extraction.
We stress that this approach does not constitute an example of two-stage sampling, since the sampling units are all at the same 'level' (there is no distinction between primary and secondary sampling units, such as in cluster sampling), nor of two-phase or double-sampling, since there is no subsampling and the survey variable is assumed to be available for all units in $\mathbf{s}$.

Denote with $\boldsymbol{\pi}_1$ and $\widehat{\boldsymbol{\pi}}_2$ the vectors of inclusion probabilities for the units available for sampling at each of the two steps; in our setting, the elements of $\boldsymbol{\pi}_1$ are $\pi_{k;1} = n_1/N, \forall k \in U$, while the elements of $\widehat{\boldsymbol{\pi}}_2$ are $\widehat{\pi}_{k;2}, k \in U \setminus \mathbf{s}_1$, and are a function of $\widetilde{Y}$.
The algorithm can be summarized as follows:

Algorithm 1

```
1. Draw $\mathbf{s}_1$ of size $n_1$ from $U$ using SRS
2. Construct $\widetilde{Y}$ based on $\mathbf{s}_1$
```

```
3. Compute $\widehat{\boldsymbol{\pi}}_2$ based on $\widetilde{Y}$
4. Draw $\mathbf{s}_2$ of size $n_2$ with probabilities $\widehat{\boldsymbol{\pi}}_2$ from $U \setminus \mathbf{s}_1$ using Pareto.
```

The final sample is then $\mathbf{s} = \mathbf{s}_1 \cup \mathbf{s}_2$, with size $n_1 + n_2 = n$.

Note that the probability function of this design is, in general, not trivial, as it encompasses two steps of randomization: the selection of $\mathbf{s}_1$, and of $\mathbf{s}_2$ conditionally on $\mathbf{s}_1$. In general terms, it should be possible to express it using chain rules in the form $p(\mathbf{s}) = p(\mathbf{s}_1)p(\mathbf{s}_2|\mathbf{s}_1)$, but the specific formula will of course be dependent on the design choices, and may not be easy to describe analytically.

3 Prediction and targeting

To implement the proposed method, choices about how to perform the prediction task and how to use the predicted values to inform the second sampling step must be made; clearly, these decisions will influence both the ability of the design to deliver the desired over-representation and its level of complexity.
In principle, any predictive machinery could be used to obtain the values $\widehat{y}_k$, from classical statistical modelling to state-of-the-art machine learning methods. In this paper, we consider a non-parametric approach common in the missing data literature, namely the distance Hot Deck imputation (HD, see [2] for a recent review); moreover, we discuss the possibility of employing the Inverse Distance Weighting interpolator (IDW, see [6], [12]), by interpreting the prediction task as a spatial mapping problem.

3.1 The HD approach

The prediction is handled as a missing data imputation problem, where no specific assumptions on how Y and X are related is made, other than accepting that units close in the X space should possess similar Y values, which is consistent with the reasons why one would choose to use a π-ps design. Moreover, imputation by distance HD is simple to implement, computationally light and possesses desirable consistency properties [21]. Following the notation introduced in Section 2, $\widehat{y}_k$ will denote the survey variable value prediction for out-of-sample units, obtained via HD imputation. Specifically,

$$\widehat{y}_k = y_j, \quad k \in U \setminus \mathbf{s}_1, j \in \mathbf{s}_1 \tag{2}$$

where

$$j = \text{argmin}_{j \in \mathbf{s}_1} \mathit{\Delta}(x_j, x_k), k \in U \setminus \mathbf{s}_1 \tag{3}$$

and $\mathit{\Delta}$ is a suitable dissimilarity function. Although in this paper we consider only auxiliary information that is measured on a numerical scale (hence, for example, we would use the city block or the euclidean distance), in principle any kind of variable could be used, under a suitable choice of Δ.

The predictions $\widehat{y}_k$ obtained via HD imputation are mapped into a set of inclusion probabilities $\widehat{\boldsymbol{\pi}}_2$ via a transformation that achieves the goal of targeting units with specific values of the dependent variable by assigning larger probabilities to such units. For the sakes of illustration, consider the target units to be those whose Y values belong to some set $\mathcal{T}$. A simple example of targeting function is

$$\varphi(\widehat{y}_k; c) = \widehat{y}_k \left[\mathbb{I}_{\mathcal{T}}(\widehat{y}_k) \cdot c + [1 - \mathbb{I}_{\mathcal{T}}(\widehat{y}_k)] \cdot c^{-1}\right] \tag{4}$$

where $c > 0$ is an arbitrary *boosting* factor. The inclusion probabilities for the second step are then defined as

$$\widehat{\pi}_{k;2} = n_2 \frac{\varphi(\widehat{y}_k; c)}{\sum_{i \in U \setminus \mathbf{s}_1} \varphi(\widehat{y}_i; c)}, k \in U \setminus \mathbf{s}_1, \tag{5}$$

