New perspectives on sampling rare and clustered populations

Emanuela Furfaro [*] Fulvia Mecatti [†]

Abstract

A new sampling design is derived for sampling a rare and clustered population under both cost and logistic constraints. It is motivated based on the example of national TB prevalence surveys, sponsored by WHO for high TB-burden countries and usually located in the poorest parts of the world. A Poisson-type sampling design named Poisson Sequential Adaptive (PoSA) is proposed with a twofold purpose: *(i)* to increase the detection rate of positive cases; and *(ii)* to reduce survey costs by accounting for logistic constraints at the design level of the survey. PoSA is derived by integrating both an adaptive component able to enhace detectability and a sequential component for dealing with costs and logistic constraints. An unbiased HT-type estimator for the population prevalence (mean) is derived by adjusting for both the over-selection bias and for the conditional structure induced by the sequential selection. Unbiased variance estimation in a closed form is also provided. Simulation results are presented and show a significant pontential of PoSA in improving the sampling methodology currently suggested by WHO guidelines.

Key Words: low prevalence surveys, adaptive design, sequential sampling, poisson sampling design

1. Introduction and motivation

Sampling a rare and clustered trait over a finite population may be challenging. In fact when using traditional sampling designs large sample sizes are needed for reasonably accurate estimation and still many cases may be missed.

This paper is inspired by the challenges the World Health Organisation (WHO) faces when carrying out tubercoulosis (TB) prevalence surveys. TB prevalence surveys are performed in those countries considered to bear a high burden of TB [3]. Although considered high TB-burden countries, the number of TB positives ranges around 150-300 per 100000 individuals. As TB is an infectious disease, the cases are expected to be clustered, configuring a sampling situation where the population of interest is rare and clustered. Moreover, in this setting, the samplers aim not only at correctly estimating the overall prevalence but also at finding ideally the largest number of cases (i.e. oversample cases) as every found case could be a treated one. The sampling strategy currently suggested in the latest WHO guidelines [13] is quite traditional: primary sampling units are geographical areas of (approximately) equal size in terms of number of eligible population. The number of areas to be sampled is chosen according to a fixed sample size computed as a function of *(i)* a prior guess of the true population prevalence, *(ii)* a chosen level of precision of the final estimate (usually around 25% relative to the true value), and *(iii)* an estimate of the variability existing between the areas' prevalences, measured in terms of coefficient of variation. A moving lab reaches the selected areas and all eligible individuals (i.e. people aged ≥ 15) are invited to undertake the medical examination. The

[*] emanuela.furfaro@unimib.it, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milano, Italy

[†] fulvia.mecatti@unimib.it, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milano, Italy

classic Horvitz-Thompson approach is then employed to estimate the population prevalence.
This is essentially an Unequal Probability Cluster Sampling (UPCS) considered sufficiently easy to implement and understand to fit general guidelines. However limitations can be pointed out. The rarity of TB positives and their uneven distribution over the inspected areas lead to the need for a very large sample size to obtain an accurate estimate of the true prevalence (the final sample size usually ranges between 30000 and 100000 individuals). Prior information on between areas variability is accounted for in the sample size computation usually leading to an increase as the between-areas variation is higher. We believe that the between areas variability might be exploited for concentrating surveying efforts in areas where spotting a case is more likely. Moreover, when the countries involved in the survey are developing countries, there may be logistic constraints due to reduced accessibility of some areas and the large survey costs may not allow for much flexibility in the final sample size. Therefore there seems to be room for improvement of the currently suggested sampling methodology especially with regards to *(i)* the number of detected cases and *(ii)* the managing of logistic constraints and costs. The purpose of this paper is to develop an improved sampling strategy with respect to both points *(i)* and *(ii)* above.

