SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of

Statistica e metodi quantitativi

PhD program Statistics and Mathematical Finance Ciclo / Cycle XXIX

Curriculum in Statistics

A SEQUENTIAL ADAPTIVE APPROACH FOR SURVEYING RARE AND CLUSTERED POPULATIONS

Cognome / Surname Furfaro Nome / Name Emanuela

Matricola / Registration number 712167

Tutore / Tutor: Prof.ssa Fulvia Mecatti

Cotutore / Co-tutor: Dott. Federico Andreis
(se presente / if there is one)

Supervisor:
(se presente / if there is one)

Coordinatore / Coordinator: Prof. Giorgio Vittadini

ANNO ACCADEMICO / ACADEMIC YEAR 2015/2016

A sequential adaptive approach for surveying rare and clustered populations

Emanuela Furfaro

Department of Statistics and Quantitative Methods
University of Milano-Bicocca
Doctoral School in Statistics and Mathematics for Finance
Ph. D. in Statistics - XXIX Cycle

Contents

Introduction

When knowledge pertaining a rare trait of the population is of interest the collection of survey data presents various challenging aspects. In fact, in order to obtain a reasonably accurate estimate, for instance the estimation of the population prevalence, very large sample sizes are needed, thus inflating survey costs. Moreover, when cases are not only rare but also unevenly distributed throughout space, i.e. they present specific spatial patterns such as, for example, clustering, traditional sampling designs may perform poorly [22]. Moreover, in many applications additional needs should be addressed by the sampling design. For instance epidemiological and environmental data collection on field usually is prone to specific logistics and costs constraints. In addition a high detection rate of the rare trait is often desirable.

This thesis 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 [6]. In these countries, typically located in developing areas of Sub-Saharan Africa and South-Eastern Asia, the prevalence of TB is measured by means of nationwide, population-based surveys that are carried out by WHO with the support of local agencies. In this setting, an accurate estimation of the true TB prevalence is of paramount importance to be able to inform public health policies aimed at reducing the burden; moreover, due to the actual presence of medical doctors on field during these surveys, every TB case that can be found can and will be cured. Although considered high TB-burden countries, the number of TB positives ranges around 150-700 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.

The sampling strategy currently suggested in the WHO guidelines for TB prevalence studies [30] is a traditonal multi-stage sampling design which is essentially meant to be sufficiently feasible and easy to implement as requested for general guidelines. The study region is divided into smaller geographical areas of about the same size (400-1000 individuals) and once a region is selected, all resident individuals are invited to undertake the medical examination. A crucial point in the WHO guidelies is the choice of the sample size. In fact, 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. Possible information on between geographical areas variability is accounted for in the sample size determination. Specifically the required sample size increases as the between areas prevalence variability gets higher. It is well understood that in this sampling setting traditional designs, although providing unbiased estimates, require a large sample size and tend to miss cases when they are clustered. Moreover, as the countries involved are developing countries, the sampling procedure may be particularly costly and some areas may be of reduced access due to, for instance, natural barriers, unusable transportation networks, war areas, etc.

WHO's practice for TB prevalence surveys may then draw benefit by more refined sampling strategies that are able, for instance, to lead to the oversampling of cases, the sample size being equal, and explicitely allow for controlling variable survey costs and possible logistic constraints.

Surveys of rare and clustered populations have motivated further advances beyond the traditional sampling designs. Among these, we consider adaptive sampling that was introduced and suggested with the aim of dealing with these sampling situations [22]. Notice that adaptive sampling is here intended as given by Thompson [19]. However literature on responsive and adaptive designs for surveys include various methods for managing data collection, tailoring data collection strategies to different subgroups, prioritizing effort according to estimated response propensities, etc (see [27] for a recent review). Out of the adaptive designs, the most suitable for our epidemiological example is the so-called adaptive cluster sampling (ACS). Introduced by Thompson in the early 90's [19], once a distance measure between units is available, the procedure for selecting units to sample is adapted to the observed values of the variable of interest. The idea is thus that the probability of sampling a unit is influenced by the value observed on nearby units. Many developments and uses of adaptive sampling strategies have been proposed in recent years (see [28] and [17] for a review), however these methods do not allow to account for logistic constraints nor to explicitily allow the planning of the survey costs.

A simple way to deal with logistic constraints and improve the planning of the survey may be to choose, beforehand, a specific route along which to visit units sequentially. The constrained route may be chosen in order to reduce costs and satisfy possible logistic constraints. Within the TB example, choosing a path across a country amounts to defining an ordered list of geographical areas that are then to be inspected one by one in the prescribed order and sequentially assigned to belong or not to the sample. Notice that this might be particularly relevant when planning a survey in developing countries: in fact it might lead to individuate a path along which transportation costs are minimized, for instance, and logistic constraints, such as reduced accessibility of some areas, can be taken into account beforehand. A renewable interest has arisen in sequential designs ([2] and [7]), and a flexible procedure for sequentially select a sample is available. However, the list-sequential setting does not allow, in its current formulation, to oversample cases nor to adaptively incorporate sample evidence.

The aim of this thesis is to develop a new sampling strategy for sampling a rare and clustered population under both cost and logistic constraints. As adaptive designs and sequential designs seem to individually meet the desirable features, we propose the integration of adaptivity in a sequential framework. The proposed integrated approach would then have to address *(i)* logistic and cost issues, *(ii)* oversampling of cases, and *(iii)* estimation of the quantity of interest via a suitable weighting-system. In fact the selection bias due to over-detection of positive cases needs to be adjusted for at the estimation stage. With reference to the inspirational example of WHO's TB surveys, once a route that minimizes transportation costs and satisfies logistic constraints has been decided, *(i)* in the previous list would be tackled. This would require a deep knowledge of the country where the survey is to take place, and possibly a suitable algorithm to choose the best route. In order to meet *(ii)*, we need to introduce adaptivity in a list-sequential framework. The idea, similarly to adaptive strategies is to employ the oberved number of TB cases to update the inclusion probabilities at each step. This will be achieved by suitably changing the updating procedure proposed for sequential sampling designs. Finally, point *(iii)* addresses estimation of the true TB prevalence given sample evidence. As a first proposal in this thesis we derive an unbiased HT-type estimator for the population prevalence by adjusting for both the over-selection bias and for the conditional structure induced by the sequential selection. The performance of the proposed strategy is then evaluated by means of an extensive simulation study aiming at comparing it with the traditional sampling design currently suggested in the WHO guidelines.

This thesis is organised as follows. In the first chapter, after giving the notation that will be used all throughout the thesis, we present and give details of the sampling strategy currently implemented by WHO. In this sampling setting, we give the details of adaptive cluster sampling and list-sequential designs. At the end of the chapter, in the light of our inspirational example of TB prevalence surveys, we show some preliminary simulation results focusing on the adavantages tha ACS and a a sequential design may bring to traditional designs. In the second chapter our first proposal for a strategy that integrates adaptivity in a list-sequential context is presented. The design together with an unbiased estimator for the population prevalence is presented. The design and estimator are also extented to the case in which the target population is structured into primary units (for instance geographical areas) and secondary units (for instance individuals) that better fits our inspirational example of TB prevalence surveys. As the proposed methodology is characterised by a random sample size, Chapter 3 is dedicated to tackle this randomness, which may be a problem in many sampling situations. A way of controlling the final sample size in adaptive sampling is discussed, stressing the reasons why they have a limited application in our sampling situation. Thus we give a way to control the final sample size in our proposed strategy and we provide an unbiased estimator for the prevalence. Chapter 4 is dedicated to studying the behaviour of the proposed strategies by means of an extensive simulation study. The proposed strategies are compared to traditional sampling designs under the profile of *(i)* logistic and cost issues, *(ii)* oversampling of cases, and *(iii)* estimators properties. Last final remarks are discussed together with some research perspectives.

Chapter 1

Some useful sampling designs

This chapter is divided into five sections. In the first section, the basic notation that will be used throughout the thesis is presented. In the second section, with reference to the motivational example of TB prevalence surveys promoted by WHO, we give the details of the survey design currently suggested in the WHO guidelines. In the third and fourth section, details about adaptive cluster sampling and list sequential sampling designs are given underlining the reasons for their utility in this context. Last we present some preliminary simulation results in order to empirically show the advantages and disadvantages of the three discussed designs.

1.1 Basic notation

Let U be a finite population $U = \{1, 2, ..., i, ..., N\}$ composed by N units. We are interested in selecting a random sample from U in order to estimate a parameter of a study variable y over the population. This is often the total $Y = \sum_{i=1}^{N} y_i$ or the mean $\bar{Y} = 1/N \sum_{i=1}^{N} y_i$

of such variable. In this thesis we focus on the case of a dicotomous variable indicating presence/absence of a certain trait, thus the population mean defines the prevalence of such a trait over the population. The random sample $\mathbf{s}$ is selected according to a sampling design which is a discrete probability distribution on the support Q of possible samples $\mathbf{s}$. The probability of getting the sample $\mathbf{s}$ is denoted by $p(\mathbf{s}) > 0 \quad \forall \mathbf{s} \in Q$ and $\sum_{\mathbf{s}\in Q} p(\mathbf{s}) = 1$.

The inclusion of unit i in the sample is formalised with the inclusion membership indicator S_i:

$$S_i = \begin{cases} 1 & \text{if unit } i \text{ is included in the sample} \\ 0 & \text{otherwise} \end{cases}$$

S_i is a Bernoulli random variable with $P(S_i = 1) = E(S_i) = \pi_i$ which defines the first order inclusion probability for unit i.