so that $\sum_{k \in U \setminus \mathbf{s}_1} \widehat{\pi}_{k;2} = n_2$. If $c = 1$ in Equation 4, the inclusion probabilities are computed in a way that is exactly proportional to the estimated Y values, regardless of the choice of the set $\mathcal{T}$. Alternatively, choosing $c > 1$ allows to boost the probability for units estimated to have values of Y in $\mathcal{T}$, while penalizing those estimated to have values outside of $\mathcal{T}$; the converse happens if $c \in (0, 1)$. Simple examples of $\mathcal{T}$ include: the set of values exceeding a fixed treshold t, $\mathcal{T} = \{y \in \mathbb{R}^+ : y > t\}$, the set of values within a certain interval (t_1, t_2), $\mathcal{T} = \{y \in \mathbb{R}^+ : y \in (t_1, t_2)\}$, and sets defined by unions of disjoint intervals, such as $\mathcal{T} = \{y \in \mathbb{R}^+ : y < t_1 \cup y > t_2)\}$. The performance of the proposed HD approach will be investigated, via simulation, in Section 5.

3.2 The spatial approach and the IDW interpolator

In many settings, the spatial component of the phenomenon under study is of primary importance, as is the case for environmental studies: our proposal naturally extends to the spatial sampling framework, by considering the geographical coordinates of the sampling units as part of the auxiliary information available. In fact, it is possible to implement our sampling strategy by relying on the sampling units locations alone as auxiliary information, as we show in Section 5.2.

The problem of obtaining good $\widehat{y}_k$ predictions is akin to the problem of constructing a good map of a spatial population based on a probabilistic sample. The task of constructing a map using an estimator that has good properties in a design-based framework has been recently addressed in [12]: the authors derive conditions under which the IDW interpolator possesses design-based asymptotic unbiasedness and consistency for the estimation of the y_k values for the whole population. The IDW interpolator of the density f_k of the survey variable for unit k is defined as

$$\widehat{f}_k = \mathbb{I}_{\mathbf{s}_1}(k) f_k + [1 - \mathbb{I}_{\mathbf{s}_1}(k)] \sum_{i \in U} w_{ik} f_i \tag{6}$$

where the w_{ik}s are suitable weights attached to the density of unit i to estimate the density of unit k. As stressed in [12], we note that usually, in practice, the area of the spatial units is known, hence drawing inference on f_k is equivalent to drawing inference on y_k itself. The conditions under which the IDW interpolator provides the asymptotic unbiasedness and consistency essentially require the underlying sampling design to provide a certain degree of spatial balance; this is the case for some designs commonly employed in environmental studies, such as SRS and stratified sampling with proportional allocation. Since we consider SRS as a sampling procedure for the first step, we reckon that it would suffice to use the IDW interpolator with a suitable choice of weights in order to obtain 'good' predictions of the $\widehat{Y}_k$s. The predictions are, in turn, used to compute the inclusion probabilities for the second sampling step as discussed in Section 3.1: ideally, the possibility of obtaining a good approximation of the Y via the predictive tool should provide us with an auxiliary variable able to improve the targeting.

We wish to stress that the presented approach may be implemented by using general variables in arbitrary spaces, not only geographical systems for physical regions: our method can be used to target subset of the population by exploring a possibly highly multidimensional covariates space with no further adaptation than the definition of a dissimilarity measure that is suitable to the nature of those covariates. In Section 5.2 we present an application to a scenario inspired by a real data problem, to highlight the spatial capabilities of our approach.

4 Estimation

In the following, we will focus on $\theta = N^{-1} \sum_{k \in U} y_k$, the population mean, as the parameter of interest. Consider its Horvitz-Thompson (HT) estimator

$$\widehat{\theta}_{HT} = N^{-1} \sum_{k \in \mathbf{s}} \frac{y_k}{\pi_k} \tag{7}$$

where $\pi_k > 0$ denotes the first-order inclusion probabilities for unit k, i.e., $\pi_k = P(k \in \mathbf{s})$. Unfortunately, in spite of the conceptual simplicity and ease of implementation of the design, the inclusion probabilities of any order are impossible to compute, since they depend on the unknown values $y_1, ..., y_N$. When the inclusion probabilities depend only on known characteristics, such as the auxiliary variable X, available for all the units in U, they can be consistently estimated via Monte Carlo as shown in [10], which in turn leads to a modified HT estimator that is asymptotically equivalent to the original one.

In the same spirit, we suggest a resampling procedure that we conjecture capable of achieving analogous results under our design. The procedure can be described as follows:

(i) obtain a 'best possible' approximation $\widetilde{Y}$ of the survey variable Y based on the final sample $\mathbf{s}$ for all units in U
(ii) apply the sampling procedure used to obtain $\mathbf{s}$ to collect a new sample from U, using $\widetilde{Y}$ in lieu of Y as the true population values, and store it
(iii) iterate point (ii) a certain number, say B, of times.