2. Steps towards an improved sampling design

A focus on case-detection may be achieved by using adaptive designs. They were originally introduced by Thompson ([10]) and are a popular tool for dealing with the problem of estimating the prevalence of a rare and clustered trait (see [11] for a review). The distinctive characteristic of adaptive designs is to make use of information collected while sampling and adaptively adjust the procedure, on the bases of a pre-defined distance measure, for oversampling in the proximity of detected cases. If the investigated trait is clustered, the number of detected cases is expected to be larger than under a traditional non-adaptive sampling design. Thus an Adaptive Cluster Sampling design (ACS in the following) seems a natural choice for enhancing detectability in the context of national TB prevalence surveys. However a disadvantage of this method is that the final sample size would be in fact a random number, thus survey costs may be hard to control and logistic constraints not considered.
On the other hand, a focus on managing costs and logistic constraints while dealing with spatial correlation would suggest to adopt a sequential sampling design based on a pre-ordered population ([2] and [6]). For instance in planning a TB survey in a Sub-Saharian African country, logistic constraints and costs control would suggest ordering areas along a specific route and then conditioning the sample selection to be sequential along such prescribed route. As opposed to adaptive designs the sequential approach offers some way to control the final sample size ([2]) and can naturally accomodate for a predefined route, but it does not allow, in its current formulation, for over-sampling positive cases. Adaptive designs and the sequential approach seem to individually overcome different limitations as discussed above when surveying a rare and clustered population, which suggests the perspective of an integrated strategy.

3. The proposed methodology: Poisson Sequential Adaptive design

We still refer to our motivational example of national TB prevalence survey. We start from the design currently suggested in the WHO guidelines and we propose a sampling design comprising both a sequential component and an adaptive component. As a first proposal and for the sake of its simplicity we consider a Poisson-type design [9] which we will name Poisson Sequential Adaptive Sampling Design, PoSA for short. Notice that a known feature of the Poisson design is the random sample size which may be a drawback with respect to survey cost planning. A proposal for controlling the final sample size in a PoSA design will be illustrated in section 4.

In the most recent WHO guidelines for TB prevalence surveys it is suggested to first select a sample of areas from a grid of M covering the national territory, and successively to collect data from all eligible individuals included in each selected area. The M areas have to be formed under the constraint of equal/very close size in terms of population eligible for the surveys. In our proposal such equal size constraint is relaxed in favour of the choice of a *sequence* of the M areas (anyhow formed), for instance as already mentioned, formed by following a specific route across the country pre-defined according to costs and logistic constraints. The sequential component of the proposed sampling design consists in following the chosen sequence in selecting the areas to be included in the final sample. At the j-th step of the sequential selection, area j is/is not selected in the sample according to a chosen adaptive rule aiming at enhancing the case-detection rate. For instance in a TB survey suppose that, at a given step of the sequential selection, a significant number of TB cases is collected, say greater/much greater than the expected national prevalence. This is assumed as an indication that further TB cases should be present in the neighbourhood, i.e. a TB cluster has been detected. As a consequence data collection would proceed by certainly selecting the subsequent area in the sample. More formally, let y_{kj} be the survey value of the k-th individual included into area j, e.g. $y_{kj} = 1$ if k is a TB-positive case and 0 otherwise. Thus the prevalence of area j is $\bar{y}_j = \sum_{k \in j} y_{kj}/N_j$ where N_j denotes the area (population) size. Finally let c, $(0 < c < 1)$ be a threshold chosen such that the following (adaptive) condition qualifies j as an area with a *significant* number of TB cases

$$\bar{y}_j = \frac{1}{N_j} \sum_{k=1}^{N_j} y_{kj} \geq c \qquad (1)$$

A set of inclusion probabilities π_j is given for the sequence of all areas $j = 1, \ldots, M$, for instance proportional to their (possibly unequal) size N_j. At the j-th step of the sequential selection, area j is certainly selected (i.e. is selected with probability 1) if the adaptive condition (1) holds for the previous area $j - 1$, otherwise it is selected with probability π_j. The proposed PoSA sampling procedure can be synthetized in a ready-to-implement set of instructions. In Algorithm 1 the usual sample membership indicator of area j is denoted by S_j, i.e. the random variable taking value 1 if area j is included in the sample and 0 otherwise. At each step of the sequential selection, S_j is updated adaptively by means of a further indicator y_j taking value 1 if the adaptive condition (1) holds in area j, and 0 otherwise.