The random sample is described by the vector of inclusion membership indicators $\mathbf{S} = (S_1, S_2, ..., S_N)$ and a sample $\mathbf{s} = (s_1, s_2, ..., s_N)$, is one of the possible outcomes of $\mathbf{S}$. The sample size n is given by $n = \sum_{i=1}^{N} S_i$ and it may be random.

Together with the first order incusion probabilities, we may consider the second order inclusion probability $\pi_{ii'}$ that is the probability that both unit i and unit i' (with $i \neq i'$) are included in the sample:

$$P(S_i = 1, S_{i'} = 1) = \pi_{ii'}$$

In many sampling situations, it may be convenient to group individuals into larger sampling units called primary units and proceed with the selection of the primary units. If a primary unit is selected, all individuals that belong to the selected primary unit are included in the final sample. This is the case, for instance, of our inspirational example where the sampling units are geographical areas and, once a geographical area is selected in the sample, all the individuals who live in the selected area are included in the final sample.

More formally, suppose the N individuals (secondary units) of the population U are grouped into $1, \ldots, j, \ldots, M$ primary units. Each primary unit contains N_j secondary units so that $\sum_{j=1}^{M} N_j = N$. Let y_{ij} be the survey value of the i-th secondary unit included into the j-th primary unit, so that the survey value associated to primary unit j is given by $y_j = \sum_{i=1}^{N_j} y_{ij}$ or $\bar{y}_j = 1/N_j \sum_{i=1}^{N_j} y_{ij}$. In the case of a binary outcome, which is the case of interest with reference to our inspirational example, $y_{ij} = 1$ if i is a positive case in primary unit j and $y_{ij} = 0$ otherwise. The number of positive cases in unit j is thus $y_j = \sum_{i=1}^{N_j} y_{ij}$. The population value to be estimated is the population total given by $Y = \sum_{j=1}^{M} \sum_{i=1}^{N_j} y_{ij} = \sum_{j=1}^{N} y_j$ and/or the population mean/prevalence given by $\bar{Y} = 1/N \sum_{j=1}^{M} y_j$. In the case of our inspirational example, primary units are geographical areas with N_j inhabitants and $\bar{y}_j = 1/N_j \sum_{i=1}^{N_j} y_{ij}$ is the area specific prevalence.

The sampling units are the primary units. Once a primary unit j is included in the sample of size m primary units, all secondary units $i = 1, \ldots, N_j$ that belong to the selected unit j are included in the sample, so that the final sample size is given by $n = \sum_{j \in \mathbf{s}} N_j$. A set of inclusion probabilities π_j is given for the sequence of all primary units $j = 1, \ldots, M$. The inclusion probabilities may be, for instance, proportional to their (possibly unequal) size N_j, so that the selection method applied is a probability proportional to size (π-ps) method. The focus is thus on the probability of selecting a primary unit j and it's inclusion/not inclusion in the sample is denoted by the usual sample membership indicator S_j. $P(S_j = 1) = E(S_j) = \pi_j$ is the first order inclusion probability for the j-th unit (but also for any secondary unit $i \in j$), while the probability that both primary units j and j' ($j \neq j'$) are included in the sample is given by $P(S_j = 1, S_{j'} = 1) = \pi_{jj'}$.

For certain applications it could be useful or even required to consider ordered populations $U_{ord} = \{(1), ..., (i), ..., (N)\}$, which means that the sampling units are ordered in some predefined way and are sampled according to the given order. In the case of our inspirational example of TB prevalence surveys, it may be convenient to order units along a predefined route along which to sample. This may be the case for instance of countries with logistics constraints due to seasonal bad weather, war zones, street conditions unfitting the transportation needs of the surveying team, etc.

Units are visited sequentially and the decision on whether to include them or not include them in the sample is made. This means that the i-th visit (or j-th in the case of primary units) is the only occasion for unit i (j in the case of primary units) to be included in the sample. For simplicity of notation, from now on, after specifying that we are considering an ordered population, we will simply use the symbol U.

1.2 Current practice in TB prevalence surveys

In TB prevalence surveys, as suggested in the WHO guidelines [30], the population is divided into a certain number of geographical areas, i.e. primary units $j = 1, ..., M$, of as homogenous population size as possible $N_j \approx N_{j'}$ for $j \neq j'$ (the guidelines indicate 400-1000 individuals per area, although in real situations it may also exceed 1500 individuals per area [30]). This working hypothesis allows to control the final sample size thus helping the planning of the survey. Once a geographical area j is included in the sample, all eligible individuals $i = 1, ..., N_j$ (i.e. people aged ≥ 15) are invited to undertake a screening interview (asking for typical TB symptoms) and a chest X-ray. If the individual is negative according to both screening tools, he/she is considered as a non-TB case. If the individual is positive in either of the two screening tools, further examination is carried out via sputum specimen. The specimens are then examined at a central location laboratory. Thanks to a new diagnostic test (called GeneXpert), diagnosis on the field is under consideration and might soon become the norm. We consider here this simplification, hence we consider that the eligible individuals are invited to undertake a medical examination at a moving lab, where any TB positive is detected and possibly treated.

The number of areas m to be sampled is chosen according to a prefixed sample size n. The sample size n is computed as a function of *(i)* a prior guess of the true prevalence, *(ii)* the desired estimation precision (usually a maximum error within 20-25% around the true national prevalence is considered), and *(iii)* an estimate of the variability existing between the areas' prevalences [30]. This computation yields sample sizes that usually range between 30000 and 100000 individuals. More specifically the sample size is calculated using the following equation:

$$n = 1.96^2 \frac{1-p_g}{d^2 p_g} \left(1 + (\bar{N}-1) \frac{\hat{k}^2 p_g}{1-p_g} \right) \tag{1.1}$$

where p_g is a guess of the true population prevalence, d is the desired precision, $\bar{N}$ is the average size of geographical areas (given by $\bar{N} = 1/M \sum_{j=1}^{M} N_j$) and $\hat{k}$ is an estimate of the coefficient of between areas prevalence variation $k = \sqrt{M^{-1} \sum_{j=1}^{M} (\bar{Y}_j - \bar{Y})^2}/\bar{Y}$. It is clear that the bigger the estimated value of the variation between the areas prevalences $\hat{k}$, the bigger is the design effect, hence the final sample size. The sample size is thus larger if the prevalence of TB varies considerably among areas.

The coefficient of between areas variation k is estimated from the results of previous surveys or by making assumptions on the distribution of the areas specific prevalence. However guidelines suggest that it should be set lower than 1 to have a better control over the final sample size. We refer the reader to [30] for further details on the estimation of k.

Additional units may also be considered in the final sample size to account for the participation rate that, although being very high (85-90%), is usually not 100%.

Once the sample size n is defined, the number of geographical areas to be selected m is calculated by dividing it by the expected size of the geographical areas, $m = n/\bar{N}$ assuming that $\bar{N} = N_j \ \forall j = 1,...,M$.

In practice, geographical areas may be selected within strata, such as urban/rural region or by administrative region/district. The WHO guidelines suggest that in the case of stratified sampling, the sample size and the number of areas to be sampled in each stratum should be proportional to the share of the national population in each stratum.

A classic Horvitz-Thompson (HT) approach (stratified if needed) is then employed to estimate the population prevalence.

We call this current practice suggested by WHO Unequal Probability Cluster Sampling (UPCS). Notice that, although stated that the inclusion probabilities should be proportional to size, the geographical areas are of as homogenous size as possible thus reducing to an (almost) equal probability sampling design. Moreover notice that here the word "cluster" is used as in traditional sampling for indicating a group of population units, that, in the context of TB prevalence surveys coincide with the geographical areas in which individuals are grouped. In other occasions, such as in adaptive clusters sampling, the term will be used to identify a group of units close in distance, that have the trait of interest, i.e. a group of geographically closed areas with many cases.

1.3 Adaptive Cluster Sampling

Different approaches are possible under the general idea of adaptive sampling: among these, the most suitable for our epidemiological example is the so-called adaptive cluster sampling (ACS). Introduced by Thompson in the early 90's [19], once a distance measure between units is available, the procedure for selecting units to sample is adapted to the observed values of the variable of interest. The basic idea is that the probability of selecting a unit is influenced by the value observed on nearby units.

Referring to the notation given in Section 1.1, in adaptive cluster sampling, it is assumed that for every primary unit j in the population of M primary units, a neighborhood is defined. The neighborhood of each unit is usually defined as a set of geographically nearest neighbors, although any distance measure can be used (e.g. social or insitutional relationships, genetics, etc ...) as long as the neighborhood relationship is symmetric (e.g. if unit j is neighbour of unit j' then unit j' is neighbour of unit j). Moreover a threshold $y_{min} \in \mathbf{N}$ is chosen so that a condition of the type $y_j > y_{min}$ is defined. An initial sample $\mathbf{s}_0$ of size $m_0 > 1$ is selected according to some probability sampling procedure and the values of the study variable y on units $j \in \mathbf{s}_0$ are observed. If a unit $j \in \mathbf{s}_0$ satisfies the given condition, all units within its neighborhood are added to the sample. The sampling continues untill no more units satisfy the given criterion. Figure 1.1, is an example of the described procedure. The objective is to estimate the number of point-object in a particular region ([19]). A 20×20 grid is overimposed over the study region and given the adaptive condition $y_j > 0$, an initial sample of $m_0 = 10$ geographical units (i.e. quadrats given by the overimposed grid) is selected. For all units that contain at least one point-object, all the neighbouring units are included in the sample untill no more neighbouring units contain at least one point-object.