An estimate of any-order inclusion probability is given by the proportion of inclusions in the B samples of the relevant units.
A thorough investigation of how to perform the task of obtaining the 'best possible' representation of Y is beyond the scope if this paper, and it will not be discussed here; for simplicity, we again consider nonparametric imputation via distance HD as a convenient way to build $\widetilde{Y}$. We are aware, however, that this step is fundamental in determining the validity of the modified estimator, and hence provide extensive empirical evidence to support our choice of using hot-deck imputation.

Let $\widetilde{Y}_k$ be defined as in Equation (1), with $\widehat{Y}_k$ obtained via HD imputation based on $\mathbf{s}$ and X (known exactly for all population units). The following algorithm formalizes the resampling strategy.

```
Algorithm 2

For $b = 1, ..., B$:

  1. select a sample $\mathbf{s}_{1b}$ of size $n_1$ using SRS from $U$
  2. build a HD prediction $\widetilde{Y}^*_{1b}$ based on $\widetilde{Y}_k, k \in \mathbf{s}^*_{1b}$
  3. compute $\widehat{\boldsymbol{\pi}}^*_{2b}$ using the same combination of targeting function and
     boosting factor as in the original design
  4. draw a Pareto sample $\mathbf{s}^*_{2b}$ of size $n_2$ with probabilities $\widehat{\boldsymbol{\pi}}_{2b}$ from $U \setminus \mathbf{s}^*_{1b}$
  5. store $\mathbf{s}^*_b = \mathbf{s}^*_{1_b} \cup \mathbf{s}^*_{2_b}$.
```

Figure 1 depicts the generic b-th run of algorithm. The first-order inclusion probabilities can then be estimated using

$$\widehat{\pi}_k = \frac{Z_k + 1}{B + 1} \tag{8}$$

where $Z_k = \sum_{b=1}^{B} \mathbb{I}_{\mathbf{s}^*_b}(k)$ is the number of times unit k enters the B samples.

The estimator returns strictly positive values and constitutes our approximation of the true π_k values; we conjecture that the goodnes of such approximation improves as $B \to \infty$, provided that $\widetilde{Y}_k, k \in U$ is a good enough representation of the original population Y. Once the estimates $\widehat{\pi}_k$ have been

$$\begin{tikzcd}
U \arrow[dr, "\text{SRS}"] & & \widetilde{Y}^*_{1b} \arrow[dr, "\pi\text{-ps}", "\text{targeting}"'] & & \mathbf{s}^*_b \\
& \mathbf{s}^*_{1b} \arrow[ur, "\text{prediction}"] & & \mathbf{s}^*_{2b} \arrow[ur, "\mathbf{s}^*_{1b} \cup \mathbf{s}^*_{2b}"] &
\end{tikzcd}$$

Fig. 1 Resampling procedure

obtained via resampling, a natural modification of the classic HT estimator is then

$$\widehat{\theta}_B = N^{-1} \sum_{k \in \mathbf{s}} \frac{y_k}{\widehat{\pi}_k} \tag{9}$$

Our sampling scheme guarantees strictly positive second-order inclusion probabilities $\pi_{kj}, \forall k \neq j \in U$, that can be estimated in analogy with Equation 8 via

$$\widehat{\pi}_{kj} = \frac{Z_{kj} + 1}{B + 1} \tag{10}$$

where $Z_{kj} = \sum_{b=1}^{B} \mathbb{I}_{\mathbf{s}^*_b}(k, j)$ is the number of times units k and j jointly enter the B samples. Again, we conjecture that such approach leads to reasonable approximations of the true unknowable π_{kj}s to be, in turn, used to obtain an estimate of the variance of $\widehat{\theta}_B$ via a modified estimator of the variance of the HT estimator $\widehat{\theta}$, as in [10]. We note, however, that satisfactorily estimating the π_{kj}s requires a very large number of runs B, thus increasing sensibly the computational effort, together with assuming that $\widetilde{Y}$ provides indeed an adequate representation of Y. A possible way to overcome this added computational burden could be to investigate approximations of second order inclusion probabilities by means of estimated first-order ones, in the spirit of [17]; this will be subject to further research.

5 Empirical evidence

This section contains the results of Monte Carlo simulations aimed at investigating the properties and potential of our proposal. We present empirical evidence concerning:

- the recovery of the true π_ks via the proposed Algorithm 2
- the sampling distribution of the modified HT estimator
- the targeting feature and its impact on over-representation

These aspects are investigated under two scenarios: i) a simulated dataset in which a univariate auxiliary variable X, highly correlated with the survey variable Y, is used; and ii) a spatial setting, where X is bivariate and represents geographical coordinates of the sampling units over a region. All computations have been carried out in `R 3.4.2` [24]; the following packages have been used:

copula [31], gstat [15], HotDeckImputation [19], sp [23], and spatstat [3]. The code for the simulations is available upon request from the authors.