Algorithm 1 PoSA Algorithm

```
procedure
Input: Ordered sequence of $M$ areas with initial inclusion probabilities $\pi_j$.
Output: A sample of random size.
    Visit unit $j = 1$ and select with probability $\pi_1$.
for j in 2 : $M$
    1. point unit $j$ & select with probability $\pi_j$ if $y_{j-1}S_{j-1} = 0$
    2. point unit $j$ & select with probability 1 if $y_{j-1}S_{j-1} = 1$
    3. if $S_j = 1$, collect $y_j$
    Return: vector of sample membership indicators of size $M$
end procedure
```

The PoSA selection procedure is expected to lead to a sample of over-represented cases. As a matter of fact this is a purpose of the sampling design at the selection stage of the survey which must be corrected at the estimation stage in order to produce an unbiased estimate of the parameter of interest, e.g. the national TB prevalence $N^{-1}\sum_{j=1}^{M}\sum_{k=1}^{N_j} y_{kj}$, where $N = \sum_{j=1}^{M} N_j$ is the population size.
We now illustrate how an unbiased Horvitz-Thompson type (HT) estimator can be derived under the PoSA sampling design. The main point is that the sequential feature of the design induces a conditional structure over each pair of subsequent sample membership indicator S_{j-1} and S_j which has to be accounted for according to the adaptive feature formallly represented by the indicator y_j. Hence an unbiased HT estimator for the population prevalence (mean) under PoSA sampling design has the following form:

$$\hat{Y}_{PoSA} = \frac{1}{N}\sum_{j=1}^{M}\sum_{k=1}^{N_j} y_{kj}\frac{S_j|S_{j-1}}{E(S_j|S_{j-1})} = \frac{1}{N}\sum_{j=1}^{M} N_j\bar{y}_j\frac{S_j|S_{j-1}}{E(S_j|S_{j-1})} \tag{2}$$

The conditional sample membership indicator for area j has the following form: for $j = 1$, S_1 is a Bernoulli random variable with parameter π_1; and for $j = 2, \ldots, M$

$$S_j|S_{j-1} = Bernoulli(\pi_j)(1 - S_{j-1}y_{j-1}) + S_{j-1}y_{j-1} \tag{3}$$

with expectation given by the following recoursive formula

$$E(S_1) = E(S_1|S_0) = \pi_1; \text{and for } \ j = 2, \ldots, M$$

$$E(S_j|S_{j-1}) = \pi_j - \pi_j E(S_{j-1}|S_{j-2})y_{j-1} + E(S_{j-1}|S_{j-2})y_{j-1} \tag{4}$$

Furthermore, it is easily proved that for every pair (j, i) such that $j = 2, \ldots, M$ and $i < j$ we have:

$$Cov[(S_j|S_{j-1}), (S_i|S_{i-1})] = \begin{cases} E(S_1)\ y_1\ [1 - E(S_2|S_1)] & \text{if } j = 2 \\ E(S_i|S_{i-1})y_{j-1}[1 - E(S_j|S_{j-1})] & \text{if } j > 2, i = j-1 \\ 0 & \text{if } j > 2, i < j-1 \end{cases} \tag{5}$$

which allows for deriving the exact variance of estimator $\hat{Y}_{PoSA}$ as given by

$$V(\hat{Y}_{PoSA}) = \frac{1}{N^2}\left[\sum_{j=1}^{M}(N_j\bar{y}_j)^2\,\frac{1-E(S_j|S_{j-1})}{E(S_j|S_{j-1})}\right.$$
$$\left. +2\sum_{j=2}^{M}\sum_{i<j}N_j\bar{y}_j\;N_i\bar{y}_i\;Cov[(S_j|S_{j-1}),(S_i|S_{i-1})]\right] \qquad (6)$$

Finally, let s be the sample of areas selected under the $PoSA$ design; an unbiased variance estimator readily follows from equations 5 and 6 as given by:

$$v(\hat{Y}_{PoSA}) = \frac{1}{N^2}\left[\sum_{j\in s}(N_j\bar{y}_j)^2\frac{1-E(S_j|S_{j-1})}{E(S_j|S_{j-1})^2}\right.$$
$$\left. +2\sum_{j\in s}\sum_{i<j\in s}N_j\bar{y}_j\;N_i\bar{y}_i\;\frac{Cov[(S_j|S_{j-1}),(S_i|S_{i-1})]}{E[(S_j|S_{j-1})(S_i|S_{i-1})]}\right] \qquad (7)$$

Being Poisson sampling-based, PoSA sampling design leads to a random number of selected areas and hence to an unplanned final sample size. In the following section a proposal will be discussed for controlling the final sample size whithout the loss of the simplicity inherent to an (unconditional) Poisson design.