The traditional HT-estimator is biased in this sampling setting

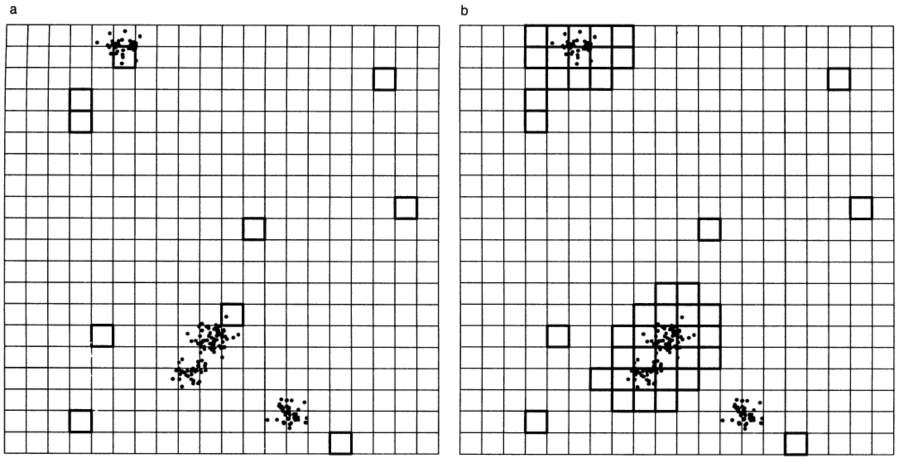

Figure 1.1: Example of adaptive cluster sampling to estimate the number of point-object in a study region. An initial sample of 10 units is taken (left panel) and, for all selected units with at least one point, neighbouring areas are added to the final sample (right panel)

[22]. In fact at the selection stage the inclusion of those units satisfying the adaptive condition $y_j > y_{min}$ has been forced, hence some bias has been introduced and it must be accounted for at the estimation stage.

The simplest unbiased estimator for the mean provided by Thompson is that based on the initial sample of size m_0. This estimator ignores all observations in the sample other than those selected initially and the traditional HT-estimatior is implemented $\hat{y}_0 = \sum_{j \in \mathbf{s}_0} y_j / \pi_j$, where π_j is the probability of being selected in the initial sample.

In order to produce a more sophisticated estimator, Thompson introduces the concept of network which is a subset of a neighbourhood including only units for which the given condition $y_j > y_{min}$ is satisfied. All units for which the adaptive condition is not satisfied, are called edge units. Networks thus create a non overlapping partition of the population, whilst neighbourhoods do not. In fact an edge unit may be edge unit of more than one neighbourhoods, resulting in overlaps between them. A unit j that satisfies the adaptive condition $y_j > y_{min}$ belongs to a network A_j and it may be sampled if it is sampled itself as part of the initial sample $\mathbf{s}_0$ or if any unit of its network A_j is sampled. An edge unit is selected if it is included in the initial sample $\mathbf{s}_0$ or if any of the neighbourhoods for which it is an edge unit is selected. It follows that the exact inclusion probability of any edge unit can be calculated only if it is selected in the initial sample. On the other hand, the inclusion probability for non-edge units can be calculated only if the structure/size of its network is known, i.e. if the network is selected. Thus observations that do not satisfy the adaptive condition are considered in the estmation only if they are selected in the initial sample.

Let the indicator variable K_j be 0 when unit j does not satisfy the given condition and was not selected in the initial sample, and 1 otherwise, an unbiased estimator for the population total is then given by the following:

$$\hat{Y}_{HT*} = \sum_{j \in \mathbf{s}_0} \frac{y_j K_j}{\pi_j^*} \tag{1.2}$$

where π_j^* is the probability that a unit is included in the computation of the estimator [19]. An estimator for the mean/prevalence is readly given by $\hat{\bar{Y}}_{HT*} = N^{-1} \sum_{j \in \mathbf{s}_0} \frac{y_j K_j}{\pi_j^*}$.

The variance $var(\hat{Y}_{HT*})$ and an unbiased estimator for it are given in [19]. We refer the reader to [22] for further details on the estimation.

As such modified HT estimtor is not a function of the minimal sufficient statistic, it can be improved by applying the Rao-Blackwell method. In ACS the minimal sufficient statistics is the set of unordered distinct units in the final sample $\mathbf{s}$ that is $D = \left\{(j, y_j) : j \in \mathbf{s}\right\}$.

Many developments and uses of adaptive sampling strategies have been proposed in recent years. These include two-stage adaptive cluster sampling ([16]), adaptive cluster double sampling ([11]), unequal probability adaptive cluster sampling ([12]; [14]; [18]). Moreover bootstrap confidence intervals for adaptive cluster sampling are discussed in Christman and Pontius (2000). We refer the reader to [22], [17] and [28] for an exhaustive review of all developments of adaptive sampling strategies.

ACS has proven to be more efficient than traditional non-adaptive sampling strategies when the population is rare and clustered and when the within areas prevalence variability is lower than the between areas variability ([22]). As compared to traditional designs, ACS would provide unbiased estimation of the population prevalence while most likely returning a larger amount of cases. However, it does not allow to account for logistic constraints nor to explicitily allow the planning of the survey costs. In the following section, a sampling design that is able to account for these two aspects is thus considered.

1.4 List-sequential sampling

A simple way to deal with logistic constraints and thus improve the planning of the survey may be to choose, beforehand, a specific route along which to visit units sequentially, as opposed to traditional non-sequential sampling designs where the route is set by the specific selected sample. The constrained route would be chosen in order to reduce costs and to satisfy possible logistic constraints. In our motivational example of TB prevalence surveys, choosing a path across a country means to define an ordered list of geographical areas that are to be inspected one by one in the prescribed order and sequentially assigned to be included or not in the sample. As the countries in which TB prevalence surveys are carried out are developing countries, it may be not uncommon for some areas to have limited access, due to, for instance, natural barriers, unusable transportation networks, war areas, etc, motivating the use of a predefined route. The choice of the route should take into account that certain characteristics may be associated with the outcome of interest. E.g. geographical areas that are hard to access due to few roads and the presence of natural barriers should not be left out but a route that goes across them should be planned as these geographical areas could in fact be hot-spots of TB (due to barriers of access to diagnosis). Moreover, although we will consider here just a route that minimises survey costs and deals with logistic constraints, notice that defining a route based on known TB epidemiology hot spots in countries could have important benefits, if there is good knowledge and understanding of the disease.

Let us consider the ordered population U_{ord} of M geographical areas. For simplicity, let us denote such ordered population with U, so that the (ordered) units can be denoted by $j = 1,\dots,M$. The sampler visits all units sequentially so that $t = 1,...,M$ visits are under-taken and decides about the inclusion of each unit. In other words, unit j occupies the j-th position in the ordered population and can only be selected at visit/step $t = j$. The sampling continues untill the decision on unit $j = M$ is made.

For populations whose units can be ordered according to some criterion, Bondesson and Thorburn ([2]) developed a general sequential method for obtaining a π-ps sample. Once a set of initial inclusion probabilities $\pi_j^{(0)} = \pi_j, \quad j = 1,...,M$ is defined, the inclusion probability of units $j \geq t$ are revised at each step t, i.e. the inclusion probabilities of units not yet selected are updated after each unit has been visited. The inclusion probability of the generic unit $j \geq t = 1,...,M$ is updated according to the following updating procedure:

$$\pi_j^{(t)} = \pi_j^{(t-1)} - (S_t - \pi_t^{(t-1)}) w_{j-t}^{(t)} \tag{1.3}$$

where $w_{j-t}^{(t)}$ constitute a weighting system able to produce any sampling design without replacement. Weights are chosen arbitrarely in the following constraints:

$$-min\left(\frac{1-\pi_j^{(t-1)}}{1-\pi_t^{(t-1)}}; \frac{\pi_j^{(t-1)}}{\pi_t^{(t-1)}}\right) \leq w_{j-t}^{(t)} \leq min\left(\frac{\pi_j^{(t-1)}}{1-\pi_t^{(t-1)}}; \frac{1-\pi_j^{(t-1)}}{\pi_t^{(t-1)}}\right)$$

that guarantee that $0 \leq \pi_j^{(t)} \leq 1$.

Once the initial inclusion probabilities are given, the decision about the inclusion/not inclusion of the first unit in the sample is made by means of a Bernoulli trial. For instance, at step $t = 1$, $S_1 = s_1$ and all units with $j > 1$ are updated as follows:

$$\pi_j^{(1)} = \pi_j^{(0)} - (s_1 - \pi_1^{(0)})w_{j-1}^{(1)}$$

At step/visit 2 ($t = 2$), the decision about the selection/not selection of unit 2 is made, $S_2 = s_2$, thus for $j > 2$ the inclusion probabilities are updated as follows:

$$\pi_j^{(2)} = \pi_j^{(1)} - (s_2 - \pi_2^{(1)})w_{j-2}^{(2)}$$

etc.