5.1 Univariate auxiliary variable

Consider an artificial population of $N = 1000$ units where the survey and the auxiliary variables Y and X are jointly generated via a Gaussian copula with parameter $\rho = 0.75$ and margins, respectively, LogNormal and Gamma. The marginal parameters are (μ-log $= 1.5, \sigma$-log $= 0.5$) for the LogNormal and $(\text{shape} = 15, \text{scale} = 0.2)$ for the Gamma.
We simulated the extraction of 20000 samples of size $n = 100$ from the population under three designs:

- simple random sampling without replacement (SRS)
- Pareto sampling, with X as the auxiliary variable (PAR)
- the new method introduced here (HD)

where for our proposal we considered a targeting set defined by the values y exceeding the median value $y_{0.5}$, i.e., $\mathcal{T} = \{y \in \mathbb{R}^+ : y > y_{0.5}\}$. We investigated three levels of boosting factor in (4), namely $c \in \{2, 4, 15\}$ and denote with HD1, HD2 and HD3, respectively, the corresponding outcomes. Moreover, the sampling effort is split equally between the two steps, i.e., $n_1 = n_2 = 50$. Lastly, $B = 2000$ runs of resampling have been employed in Algorithm 2.

Figure 2 contains the results of the simulation study. The artificial population, for which the joint (X, Y) distribution is shown in the top-left panel, has been generated using positive margins, which is often the case when dealing with real world physical measurements such as those common in the environmental surveys (concentration of a pollutant, abundance of a species, deadwood volume, etc). Clock-wise, the second panel shows the Monte Carlo distributions of the HT estimators under SRS and Pareto designs, together with the modified HT estimators that we have proposed for our design; the percentage relative bias with respect to the true mean of Y (represented by the dashed horizontal line) are reported below each boxplot. The HT estimator is unbiased by construction, while our modified version appears to be only approximately so; our conjecture that the magnitude of the bias can be somewhat controlled with a large enough number of resampling runs B seems to be confirmed in the present scenario: the bias appears to be decreasing with B (results not shown here), although we suspect that a better approximation could be obtained by accounting in a more precise way for the specific value of the boosting factor when computing the estimates. For what concerns the variability of the modified HT estimator, further simulative results (not shown) seem to indicate a non trivial relationship with c which appears, however, to be negligible in this example. Interestingly, the variance seems comparable to that of the simple π-ps approach. The third panel shows the ability of Algorithm 2 to recover the true inclusion probabilities for each unit: for each $c \in \{2, 4, 15\}$, and for each Monte Carlo run, $B = 2000$ resampling runs have been used to obtain the $\widehat{\pi}_k$s. Since a closed form expression for the first-order inclusion probabilities is not available, we have run a separate Monte Carlo simulation with the same sets of parameters to select 200000 samples with our method and we have counted the relative frequency for each unit in the samples. These values have been taken as the true inclusion probabilities π_k for comparison purposes: the plot shows the Monte Carlo percentage relative bias $(\widehat{\pi}_k - \pi_k)/\pi_k \cdot 100\%$ for HD1-3. The Monte Carlo distributions of the biases seem to be centered around zero, although we observe an increase in skewness and variability with