4. Controlling the final sample size

An appreciable feature of the customary (unconditional) Poisson sampling design is its simplicity due to independence between unit selections [9] . In the proposed PoSA design such a feature leads to exactly unbiased formulae for point and variance estimation in a simple closed form despite the conditional structure between subsequent sample membership indicators. The need for planning the final sample size at the design level of the survey and formal easiness of PoSA estimation as illustrated in the previous section, are actually conflicting objectives. A reasonable compromise for sample size controlling may be that of fixing an upper bound n, (integer, > 1) that must not be exceeded while performing a PoSA selection. In other words, once the maximum sample size n is attained, the sampling procedure simply stops. Algorithm 2 illustrates how to implement a Poisson Sequential Adaptive design with maximum sample size n ($PoSA_n$). The procedure is analogous to that listed in Algorithm 1 with the additional definition of an indicator $\Lambda_j,\ j = 1,\ldots,M$. At each step j, Λ_j takes value 1 if the pre-stated maximum sample size n has not been attained yet, so that sampling selection should proceed at step j; it equals 0 if the pre-stated sample size n has been reached at (any of) the previous step(s) so that the sampling procedure must be stopped.

Algorithm 2 $PoSA_n$ algorithm

```
procedure
  Input: Ordered sequence of $M$ areas with initial inclusion probabilities $\pi_j$ and the sample size $n$ that does not have to be exceeded.
  Output: A sample of size $\leq n$.

    Visit unit $j = 1$ and select with probability $\pi_1$, set $\Lambda_1 = 1$.

     for j in $2 : M$

    define $\Lambda_j = \begin{cases} 1 & \text{if } \sum_{i=1}^{j-1} S_i|S_{i-1} < n \\ 0 & \text{otherwise} \end{cases}$
     while $\Lambda_j = 1$
    1. point unit $j$ & select with probability $\pi_j$ if $y_{j-1}S_{j-1} = 0$
    2. point unit $j$ & select with probability 1 if $y_{j-1}S_{j-1} = 1$
    3. if $S_j = 1$, collect $y_j$

    Return: vector of sample membership indicators of size $M$
end procedure
```

Equations 2-7 shall be adjusted with the introduction of the stopping rule Λ_j, which affects the inclusion probabilities of all areas $j > n$. An unbiased HT-type estimator for the population prevalence (mean) replacing equation 2, is:

$$\hat{Y}_{PoSA_n} = \frac{1}{N} \sum_{j=1}^{M} \sum_{k=1}^{N_j} y_{kj} \frac{S_j|(S_{j-1}, \Lambda_j)}{E(S_j|(S_{j-1}, \Lambda_j))} \tag{8}$$

where the conditional sample membership indicator for area j has the following form

$$S_j|(S_{j-1}, \Lambda_j) = \Lambda_j \left[Bernoulli(\pi_j)(1 - S_{j-1}y_{j-1}) + S_{j-1}y_{j-1} \right] \tag{9}$$

and its expectation is given by the following recoursive formula:

$$\text{for } j = 1 \ E(S_1) = E(S_1|S_0, \Lambda_1) = \pi_1; \text{and for } j = 2, \ldots, M$$

$$E(S_j|S_{j-1}, \Lambda_j) = \begin{cases} \pi_j & \text{if } S_{j-1} = 0 \cup \{S_{j-1} = 1 \cap y_{j-1} = 0\} \cap j \leq n \\ 1 & \text{if } S_{j-1} = 1 \cap \{S_{j-1} = 1 \cap y_{j-1} = 1\} \cap j \leq n \\ w_j\pi_j & \text{if } S_{j-1} = 0 \cup \{S_{j-1} = 1 \cap y_{j-1} = 0\} \cap \Lambda_j = 1 \cap j > n \\ w_j & \text{if } S_{j-1} = 1 \cup \{S_{j-1} = 1 \cap y_{j-1} = 1\} \cap \Lambda_j = 1 \cap j > n \end{cases}$$

where

$$w_j = 1 - \Pi_{i \in A_j} P(S_i = 1|S_{i-1}) \Pi_{i \in \bar{A}_j} P(S_i = 0|S_{i-1})$$

and A_j is an index set including areas selected at any of the previous steps $1, \ldots, j-1$ while $\bar{A}_j$ is the complement set referring to the un-selected areas.