In general this procedure can be represented with an updating matrix:

visit / unit	$\pi_1^{(0)}$	$\pi_2^{(0)}$	$\pi_3^{(0)}$	$\pi_4^{(0)}$	...	$\pi_M^{(0)}$
t=1	$S_1 = s_1$	$\pi_2^{(1)}$	$\pi_3^{(1)}$	$\pi_4^{(1)}$	...	$\pi_M^{(1)}$
t=2	s_1	$S_2 = s_2$	$\pi_3^{(2)}$	$\pi_4^{(2)}$	...	$\pi_M^{(2)}$
t=3	s_1	s_2	$S_3 = s_3$	$\pi_4^{(3)}$	...	$\pi_M^{(3)}$
...			...			
t=M	s_1	s_2	s_3	s_4	...	$S_M = s_M$

Notice that the Bernoulli experiment is defined on the diagonal of the updating matrix. This means that when $t = j$ the choice about the inclusion of unit j is made upon the performance of a Bernoulli trial with probability $\pi_t^{(t-1)}$, that is at the t-th visit, the $j - th$ unit is/is not included. The last row of the table defines the final sample **s**.

The choice about the weighting system to be used defines the correlation between the sample memberhip indicators S_j. In fact, Bondesson and Thorburn show that the weights can be represented as:

$$w_{j-t}^{(t)} = -\frac{Cov(S_t, S_j|S_1, ..., S_{t-1})}{Var(S_t|S_1, ..., S_{t-1})}$$

Therefore it is possible to choose weights for inducing positive/negative correlation between the sample membership indicators and represent any without replacement sampling design. In general, positive weights induce negative correlations and negative weights induce positive correlations.

The simplest way to sequentially select a sample is by Poisson sampling (see [26] for further details). Visits start from the first unit in the sequence. The sampler makes the decision on whether to include unit 1 with probability π_1. Unit $j = 2$ is thus visited and it is included in the sample with probability π_2. This easy sampling design can be represented with the updating procedure given in Equation 1.3. In fact if $w_{j-t}^{(t)} = 0 \ \forall j > t, t \geq 1$, we get thet $\pi_j^{(t)} = \pi_j^{(0)} \ \forall j = t, ..., M$. This means that $Cov(S_t, S_j|S_1, ..., S_{t-1}) = 0$, that is there is independence within selections. This is the main feature of Poisson sampling and it is also the reason of its simplicity. However it is characterised by a random sample size that may limit its use.

A list-sequential approach allows for setting additional conditions on the weights for creating a sequential sampling design with other desirable features. For instance, in order to have a sampling design with fixed sample size, at each given t the sum of weights must be equals to 1:

$$\sum_{j=t+1}^{M} w_{j-t}^{(t)} = 1$$

If some measure of distance between the population units is available (not necessarily beforehand), it is possible to incorporate it in the weights determination and induce the procedure to produce samples with desired spatial behaviours. This method has been called Spatially Correlated Poisson Sampling (SCPS, [9]). For example, if the researcher wishes to obtain a sample that is well-spread over the geographical area of interest, the so-called maximal weights strategy ([2], [7]) can produce such result; likewise, it is possible to obtain samples that are more or less spatially clumped by choosing an appropriate set of weights. It is worth noting that, although the proposal originated in the field of real-time sampling with application to forestry, it is in principle applicable to a much wider range of situations. As noted for ACS, the defition of distance does not need to be geographical, but it can be thought of a more general measure of similarity (social, genetical, etc.).

With an appropriate choice of weights, any sampling design without replacement can be described under the list-sequential approach. An HT approach to estimation can then be considered, thanks to the fact that the unconditional updated inclusion probabilities are equal to the initial inclusion probailities themselves (see [2] for further details).

1.5 Comparing key characteristics via a simulation study

In the following paragraph we present a preliminary simulation study in which the three designs presented above are empirically compared to highlight their limitations and advantages and to delineate the advatnages of an integrated strategy in the context of our inspirational example of TB prevalence survey. Simulations focus on the comparison between UPCS, ACS and a sequential approach (SCPS). The three designs are compared with respect to *(i)* survey costs, *(ii)* cases detection and *(iii)* estimators properties.

The simulation has been carried out completely in the R environment ([15]), and the packages spatstat ([1]) and BalancedSampling ([10]) have been used to implement spatial patterns generation and SPCS, respectively. ACS was implemented thanks to the functions kindly provided by Mary Christman and Kristen Sauby.

Different artificial populations assumed as possible TB prevalence surveys scenarios have been simulated. For the sake of illustration and to stress limitations and advantages of each of the considered strategies, we here show two of the simulated scenarios (Figure 1.2). Both simulated populations are composed by $N = 200000$ individuals evenly spread over a two-dmensional space. The study variable y has value 1 for population units that are TB cases (represented as dots in the Figure) and value 0 otherwise. The actual population TB prevalence $\bar{Y} \approx 0.01$. This is in line with the estimated number of cases in most of the countries where TB prevalence surveys are carried out (prevalence usually ranges between 0.1 % and 1.5% [31]).

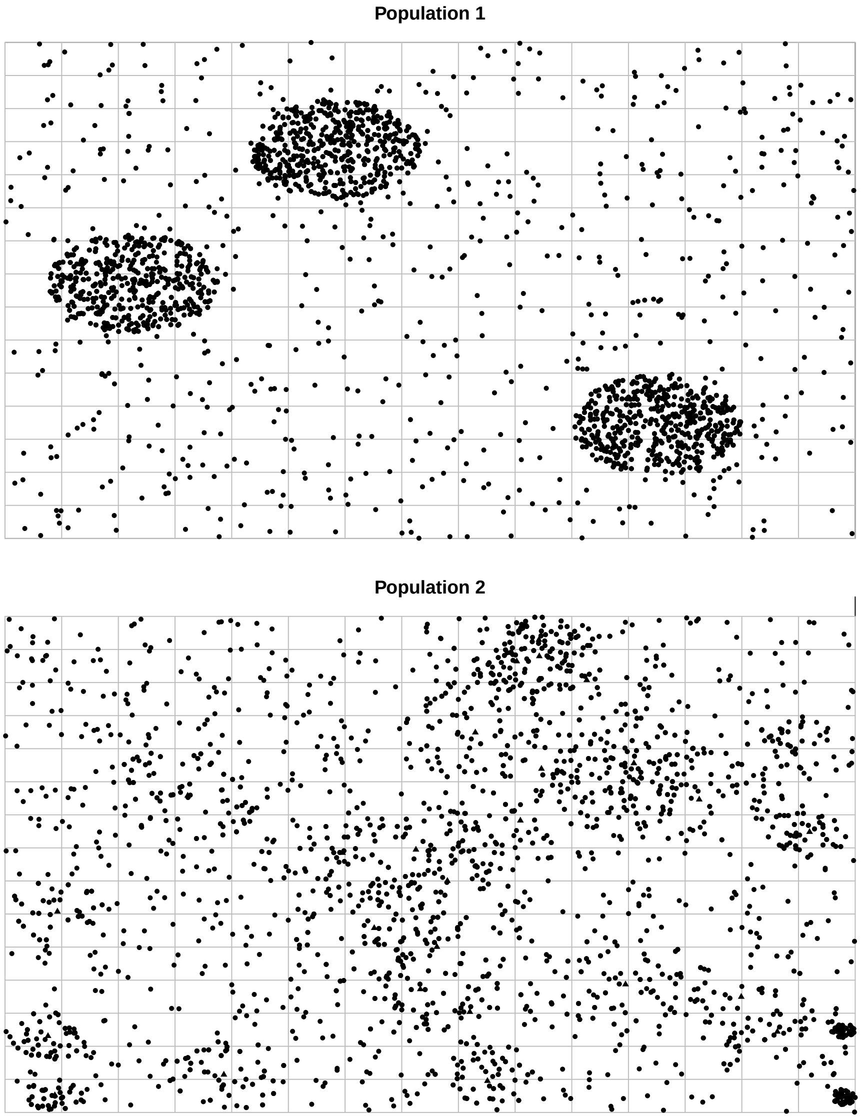

Figure 1.2: Two of the simulated scenarios. Cases are represented as dots, $N = 200000$. Upper panel: true prevalence $\bar{Y} \approx 0.01$, total number of cases $Y = 1980$. Lower panel: true prevalence $\bar{Y} \approx 0.01$, total number of cases $Y = 1956$.

In population 1 cases are mostly clustered in 3 groups homogeneous in terms of prevalence, while in the second scenario, cases are clustered in 30 small homogenous groups, thus population 2 appear to be less clustered. In both populations the overimposed 15×15 grid generates a set of $M = 15^2 = 225$ areas from which to sample.

According to WHO guidelines we assumed to have a perfect guess of the true prevalence ($p_g = 0.01$). Under the suggested UPCS design, the required sample size is thus, in both situations, of $n = 19729$ individuals; this means that a sample of $m = 23$ areas is selected with 100% participation rate within selected areas (which is a good approximation of the actual participation rate for TB prevalence surveys, usually in the range 85- 90% [30]).

The four methods compared are the design currenly implemented by WHO (with the sample size $m = 23$ calculated as discussed above), ACS with an initial sample of size $m_0 \approx \frac{1}{2}m = 12$ (denoted by ACS_1), ACS with an initial sample of size $m_0 \approx \frac{2}{3}m = 18$ (denoted by ACS_2) and SCPS with maximal weights strategy (see [9] for further details). The adaptive condition was set to $y_j > p_g \cdot N_j$, meaning the sampling efforts are concentrated close to areas that exceed the number of cases according to the prevalence guess.