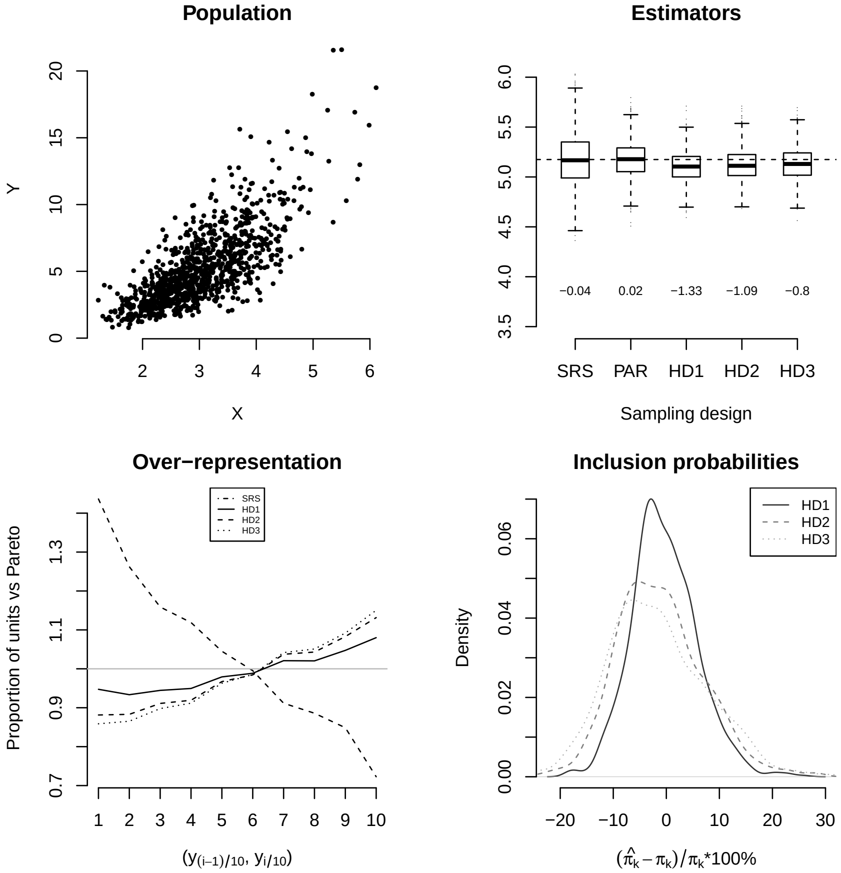

Fig. 2 Simulated population results. Clockwise from top-left panel: scatterplot of simulated population values – Monte Carlo distribution of the HT estimators of the mean under SRS and PAR, and the modified HT under HD1-3 (percentage relative bias shown under each boxplot) – kernel density estimates of percentage relative bias in estimating the true π_ks under HD1-3 – Monte Carlo proportion of units per 10% Y-quantiles intervals for SRS and HD1-3, with respect to PAR.

c, likely due to the stronger imbalancing induced by the larger boosting factor. The fourth panel explores the oversampling feature by contrasting the Monte Carlo proportion of units in the final sample that fall within 10% intervals of the distribution of Y under SRS and HD1-3 to the same quantity under PAR; specifically, the values corresponding to $i = 1, ..., 10$ on the abscissa are the number of units in the intervals $(y_{(i-1)/10}, y_{i/10})$, where y_p denotes the p-th percentile of Y, divided by the corresponding number of units obtained under Pareto sampling. Note that the Pareto design possesses a natural over-representation feature of its own: units with larger X values are expected to possess larger Y values, because of the strong positive correlation existing between X and Y, and since under classic PAR the π_ks are exactly proportional to the x_k values, we can expect, on average, to undersample units with a low y_k, and oversample units with a large y_k. It appears clear from the plot that SRS tends to oversample (respectively, undersample) units with low (high) Y values with respect to PAR, given the complete absence of targeting - the inclusion probabilities are all equal. On the other hand, HD1-3 manage to always deliver more units with Y values beyond $y_{0.5}$ (the arbitrary targeting threshold for this scenario), up to 15% more on the right tail when $c = 15$, than PAR does. Values below the threshold are under-represented by HD1-3 in the final sample as compared to PAR, with approximately the same magnitude as the over-representation. Finally, the magnitude of the imbalance seems to be proportional to c: this provides us with a way to control the amount of desired over- and under-representation, given our simple targeting function.

5.2 Spatial example

The map in Figure 3 depicts the topsoil lead concentration in mg/kg (ppm) in a flood plain of the river Meuse in the Netherlands. The population for this scenario is composed by N=3103 $40m \times 40m$ quadrats for which the lead concentration levels (variable Y) have been computed via IDW interpolation of 155 soil samples data, as described in [4], p. 216 (the original dataset was introduced in the literature in [7]). Existing literature on the dataset describes the process governing the distribution of heavy metals as being driven by polluted sediments carried by the river, and deposited close to the bank in areas of low elevation; the spatial distribution of lead concentration can be seen to be present a markedly clustered pattern. For the present example we make use, as auxiliary information, of the geographical location of the centroids of the quadrats only, described in terms of Easting X_1 and Northing X_2 (in metres) in Rijksdriehoek map coordinates (the coordinate system in use at the national level in the Netherlands). We choose to ignore the existing data on elevation to investigate the potential of our proposal in a situation where no other information than the locations of the population units is available.

In this simulation, we extracted 10000 samples of size 250 with the new method ($n_1 = n_2 = 125$) for different values of c, and stored the Monte Carlo distribution the modified HT estimator for the mean level (137.8221 mg/kg) of lead concentration in the region and the corresponding oversampling evidence; the latter aspect will be presented and discussed via maps of the Monte Carlo estimates of the quadrats inclusion probabilities.