Finally for variance estimation purposes, notice that, for every pair (j, i) such that $j = 2, \ldots, M$ and $i < j$, equation 5 shall be replaced by:

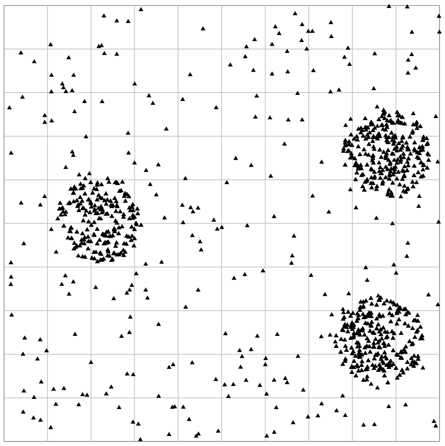

Figure 1: Simulated population of $N = 100000$ individuals unevenly spread with three clusters of positive cases (black triangles) and ≈ 0.015 true population prevalence

$$Cov[(S_j|S_{j-1},\Lambda_j),(S_i|S_{i-1},\Lambda_i)] =$$

$$= \begin{cases} E(S_1)\ y_1\ [1 - E(S_2|S_1,\Lambda_2)] & \text{if } j = 2 \cap \Lambda_2 = 1 \\ E(S_i|S_{i-1},\Lambda_i)y_{j-1}[1 - E(S_j|S_{j-1},\Lambda_j)] & \text{if } j > 2, i = j-1 \cap \Lambda_j = 1 \\ 0 & \text{if } j > 2, i < j-1 \cup \Lambda_j = 0 \end{cases}$$

5. Some empirical evidence

In this Section some significant results are presented from a simulation study aiming at evaluating the performance of the proposed PoSA sampling design. Simulations focus on the key features discussed in the previous sections, namely the over-sampling of positive cases when surveying a rare and clustered trait under both cost and logistic constraints, as it is the case for our inspirational example of a national TB prevalence survey. The proposed PoSA design has been empirically compared with *i)* a purely Adaptive Cluster Sampling (ACS); *ii)* a purely sequential design known as Spatially Correlated Poisson Sampling (SCPS with maximal weights strategy; see [4] and [5] for further details): and *iii)* a traditional Unequal Probability Cluster design (UPCS) as currently suggested in WHO guidelines. Figure 1 depicts the simulated population composed by $N = 100000$ individuals evenly spread over a two-dimensional space, with three clusters of cases (for instance TB positive) and ≈ 0.015 true population prevalence. The overimposed 10×10 grid generates a set of $M = 100$ areas to be sampled.

According to WHO guidelines we assumed to have a good guess of the true prevalence (0.01) and a 0.5 coefficient of between areas' prevalences variation. Under the suggested UPCS, a sample of 23 areas is selected with 100% participation rate within selected areas (which is a best-scenario approximation of the actual participation rate for TB prevalence surveys, usually in the range $85 - 90\%$ [13]). For all the compared designs, the total survey cost has been computed based on a linear cost function including *(i)* a fixed cost, for instance for equipment and staff; *(ii)* a unitary cost for each selected area, for instance for transportation and installation of the moving lab in the selected location; and *(iii)* a unitary cost for each individual data collected in every selected area. Under both the purely sequential design SCPS and the proposed PoSA design, a 30% reduction of the unitary cost for selected area has been considered as allowed by planning the route (area sequence) in advance. Under both purely adaptive design ACS and the proposed PoSA design the threshold $c = 0.01$ has been set for the adaptive condition (1).