For all the compared designs, the total survey cost has been computed based on the following linear cost function:

$$C = c_0 + c_1 m + c_2 n \tag{1.4}$$

where c_0 is a fixed cost, for instance for equipment and staff ($c_0 = 100000\$$); c_1 is the unitary cost for each selected area in the sample for instance for transportation and installation of the moving lab in the selected location ($c_1 = 1000\$$); and c_2 is a unitary cost for each individual data collected in every selected area ($c_2 = 10\$$).

We considered the advantage of a careful route planning, easily accomodated by the sequential approach, by reducing by 20% the area sampling cost for SCPS. With regards to ACS, we considered the cost function given by Thompson [22], thus considered the "complete" area cost of $c_1 = 1000\$$ only for areas selected in the initial sample and applied a 20% discount to the cost c_2 for each area added adaptively.

Figure 1.3 summarizes the results of 10000 Monte Carlo runs on the two simulated populations (upper panel for population 1 and lower panel for population 2). The parameter to estimate is the population prevalence that, as expected from the theory, is unbiasedly estimated by all methods here compared. As expected ACS outperforms the other sampling designs in terms of detection power, although the final sample size is extremely variable as showed in Table 1.1. Due to the distribution of y in the population, the sample size is more variable and larger in population 2 with a final sampling fraction reaching 70%. Thanks to a moderate choice of the initial sample size, ACS_1 seems to be able to oversample cases while moderately increasing the final sample size. This suggests that with an accurate choice of the initial sample size, possible for instance thanks to a priori information, and in a situation with no logistic constraints, ACS may be recommended. A large and variable sample size leads to very large and variable final costs as they are a linear function of the number of areas and individuals selected. On the other hand, the sequential approach presents a very stable behaviour in terms of costs, also highlighting the gain in planning a cost-minimizing route

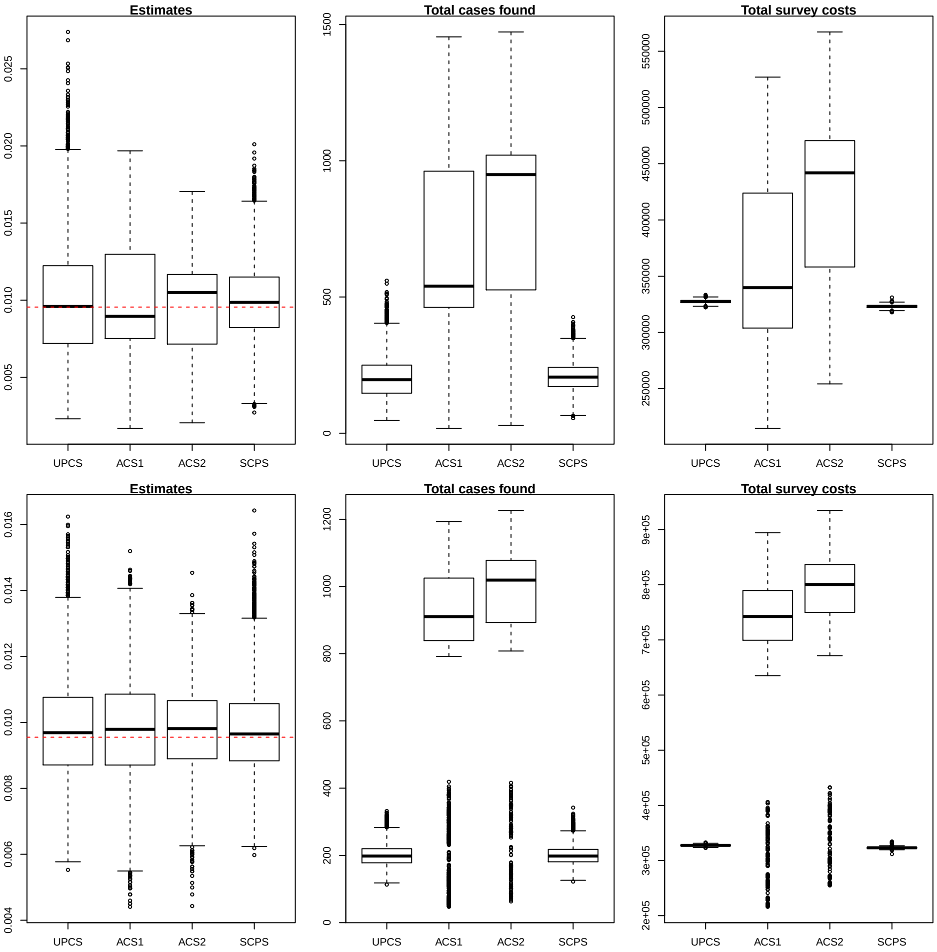

Figure 1.3: MC distributions of the estimators under the four sampling designs (the dashed line indicates the true prevalence), the number of detected cases, and the total costs (plotted on the log scale for better readability). Upper panel: population 1; lower panel: population 2

beforehand, but is unable, as expected, to oversample cases as compared to the traditional UPCS. The costs reduction in planning the route, is given by the discount applied to each added area. In this situation the overall reduction is 5% given that each added area is applied only a 20% discount.

The use of adaptive cluster sampling has been suggested as a natural alternative to the current methodology used by WHO in TB surveys when the primary aim, along with unbiased estimation of the population TB prevalence, is to overdetect TB cases. Simulation results show that ACS manages to oversample cases in both the depicted scenarios, although the number of areas selected, thus the final sample size, strongly depend on the spatial pattern and may result in uncontrollable final costs. On the other hand, in both the scenarios depicted, the use of a sequential procedure lower costs while logistic constraints are accounted for. Therefore the use of an adaptive design and the sequential approach seem to individually meet different desirable features in sampling a rare and clustered trait suggesting that an integrated strategy may be able to both address logistic and cost issues as well as the oversampling of cases. Starting from the designs discussed in the above paragraphs, in the next few chapters we propose a new sampling strategy comprising both a sequential component and an adaptive component.

Elementary Statistics	Population 1				Population 2			
	UPCS	ACS_1	ACS_2	SCPS	UPCS	ACS_1	ACS_2	SCPS
Min.	19934	10277	13822	19947	19981	10396	3870	9405
1st Qu.	20341	18207	23202	20374	20350	54657	59085	20377
Median	20442	21560	30611	20472	20445	58288	63592	20467
Mean	20444	22997	29029	20472	20445	57366	62611	20469
3rd Qu.	20546	29003	33263	20569	20538	62732	66994	20557
Max.	21039	38253	41916	21187	20944	71729	75263	21518

Table 1.1: Final sample size: elementary MC statitistics over 10000 runs

Chapter 2

The proposed strategy: Poisson Sequential Adaptive (PoSA)

In this chapter we outline our proposal of an integrated strategy, able to combine adaptivity in a sequential framework to obtain a sampling design that is able to over-detect cases while considering costs and logistic constraints. In the first section, after the main features of the proposed design are given, an unbiased estimator for the population mean/prevalence is provided together with its exact variance and an estimator for it. In order to apply the proposed methodology to our inspirational example of TB prevalence surveys, in the second section of this chapter the proposed strategy is extended to a sampling structure where primary units are geographical areas. The chapter ends by presenting some preliminary simulation results to highlight the main advantages and limitations of the proposed strategy.

2.1 PoSA sampling design

The preliminary simulation results illustrated at the end of Chapter 1, show that an adaptive design and a sequential design may separately achieve enhancement of cases detection and may allow for considering costs and logistic constraints. The integration of adaptivity in a sequential framework thus seems a reasonable way to achieve both the desired features.

As a first proposal and for the sake of its simplicity we consider a Poisson-type design [26] which we named Poisson Sequential Adaptive Sampling Design, PoSA for short. The proposed design is composed of a sequential component for dealing with logistic and cost constraints and of an adaptive component for enhancing cases detection. The sequential component consists in choosing, beforehand, a specific route, which best allows for reducing (possibly minimizing) on-field survey costs and of acknowledging possible logistic constraints. Along the chosen route units are sampled (visited) sequentially, as illustrated in Section 1.4. The infectivity of a disease, such as TB, may result in groups of TB positives that are geographically close. Thus, once a TB positive is found, it is desirable to force the inclusion of units that are (geographically) close to such individual and possibly positive cases to treat. The adaptive component is therefore based on the idea that the probability of including a unit in the sample may depend on the values observed on nearby units. Notice that our proposal is a Poisson based design and 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 Chapter 3.

Let us make one additional consideration regarding our inspirational example, before giving the details of the proposed design. As mentioned in Section 1.2, non-TB cases are detected on the field, while for individuals that were found positive in at least one of two screening tools, additional specimens are taken and sent to a central lab. However, in a couple of years the detection and treatment of TB cases will be on the field thanks to a the GeneXpert test. We hence refer to the case of on field detection for simplicity. Notice that, even if we do not consider the new test, the proposed method still applies if by 'TB-positives' we consider individuals that were found positive to one of the two screening tools.