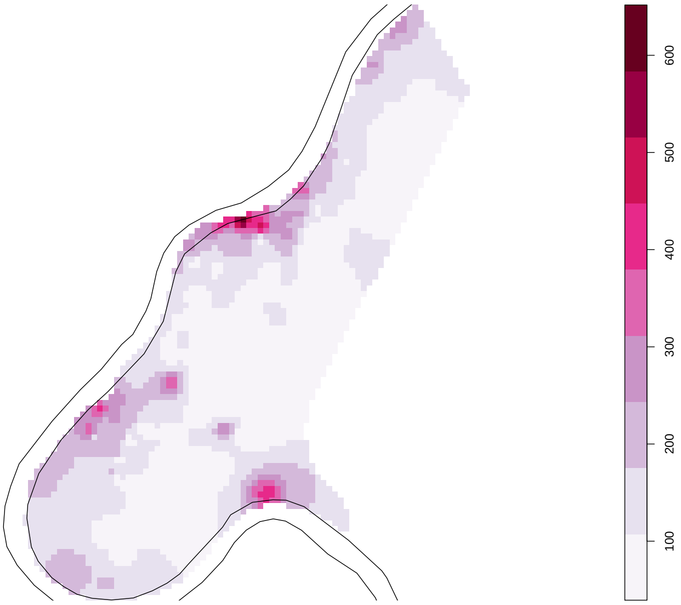

Fig. 3 Spatial example: interpolated lead concentration from the Meuse dataset

The targeted units, as in the previous example, are those with Y values in the set $\mathcal{T} = \{y \in \mathbb{R}^+ : y > y_{0.5}\}$; we include results from SRS as benchmark. Figure 4 contains the results concerning estimation: $B = 1000$ resampling runs have been used to to obtain the estimates $\widehat{\pi}_k$ of the inclusion probabilities for each Monte Carlo run (equivalently, for each of the 10000 samples extracted) and each value of c via Algorithm 2. The estimates have been, in turn, used to obtain the modified HT estimates of the mean. The percentage relative bias with respect to the true mean of Y (represented by the dashed horizontal line) are reported below each boxplot. Also in this case, the modified HT estimators exhibit some bias, albeit small; possibly, this is an indication that the reconstruction of the population using a distance-based method, such as distance HD, has been a good choice in this spatial setting, allowing for an adequate recovery of the spatial relationship between X and Y.

Figure 5 reports the maps of the Monte Carlo proportion of inclusions of each quadrat in the final sample $\mathbf{s}$, for SRS and HD1-3. For purely aesthetic reasons, a mild spatial smoothing with Gaussian kernel has been applied to the Monte Carlo empirical values. SRS provides the expected benchmark of uniform selection of units over the region, the negligible departures being imputable to

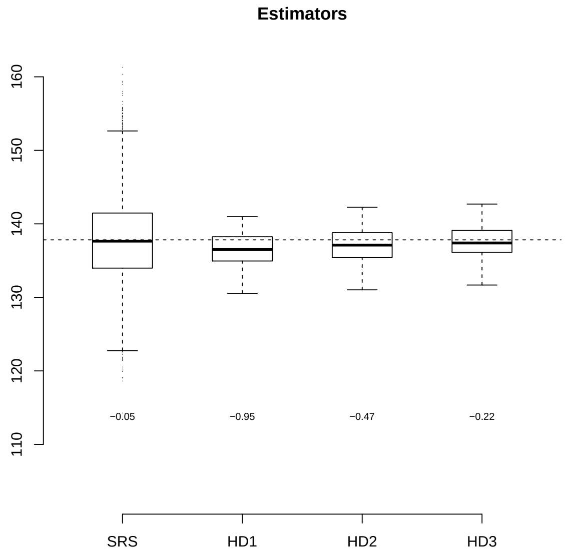

Fig. 4 Spatial example results: Monte Carlo distributions of the estimator of the mean under SRS and of the modified HT estimators of the mean under HD, for $c \in \{2, 7, 15\}$ (HD1-3), percentage relative bias shown under each boxplot; the total number of runs is 10000 and $B = 1000$

Monte Carlo error alone. The oversampling induced by HD1-3 is immediately evident and appears to become more marked as c increases; if compared with the original population map in Figure 3, these depictions indicate that the new method is indeed targeting, on average, areas where the lead concentration is larger.

6 Conclusions

Adaptive sampling methods have gained popularity in the environmental setting in recent years and have been applied to a variety of situations where the over-representation of units responding to prescribed characteristics in a sample is of interest. In this paper we proposed a novel fixed-size design-based probabilistic approach to sampling that aids the task of targeting specific subsets of a finite population in a controlled way; specifically, we introduced a two-steps sampling design that makes use of the information obtained at the first step to inform the sampling effort in the second, via predictive techniques. Our method offers ease of implementation and great flexibility in the choice of its building blocks: indeed, the structure of the design is very general and sampling procedures other than those we used (SRS at the first step and Pareto sampling at the second step) can be employed. Similarly, there is flexibility on