Tables 1- 3 show elementary statistics (quartiles and avarage), over 5000 Monte Carlo runs, with a focus on the selection stage of the survey, namely final sample size, number of cases detected and the total cost. Under these respects, empirical results indicate that the proposed $PoSA$ sampling design uniformly improves the WHO's guidelines suggested design (UPCS); notice that PoSA shows an intermediate performance between ACS and SCPS as the result of integrating both a sequential and an adaptive component.

The last column of each table shows the results for $PoSA_n$ (see section 4) with the same UPCS choice $n = 23$. As compared to the $PoSA$, $PoSA_n$ appears to succeed in reducing survey costs by better controlling the final sample size. However its case-detection ability, though still improved with respect to UPCS, can decrease significantly.

Table 1: Final sample size: elementary MC stats

	UPCS	ACS	SCPS	PoSA	$PoSA_n$
0.25 percentile	22916	32352	22941	26485	22875
Median	23029	35486	23041	30501	23001
Average	23034	34633	23041	30525	22784
0.75 percentile	23144	37662	23143	34382	23129

Table 2: Number (%) of detected cases (total number of cases=1518): elementary MC stats

	UPCS	ACS	SCPS	PoSA	$PoSA_n$
0.25 percentile	291 ($\approx$ 19%)	797 ($\approx$ 0.53%)	313 ($\approx$ 20%)	500 ($\approx$ 33%)	402 ($\approx$ 26%)
Median	357 ($\approx$ 24%)	1121 ($\approx$ 74%)	364 ($\approx$ 24%)	660 ($\approx$ 44%)	494 ($\approx$ 33%)
Average	364 ($\approx$ 24%)	953 ($\approx$ 63%)	365 ($\approx$ 27%)	654 ($\approx$ 43%)	488 ($\approx$ 32%)
0.75 percentile.	432 ($\approx$ 28%)	1141 ($\approx$ 75%)	416 ($\approx$ 28%)	810 ($\approx$ 53%)	574.0 ($\approx$ 38%)

Table 3: Total survey costs: elementary MC stats ($100000 of fixed cost, $ 1000 unitary cost per area, $ 20 unitary cost per individual)

	UPCS	ACS	SCPS	PoSA	$PoSA_n$
0.25 percentile	352160	466248	340910	377848	340250
Median	353290	504755	341910	420005	341510
Average	353338	493773	341907	420412	339215
0.75 percentile	354440	529923	342930	460815	342790

Figure 2 shows simulation results with regards to the estimation stage. The Monte Carlo distribution of the HT-type estimator under each of the simulated sampling design is summarized via boxplots. Although all unbiased, the compared sampling strategies can differ with respect to variability, i.e. stability/efficiency. Simulation results confirm that ACS estimator may be heavily asymmetric due to its tendency to extreme over/under cases detection. They also show that a purely sequential design, though less able to meet the over-detection objective, can be more efficient among the compared designs. The proposed PoSA design (here limited to its original version with random sample size) shows to be able to maintain a comparable estimator efficiency while both enhancing cases detectability and costs/logistics management. Also notice that, despite its adaptive component, PoSA appears to provide a symmetric estimator's distribution, which is appreciable for constructing confidence intervals.

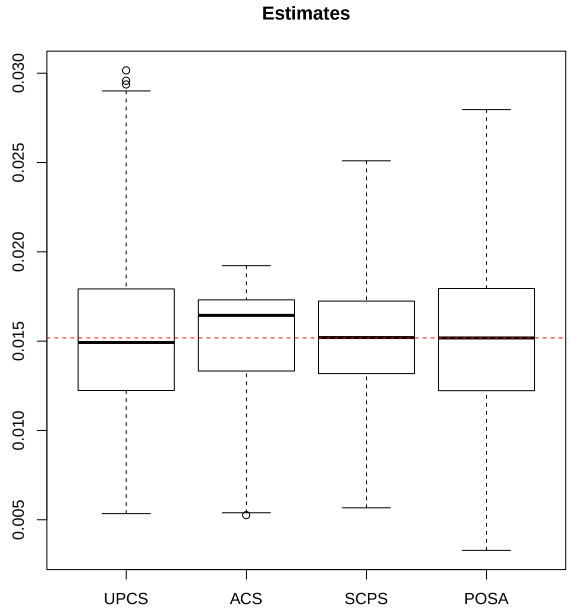

Figure 2: MC distribution of estimates for the true population prevalence ≈ 0.015 over 5000 runs.