Given the population $U = \{1, \ldots, i, \ldots, N\}$ ordered according to the above considerations, units are thus visited following the chosen route. At each step of the sequential selection, unit i is/is not selected in the sample with probability given according to a chosen rule which integrates the adaptive feature of the design. More formally let y_i be the survey value of the i-th unit and let y_{min} be a threshold chosen such that the following adaptive condition qualifies i as a unit with a significant value of y:

$$y_i > y_{min} \tag{2.1}$$

We consider here the case in which y is a dichotomous variable, thus the adaptive condition can be written as $y_i > 0$, i.e. $y_i = 1$ versus the case in which the condition is not satisfied that is when $y_i = 0$. In the case of our inspirational example the study variable y indicates presence/absence of TB.

Given a set of initial inclusion probabilities $\pi_i^{(0)} = \pi_i$ for the sequence of units $i = 1,\ldots,N$, at the t-th step of the sequential selection, unit $i = t$ is certainly selected (i.e. is selected with probability 1) if the previous unit $i-1$ was selected and adaptive condition in Equation 2.1 holds for it, otherwise it is selected with probability π_i.

The inclusion membership indicator S_i deserves special attention. In fact it interprets the Bernoulli trial upon which the decision to select/not select unit $i = t$ is made, at t-th visit, but in our PoSA proposal, such decision also includes the adaptive condition. Unit i is certainly included into the sample (i.e. with unitary probability) if the previous $i-1$ selection has resulted into a selected case; otherwise it is selected with (unaltered) probability π_i. In other word, S_i is a mixture of a bernoulli random variable and of a degenerate (unitary) random variable depending on whether or not the adaptive condition holds at the previous step of the selection sequence. Thus the probabililty function of the random variable S_i is given by:

$$P(S_i = s_i) = y_{i-1}P(S_{i-1} = 1)\cdot 1 + [1 - y_{i-1}P(S_{i-1} = 1)]\,\pi_i^{s_i}(1-\pi_i)^{1-s_i} \tag{2.2}$$

where $s_i = \{0,1\}$. Notice that Equation 2.2 coincides with the probability distribution of a Bernoulli random variable with parameter π_i if the adaptive condition is not satisfied for the previous unit $i-1$. For simplicity we can thus write, for $i = 1$, $S_i \sim Bernoulli(\pi_1)$ and for $i = 2,\ldots,N$:

$$S_i \sim Bernoulli(\pi_i)(1 - y_{i-1}P(S_{i-1} = 1)) + y_{i-1}P(S_{i-1} = 1) \tag{2.3}$$

It follows straightforward that, for $i = 2,...,M$ the expectation of S_i is:

$$E(S_i) = E(Bernoulli(\pi_i)(1 - y_{i-1}P(S_{i-1} = 1)) + 1 \cdot y_{i-1}P(S_{i-1} = 1) \tag{2.4}$$

$$= \pi_i - \pi_i y_{i-1}E(S_{i-1}) + 1 \cdot y_{i-1}E(S_{i-1})$$

$$= \pi_i + y_{i-1}E(S_{i-1})(1 - \pi_i)$$

and for $i = 1$, $E(S_1) = \pi_1$. For simplicity Equation 2.4 can be written as:

$$E(S_i) = \begin{cases} 1 & \text{if } s_{i-1}y_{i-1} = 1 \\ \pi_i & \text{otherwise} \end{cases} \tag{2.5}$$

Hence the variability of the sample membership indicator S_i is trivially given by:

$$V(S_i) = E(S_i)(1 - E(S_i)) \tag{2.6}$$

where $E(S_i)$ is given in Equation 2.5.

It is important to notice that, for the PoSA design, the addition of the adaptive component to the sequential selection implies a partial loss of the independent selection caracterizing a simple (non-adaptive) Poisson sampling. In particular for a PoSA design and limited to a pair of subsequent units, selection are not independent. If unit $i' = i-1$, there exists dependence induced by the value of y_{i-1}. In fact if unit $i - 1$ is selected and $y_{i-1} > y_{min}$, that in our simplification coincides with $y_{i-1} = 1$, $P(S_i = 1|S_{i-1} = 1) = 1$, while they are independent otherwise. By using symmetry we consider $i\prime < i = 2,\ldots,N$, we can thus write the mixed moment under PoSA design:

$$\begin{aligned} E(S_i, S_{i-1}) &= P(S_i = 1, S_{i-1} = 1) \\ &= P(S_i = 1|S_{i-1} = 1)P(S_{i-1} = 1) \\ &= [1y_{i-1} + \pi_{i-1}(1 - y_{i-1})]E(S_{i-1}) \end{aligned} \tag{2.7}$$

Therefore, two cases hold depending on whether or not S_i and $S_{i'}$ refer to subsequent units $i-1$ and i. Otherwise, for any pairs of non-subsequent units selection independence holds. The mixed moment can thus be rewritten as follows:

$$E(S_i, S_{i'}) = \begin{cases} E(S_i)E(S_{i'}) & \text{if } i' \neq i-1 \\ [y_{i-1} + (1 - y_{i-1})\pi_i]E(S_{i-1}) & \text{if } i' = i-1 \end{cases} \tag{2.8}$$

It follows straightforward that for every pair (i, i') such that $i = 2,\ldots,N$ and $i' < i$ we have:

$$cov(S_i, S_{i'}) = \begin{cases} 0 & \text{if } i' \neq i-1 \\ E(S_{i-1})[y_{i-1} + \pi_i(1 - y_{i-1}) - E(S_i)] & \text{if } i' = i-1 \end{cases} \tag{2.9}$$

With little algebra, by substituing Equation 2.4 in Equation 2.9, we find that the covariance is given by:

$$cov(S_i, S_{i'}) = \begin{cases} 0 & \text{if } i' \neq i-1 \\ E(S_{i-1})[1 - E(S_{i-1})]y_{i-1}(1 - \pi_i) & \text{if } i' = i-1 \end{cases} \tag{2.10}$$

which is readily implementable.

The proposed PoSA sampling procedure is synthetized in Algorithm 1.

```
Input: Ordered sequence of $N$ units & a set of inclusion
  probs $\pi_1, ..., \pi_N$.
Return: vector of sample membership indicators of size $N$
Procedure: Visit unit $i = 1$ and select with probability $\pi_1$.
  If $s_1 = 1$, collect $y_1$
for $i$ in $2 : N$ do
    point unit $i$
    if $y_{i-1}s_{i-1} = 1$ then
        select with probabilty 1
    else
        select with probability $\pi_i$
    end
    if $s_i = 1$ then
        collect $y_i$
    end
end
```

Algorithm 1: PoSA Algorithm

The set of sample membership indicator S_i and of expectations above, provides us with a complete formal description of the PoSA sampling design and its implementation at the selection stage of the survey. Prior to considering step ahead to the estimation stage, we make some important remarks.

First notice that, as a first proposal PoSA sampling design is expected to meet the objective of improving over traditional sampling design in case of a rare and spatially clustered trait as well as in presence of budget and logistic constraints, particularly by oversampling the cases $y_i = 1$. However, being Poisson design based, the procedure leads to a random sample size. Unlike to a simple (non-adaptive) Poisson design, an accurate analysis of the variability of the PoSA sample size in complicated by its adaptive component, namely n depend also on the distribution of the study variable y on the (ordered) population. In a simple Poisson design, the random sample size has a so called Poisson-Binomial probability distribution [26]. In fact the final sample size $n = \sum_{i=1}^{N} S_i$ is the sum of independent Bernoulli random variables with possibly different parameters π_i. In a PoSA design the independence is partially lost since subsequent S_{i-1} and S_i are actually dependent. Moreover, as detailed above, every S_i, $i = 2, \ldots, N$ is either a Bernoulli and or a degenere (unitary) random variable depending upon the adaptive condition, i.e. the distribution of y in the (ordered) population. Some trivial extreme cases can be easily discussed.

i) No cases are present in U, i.e. y is constantly null for every population unit; and *ii)* U is composed completely by cases, i.e. y is constantly unitary for every population unit. For case *i)* our design is a traditional Poisson design thus the results on the distribution of n in the traditional Poisson design hold. For case *ii*), the probability of sampling the whole population is given by $P(n = N) = P(S_1 = 1) = \pi_1$, while $P(n = 0)$ is given by a Poisson Binomial distribution with the set of parameters $\pi_1, ..., \pi_N$.

Future research will focus on studying the variability of PoSA random sample size in non-trivial cases.

Secondly notice that, as already observed, random sample size may be a practical issue for the survey planning. In Chapter 3 a way to control the final sample size is proposed without losses in the ability to over detect cases.

Finally notice that the sequential feature of PoSA allows us to also formalize it according to the general list-sequential formulation as illustrated in Section 1.4. Particularly, by using a list seqeuntial approach, the PoSA updating-matrix can be written as:

visit / unit	$\pi_1^{(0)}$	$\pi_2^{(0)}$	$\pi_3^{(0)}$	$\pi_4^{(0)}$	...	$\pi_M^{(0)}$
t=1	$S_1 = s_1$	$\pi_2^{(1)}$	$\pi_3^{(1)}$	$\pi_4^{(1)}$	...	$\pi_M^{(1)}$
t=2	s_1	$S_2 = s_2$	$\pi_3^{(2)}$	$\pi_4^{(2)}$	...	$\pi_M^{(2)}$
t=3	s_1	s_2	$S_3 = s_3$	$\pi_4^{(3)}$	...	$\pi_M^{(3)}$
...			...			
t=M	s_1	s_2	s_3	s_4	...	$S_M = s_M$

where sample membership indicator S_i are defined by Equation 2.3. For any row $t = 1, ..., N$ and unit $i < t$, the (updated) inclusion probabilities are given by:

$$\pi_i^{(t)} = \begin{cases} 1 & s_{i-1}y_{i-1} = 1 \text{ and } i = t+1 \\ \pi_i^{(t-1)} - (S_t - \pi_t^{(t-1)})w_{i-t}^{(t)} & \text{otherwise} \end{cases} \tag{2.11}$$

with the trivial choice for the weights $w_{i-t}^{(t)} = 0 \ \forall t, i$.