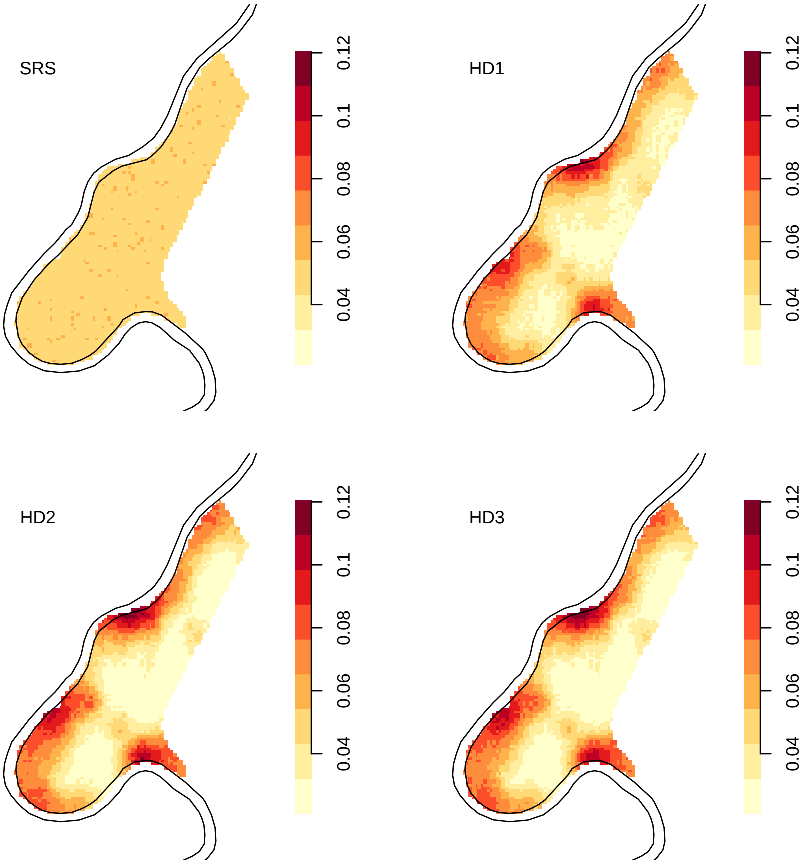

Fig. 5 Spatial example results: quadrat-specific Monte Carlo inclusion probabilities for SRS and HD, $c \in \{2, 4, 15\}$ (HD1-3), darker areas are more frequently selected in the final sample; the total number of runs is 10000.

how: i) the predicted survey variable $\widehat{Y}$ is constructed, and ii) the predicted values are transformed into inclusion probabilities for the second step. Moreover, the method is suited to envisage more than two steps: it is immediate to extend the algorithm beyond two sampling occasions, which may be helpful in improving the targeting. Clearly, all these choices will influence the performances of the method, to an extent that is still under study.

We addressed the problem of estimating a parameter, the population mean, and proposed a numerical procedure to overcome the impossibility of obtaining the true inclusion probabilities under the new design. We conjectured that, in front of an adequate reconstruction of the original population based on the sample evidence, the resampling approach we discussed should provide a good approximation of the π_ks to, in turn, used with a modified version of the HT estimator. We also showed that the extension of the new design to spatial applications is straightforward; indeed, the only adaptation required is to include the geographical coordinates of the sampling units in the auxiliary information. Moreover, we discussed how some recent results on consistent map reconstruction based on probabilistic samples may help improve the predictive part, needed to inform the second step sampling. The flexibility of the method also allows for more complex targeting functions than the one considered here: for example, the researcher may be interested in targeting units whose surroundings present larger-than-average variability with respect to the survey variable; this would only require to adjust the targeting function accordingly. The simulation results indicate that our proposal is successful in delivering over-representation of units responding to prescribed characteristics, if compared with SRS and classic π-ps sampling designs, such as Pareto. The complete control on final sample size and on the the magnitude of the over-representation are desirable features that are currently not provided by competing adaptive designs such as ACS and AGD; further research is needed, however, to directly compare the new method with these competing approaches. Finally, an aspect that has received much attention in the survey literature recently is that of spatial balance; we reckon that by suitably combining design choices and targeting functions, features of spatial balance with respect to general (not only geographical) auxiliary information, in the sense described in, for example, [16], can be achieved while still retaining oversampling (although some trade-off is clearly expected).

Acknowledgements The idea of using Hot Deck imputation to construct a resampling population as described in Section 4 originates from unpublished work on bootstrapping non-iid samples developed in collaboration among Fulvia Mecatti, Pierluigi Conti and the first author. A draft version of that work, which will appear elsewhere, is available at https://arxiv.org/abs/1705.03827v2.
The authors also wish to thank two anonymous referees for their precious comments, that helped improve the quality of this work.

References

1. Andreis F, Furfaro E and Mecatti F (2017). Methodological perspectives for surveying rare and clustered population: towards a sequentially adaptive approach. Studies in Theoretical and Applied Statistics (Eds. Cira Perna, Monica Pratesi, Anne Ruiz-Gazen). Springer.
2. Andridge RR, and Little RJA (2010). A Review of Hot Deck Imputation for Survey Non-response. International Statistical Review. 78, 40–64.
3. Baddeley A, Rubak E and Turner R (2015). Spatial Point Patterns: Methodology and Applications with R. London: Chapman and Hall/CRC Press, 2015. http://www.crcpress.com/Spatial-Point-Patterns-Methodology-and-Applications-with-R/Baddeley-Rubak-Turner/9781482210200/.
4. Bivand RS, Pebesma E and Gómez-Rubio V (2013). Applied Spatial Data Analysis with R. Second edition. UseR! Series, Springer.
5. Brown JA and Manly BFJ (2016). Restricted adaptive cluster sampling. Environmental and Ecological Statistics, 5: 47-62.
6. Bruno F, Cocchi D and Vagheggini A (2013). Finite population properties of individual predictors based on spatial pattern. Environmental and Ecological Statistics, 20:457-494.