6. Conclusions and research perspectives

In this work a new sampling design is proposed for sampling a rare and clustered population under both cost and logistic constraints. It is motivated based on the example of national TB prevalence surveys promoted by WHO for high TB-burden countries primarily concentrated in the poorest areas of the world. We proposed a Poisson-type sampling design named Poisson Sequential Adaptive ($PoSA$) with the two main purposes of *i)* increasing the detection rate of positive cases; and *ii)* reducing survey costs by accounting for logistic constraints at the design level of the survey. PoSA has been derived by integrating both an adaptive component able to enhace cases detectability and a sequential component for dealing with costs and logistic constraints. An unbiased HT-type estimator for the population prevalence (mean) is derived based on adjusting for both the over-selection bias and for the conditional structure induced by the sequential selection. A slightly modified version of PoSA for a chosen maximum sample size is also illustrated. Simulaton results show a significant pontential of $PoSA$ in improving the methodology currently suggested by WHO guidelines with respect to both case-detection and costs controlling. Interesting perspectives for future research are also opened by empirical evidence. Particularly the control over the final sample size needs to be further investigated, as well as the effect of the population ordering defining the sequentiality of selection. Simulation results confirm that the ordering choice coupled with the choice of a maximum sample size affects selection probabilities which tend to decrease as the selection proceed thus interacting with the adaptive feature. The effect of more general non linear cost functions tailored for different sources of expenditure and budgeting flexibility appears worth investigating. Room for improving the sequential component of PoSA is offered by fully exploiting the potential of a probability updating system. In fact, instead of having areas with inclusion probability either (initial) π_j or (updated) 1, they could be finely tuned for example by considering costs or any other available area info. The availability of auxiliary variable and other meta-data will also be considered for improving PoSA estimation via regression.

Aknowledgments

The authors are grateful to Kristen Sauby and Mary Christman for providing their R functions for generating ACS procedure and the ACS estimator.

References

[1] Baddeley, A. and Turner, R.: spatstat: An R Package for Analyzing Spatial Point Patterns. Journal of Statistical Software 12(6), 1-42. (2005)

[2] Bondesson, L., Thorburn, D.: A list sequential sampling method suitable for real-time sampling. Scandinavian Journal of Statistics. **35**, 466–483 (2008)

[3] Glaziou, P., van der Werf, M. J., Onozaki, I., Dye, C.:Tuberculosis prevalence surveys: rationale and cost. International Tuberculosis Lung Disease. **12(9)**, 1003–1008 (2008)

[4] Grafström, A.: On a generalization of Poisson sampling. Journal of Statistical Planning and Inference. **140**, 4, 982–991 (2010)

[5] Grafström, A., Lundström, N.L.P. and Schellin, L.: Spatially Balanced Sampling through the Pivotal Method. Biometrics. **68**, 2, 514–520 (2011)

[6] Grafström, A.: Spatially Correlated Poisson sampling. Journal of Statistical Planning and Inference. **142**, 1, 139–147 (2012)

[7] Grafström, A. and Lisic, J.: BalancedSampling: Balanced and Spatially Balanced Sampling. R package version 1.5.1. (2016)

[8] Seber, G.A.F., Salehi M.M.: Adaptive Sampling Designs. Springer, Berlin (2012)

[9] Tillé, Y.: Sampling Algorithms. Springer, USA (2006)

[10] Thompson, S.K.: Adaptive cluster sampling. Journal of the American Statistical Association. **85**, 1050–1059 (1990)

[11] Thompson, S.K., Seber, G.A.F.: Adaptive Sampling. John Wiley & Sons, Inc. (1996)

[12] Thompson, S.K.: Sampling. John Wiley & Sons, Inc., Hoboken, New Jersey (2012)

[13] The World Health Organisation: Tubercoulosis PREVALENCE SURVEYS: a handbook. WHO Press, Geneva (2011)