Different non-trivial choices for the weighting system would lead to designs with a different sequential component, more complex that the basic one contemplated in PoSA. In this work we chose Poisson sampling for its simplicity but the choice of a design with other characteristics, yet accomodating a predefined route, is still possible. In Chapter 3 we will better address this possibility.

Estimation under PoSA sampling design

We now illustrate how an Horvitz-Thompson type (HT) estimator can be derived under the PoSA sampling design. Unlike for a simple (non-adaptive) Poisson sampling with independent (though sequential) selections, a main issue to be acknowledged here is that the adaptive feature of the design. In fact it induces a conditional structure over each pair of subsequent sample membership indicators S_{i-1} and S_i. Hence an unbiased HT estimator for the population total prevalence (mean) under PoSA sampling design has the following form:

$$\hat{\bar{Y}}_{PoSA} = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i S_i}{E(S_i)} \tag{2.12}$$

which is unbiased by construction.

In practice Equation 2.4 represents the design weight and it may be calculated for all units of the population (although it is not necessary for estimation purposes) according to the sample selected. Table 2.1 represents the values to be used as design weights according to the selected sample.

The exact variance for the estimator $\hat{\bar{Y}}_{PoSA}$ can easily be calculated as follows:

Table 2.1: deign weights in PoSA estimator

s_{i-1}	y_{i-1}	design weight for unit i
0	0	π_i
0	1	π_i
1	0	π_i
1	1	1

$$V(\hat{\bar{Y}}_{PoSA}) = \frac{1}{N^2}\left[\sum_{i=1}^{N} y_i^2 \frac{E(S_i)(1-E(S_i))}{E(S_i)^2} + 2\sum_{i'<i}\sum_{i=2}^{N} y_i y_{i'} \frac{cov(S_i, S_{i'})}{E(S_i)E(S_{i'})}\right] \tag{2.13}$$

By substituting Equation 2.9 into 2.13, and with little algebra, the following expression for the exact variance of the $Po\hat{\bar{S}}A$ estimator is found:

$$V(\hat{\bar{Y}}_{PoSA}) = \frac{1}{N^2}\left[\sum_{i=1}^{N} y_i^2 \frac{(1-E(S_i))}{E(S_i)} + 2\sum_{i=2}^{N} y_i y_{i'}^2 \frac{(1-E(S_{i-1})(1-\pi_i))}{E(S_i)}\right] \tag{2.14}$$

PoSA design can be represented with the updating matrix and the updating prodecure as given in Equation 2.11. Notice that $\forall \quad t \leq i, \quad i = 2, ..., N$, $E(S_i) = \pi_i^{(i-1)}$, while for $i = 1$, $E(S_1) = \pi_1$. It follows that a formula for easily calculating the value for the PoSA estimator is the following:

$$\bar{\hat{Y}}_{PoSA} = \frac{1}{N}\sum_{i \in \mathbf{s}} \frac{y_i}{\pi_i^{i-1}}$$

In other words, for calculating the PoSA estimator we use the inclusion probabilities for the selected units updated at step $i-1$. Notice then that the updating matrix contains all the information for calculating the point estimate for the estimator and for the variance of the estimator.

In particular the variance estimation can be simplified and calculated as follows:

$$v(\hat{\hat{Y}}_{PoSA}) = \frac{1}{N^2}\left[\sum_{i \in \mathbf{s}}^{N} y_i^2 \frac{(1-E(S_i))}{E(S_i)^2} + 2 \sum_{(i,i') \in \mathbf{s}}^{N} y_i y_{i'}^2 \frac{(1-E(S_{i-1}))(1-\pi_i)}{E(S_i)E(S_i, S_{i-1})}\right] \tag{2.15}$$

where (i,i') refers to all the pairs of units in the sample that are subsequent units in the (ordered) population and $E(S_i, S_{i-1})$ corresponds to the joint inclsuion probability. Again, using information of the updating matrix, Equation 2.15 can be written as follows:

$$\begin{aligned} v(\hat{\hat{Y}}_{PoSA}) =& \frac{1}{N^2}\left[\sum_{i \in \mathbf{s}}^{N} y_i^2 \frac{(1-\pi_{i-1}^{(i-1)})}{(\pi_i^{(i-1)})^2}\right. \\ & \left. + 2 \sum_{(i,i') \in \mathbf{s}}^{N} y_i y_{i'}^2 \frac{(1-\pi_{i-1}^{(i-2)})(1-\pi_i)}{\pi_i^{(i-1)}\pi_{i-1}^{(i-2)}(y_{i-1}+y_{i-1}(\pi_i-\pi_i))}\right] \end{aligned} \tag{2.16}$$

2.2 PoSA in the example of TB prevalence survey

In some applications, as it is the case of our motivational example of TB prevalence survey, population units naturally appear as grouped into larger primary sampling units such as geographical areas. The sampling design, according to the current WHO guidelines ([30]), is thus composed of two selection stages: *i)* a random selection of primary sampling units; and *ii)* the complete selection of all population units belonging to the selected primary sampling units. The sampling design that is currently used in TB prevalence surveys considers the target population grouped into geographical areas to be sampled. Once the geographical areas (primary units) are sampled, all individuals are invited to undertake a medical examination aiming at the diagnosis of TB. It is thus natural to extend the PoSA method to this sampling situation, and we consider the extension of PoSA to cluster sampling. By "cluster" we denote here the group of individuals that belong to the same primary unit, in accordance to traditional cluster sampling and in accordance to UPCS design. However when giving details about the sampling design we will refer to "primary units" rather than "clusters" in order not to generate confusion with "clusters" as a group of (geographically) close cases.

Recalling the notation given in Section 1.1, each primary unit $j = 1, ..., M$ contains N_j individuals $i = 1, ..., N_j$ so that $\sum_{j=1}^{M} N_j = N$. The number of positive cases in unit j is $y_j = \sum_{i \in j} y_{ij}$. Let y_{min} be a threshold chosen such that the following adaptive condition qualifies j as a (primary) unit with a significant number of positive cases:

$$y_j = \sum_{i=1}^{N_j} y_{ij} > y_{min}$$

Notice that if we look at the prevalence instead of the total number of cases, we can refer to the prevalence for unit j as $\bar{y}_j = \sum_{i \in j} y_{ij}/N_j$. The threshold that qualifies unit j as a unit with a significant number of cases is thus $\bar{y}_{min} \in (0, 1)$ and the adaptive condition is as follows:

$$\bar{y}_j = \frac{1}{N_j}\sum_{i=1}^{N_j} y_{ij} > \bar{y}_{min} \tag{2.17}$$

The primary units to be sampled are here geographical areas and are ordered, for instance, along a route minimizing fixed survey costs and acknowledging logistics constraints, as discussed in Section 1.4. The geographical unit j is thus the unit that occupies the j-th position in the ordered population.

A set of inclusion probabilities π_j is given for the sequence of all units $j = 1,\ldots,M$, for instance proportional to their (possibly unequal) size N_j. At the j-th step of the sequential selection, unit j is certainly selected (i.e. is selected with probability 1) if the previous unit $j-1$ is selected and the adaptive condition given in Equation 2.17 holds for the previous unit $j-1$. Otherwise area j is selected with probability π_j.

The PoSA sampling procedure, adapted to the context of TB prevalence surveys, can be synthetized in a ready-to-implement set of instructions. The vector of sampled units is composed by the inclusion memberhip indicators S_j of areas $j = 1,...,M$, taking value 1 if unit j is included in the sample and 0 otherwise. At each step of the sequential selection, π_j is updated adaptively by means of a further indicator I_j taking value 1 if the adaptive condition in Equation 2.17 holds in unit j, and 0 otherwise. Algorithm 1 is thus slightly modified as showed in Algorithm 2.

```
Input: Ordered sequence of $M$ units & a set of inclusion
  probs $\pi_1,...,\pi_M$.
Return: vector of sample membership indicators of size $M$
Procedure: Visit unit $j = 1$ and select with probability $\pi_1$.
  If $S_1 = 1$, collect $y_1$ and $I_1$
for $j$ in $2 : M$ do
    point unit $j$ if $I_{j-1}s_{j-1} = 1$ then
        select with probabilty 1
    else
        select with $\pi_j$
    end
    if $S_j = 1$ then
        collect $y_j$ and $I_j$
    end
end
```

Algorithm 2: Cluster PoSA Algorithm

All the results showed in paragraph 2.1 still hold, but formulas are sligthly modified.