7. Burrough PA and McDonnell RA (1998). Principles of Geographical Information Systems. Oxford University Press.
8. Chipeta MG, Terlouw DJ, Phiri KS and Diggle PJ (2016). Adaptive geostatistical design and analysis for prevalence suveys. Spatial Statistics, 15, 70-84.
9. Di Battista T (2003). Resampling methods for estimating dispersion indices in random and adaptive designs. Environmental and Ecological Statistics, 10 (1), 83-93.
10. Fattorini L (2006). Applying the Horvitz-Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. Biometrika, 93 (2), 269–278.
11. Fattorini L, Corona P, Chirici G and Pagliarella MC (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. doi:10.1002/env.2332
12. Fattorini L, Marcheselli M and Pratelli L (2017). Design-Based Maps for Finite Populations of Spatial Units. Journal of the American Statistical Association, doi: 10.1080/01621459.2016.1278174.
13. Gattone S and Di Battista T (2011). Adaptive cluster sampling with a data driven stopping rule. Statistical Methods and Applications, 20 (1), 1-21.
14. Gattone S, Mohamed E and Di Battista T (2016). Adaptive cluster sampling with clusters selected without replacement and stopping rule. Environmental and Ecological Statistics, 23: 453-468.
15. Gräler B, Pebesma EJ and Heuvelink G (2016). Spatio-Temporal Interpolation using gstat. The R Journal 8(1), 204-218.
16. Grafström A, Saarela S and Ene LT (2014). Efficient sampling strategies for forest inventories by spreading the sample in auxiliary space. Canadian Journal of Forest Research 44 (10), 1156-1164.
17. Hájek J (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. The Annals of Mathematical Statistics, 35, 4, 1491–1523.
18. Hájek J (1981). Sampling from a finite population. Statistics: Textbooks and Monographs 37. Marcel Dekker Inc., New York. Edited by Václav Dupač, With a foreword by P. K. Sen. ISBN: 0-8247-1291-9.
19. Joenssen DW (2015). HotDeckImputation: Hot Deck Imputation Methods for Missing Data. R package version 1.1.0. https://CRAN.R-project.org/package=HotDeckImputation.
20. Kabaghe AN, Chipeta MG, McCann RS, Phiri KS, van Vugt M, Takken W, Diggle P and Terlouw AD (2017). Adaptive geostatistical sampling enables efficient identification of malaria hotspots in repeated cross-sectional surveys in rural Malawi. PLoS ONE 12(2):e0172266. doi:10.1371/journal.pone.0172266
21. Marella, D, Scanu, M, and Conti, PL (2008). On the matching noise of some nonparametric imputation procedures. Statistics and Probability Letters, 78, 15931600.
22. Pacifici K, Reich BJ, Dorazio RM and Conroy MJ (2016). Occupancy estimation for rare species using a spatially-adaptive sampling design. Methods in Ecology and Evolution, 7, 285–293. doi:10.1111/2041-210X.12499.
23. Pebesma EJ and Bivand RS (2005). Classes and methods for spatial data in R. R News 5 (2), https://cran.r-project.org/doc/Rnews/.
24. R Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, url: http://www.R-project.org.
25. Rosen B (1997). On sampling with probability proportional to size. Journal of Statistical Planning and Inference 62, 159-191.
26. Salehi MM and Seber GAF (2017). Two-stage complete allocation sampling. Environmetrics.
27. Salehi MM, Moradi M, Al Khayat JA, Brown . and Yousif AEM (2015). Inverse Adaptive Cluster Sampling with Unequal Selection Probabilities: Case Studies on Crab Holes and Arsenic Pollution. Aust. N. Z. J. Stat., 57: 189–201. doi:10.1111/anzs.12118.
28. Seber GAF and Salehi MM (2013). Adaptive Sampling Designs: Inference for Sparse and Clustered Populations. Heidelberg: Springer.
29. Seber GA and Thompson SK (1994). 6 Environmental adaptive sampling. Handbook of statistics. Dec 31;12:201-20.
30. Thompson SK (1990). Adaptive Cluster Sampling. Journal of the American Statistical Association, 85 (412), 1050-1059.

31. Yan J (2007). Enjoy the Joy of Copulas: With a Package copula. Journal of Statistical Software, 21(4), 1-21. http://www.jstatsoft.org/v21/i04/.