For every pair (j, j') such that $j = 2, \ldots, M$ and $j' < j$ we have:

$$
\begin{aligned}
E(S_j, S_{j-1}) &= P(S_j = 1, S_{j-1} = 1) \\
&= P(S_j = 1|S_{j-1} = 1)P(S_{j-1} = 1) \\
&= \Big[1I_{j-1} + \pi_{j-1}(1 - I_{j-1})\Big]E(S_{j-1})
\end{aligned}
$$

which allows for deriving:

$$cov(S_j, S_{j'}) = \begin{cases} 0 & \text{if } j' \neq j-1 \\ E(S_{j-1})\left[I_{j-1} + \pi_j(1-I_{j-1}) - E(S_j)\right] & \text{if } j' = j-1 \end{cases} \tag{2.18}$$

An unbased estimator for the population mean/prevalence is then given by:

$$\hat{\bar{Y}}_{PoSA} = \frac{1}{N}\sum_{j=1}^{M}\sum_{i=1}^{N_j} y_{ij}\frac{S_j}{E(S_j)} = \frac{1}{N}\sum_{j=1}^{M} y_j\frac{S_j}{E(S_j)} \tag{2.19}$$

and the results on its variance still hold, but are slightly modified as in Equation 2.18.

2.3 Preliminary empirical results on the PoSA design

In this section we show some preliminary simulation results on the PoSA design in order to stress its advantages and limitations and compare it with the designs presented in Chapter 1. More specifically simulations focus on comparing UPCS, ACS, a sequential approach (SCPS) and PoSA with respect to *(i)* total survey costs, *(ii)* ability to detect positive cases, *(iii)* sample size (which is fixed for UPCS and SCPS and it is a random variable for PoSA and ACS) and *(iv)* estimators properties.

As in Section 1.5, we simulated two artificial populations assumed as possible TB prevalence survey scenarios. For the sake of stressing the main advantages and limitations of our new strategy, we show the same scenarios as in Section 1.5 (please refer to

Figure 1.2). Recalling the characteristics of the populations, both are composed by $N = 200000$ individuals evenly spread over a two-dmensional space; the overimposed 15×15 grid generates a set of $M = 15^2 = 225$ areas assumed as primary sampling units. The study variable y has value 1 for population units that are TB positive cases (represented as dots in the figure) and value 0 otherwise. The actual population TB prevalence $\bar{Y} \approx 0.01$. We assumed to have a perfect guess of the true prevalence (0.01) and, according to the guidelines given by WHO. According to WHO we determined that the required sample size is $n = 19729$ individuals, that is a sample of $m = 23$ areas. The adaptive condition applied to ACS and PoSA is given based on the prevalence guess. In fact, if p_g is a good guess for the population prevalence, it is reasonable to assume that if we encounter a unit with area specific prevalence higher than that expected throughout the population ($y_j > N_j p_g$), we can assume to have encountered a cluster of positive cases.

For all the compared designs, the total survey cost has been computed based on the linear cost function given in 1.4 including *(i)* a fixed cost, for instance for equipment and staff ($c_0 = 100000\$$); *(ii)* a unitary cost for each selected area, for instance for transportation and installation of the moving lab in the selected location ($c_1 = 1000\$$); and *(iii)* a unitary cost for each individual data collected in every selected area ($c_2 = 10\$$). We considered the advantage of a careful route planning, easily accomodated by the sequential approach, by reducing by 20% the area sampling cost for SCPS and PoSA, as well to those areas added adaptively in ACS.

In Figure 2.1 the main results on designs and on estimators are synthetised. Top panel refers to population 1 while bottom panel refers to population 2. Notice that, as opposite to the modified HT-estimator used for ACS, the distribution of the proposed estimator $\hat{\hat{Y}}_{PoSA}$ seems symmetric regardless of the distribution of variable y in the population, that may be desirable, for instance, for interval estimation. On the other hand, there seems to be a loss of efficiency, especially in the less clustered population, with respect to the traditional HT estimator used in UPCS and SCPS. The latter seems to be the most stable as it is based on a spatially balanced sample. The loss in efficiency in PoSA estimator is due to the fact that, very small samples with only few cases are possible, as well as large samples with possibly a large number of cases. In fact, as showed in table 2.3, the final sample size, especially in population 2, is very variable. Moreover, notice that the variability in the final sample size in ACS, is due to additional units added because they satisfy a given adaptive condition, while in PoSA, large samples are not always associated with additional cases. A control over the final sample size may thus be desirable for reducing the variability of the proposed estimator and ensuring that additional areas are areas with a large number of cases.

The main objectives pursued with our proposed PoSA strategy were to enhance case-detection and to control over final costs while dealing with logistic constraints. With regards to the first characteristic, we notice that the PoSA design is able to detect a much larger number of cases as compared to traditional sampling designs, however its detection rate is not as high as that of ACS. This is due to the fact that once a route is chosen, the PoSA design, as thought in this first and simplest formulation, does not allow for deviations, hence sampling only subsequent units. In fact the route is tailored to deal with logistic and cost constraints and it is not tailored on the shape of the clusters (here we chose an up-and-down route, while the clusters are circular). On the other hand, ACS follows the shape of the clusters regardless of logistic constraints. As a consequence cases might be missed more likely under PoSA than than under ACS. The final sample size, as showed in table 2.3 is on average smaller than that of ACS, although, as discussed in Section 2.1, its variability depends on the structure of the study variable y in the population and it is not on average equal to the sum of the initial inclusion probabilities (here $\sum_{j=1}^{M} \pi_j = 23$ which is the number of areas to be sampled according to the WHO design). More specifically when cases are more evenly distributed throughout space (population 2) the average sample size grows larger as well as its whole distribution. With regards to the final survey costs, PoSA design manages to lower costs as compared to ACS and it is able, differently from both UPCS and ACS to deal with logistic constraints. However the large variability in the final sample size leads to unpredictable final survey costs. In many applications, especially when the survey cost per sampled unit is large, it may be desireable to control the final survey costs by fixing the sample size. In our inspirational example for instance, the planning of the survey costs is important for ensuring the right budget, although during the survey the budget needed may be revised and adjusted.

This preliminary empirical evidence give some clear indications about the advantages of our proposed PoSA strategy as well as highlighing some weaknesses. The PoSA design at both the selection and

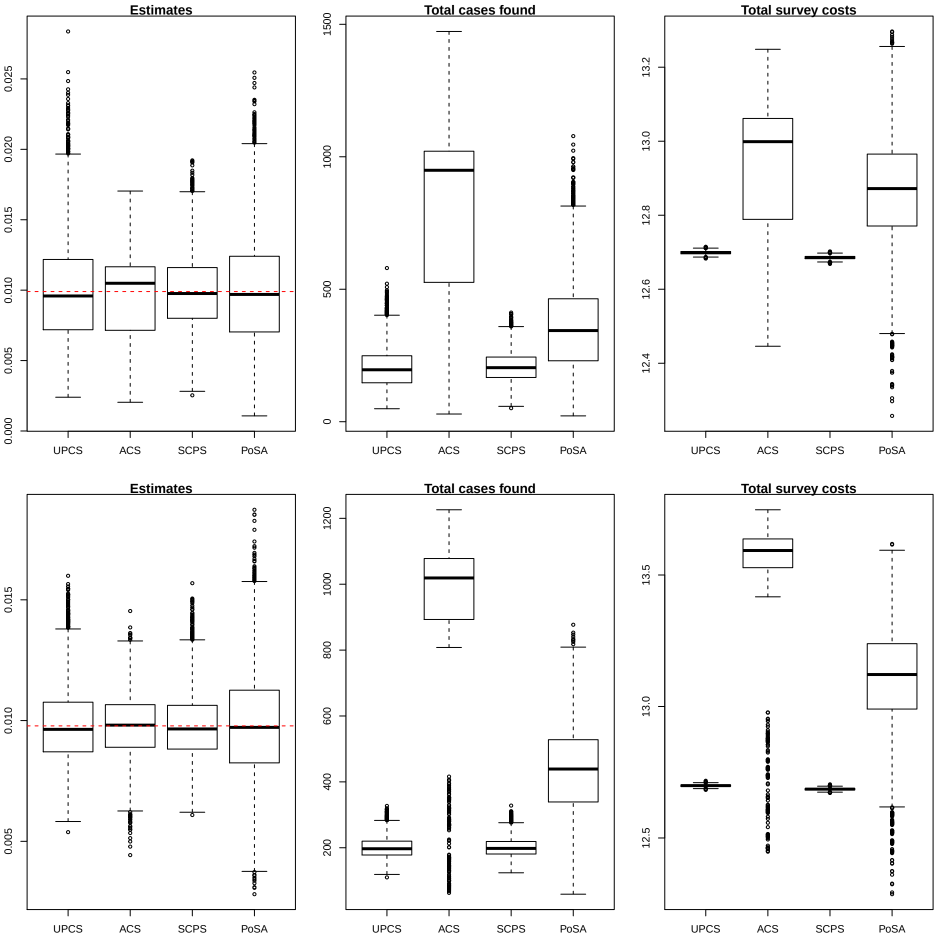

Figure 2.1: MC distribution of estimators, of number of cases found and of total survey costs for UPCS, ACS, SCPS and PoSA. Upper panel refers to populaion 1, while lower panel refers to population 2

the estimation stage is easy to implement and thanks to adaptivity it gives a larger number of detected cases as compared to traditional sampling designs. Moreover it requires a lower budget compared to ACS and there seems to be no large losses in efficiency compared to the traditional estimator. However the final sample size, that, as discussed, can potentially be as low as 0 and as large as N, is very variable hence adding instability to the total survey costs. The number of cases found is also very unstable, again due to the very large variation in the final sample size. A larger stability in the final survey costs may be achieved with a control over the final sample size although it is essential to keep the cases detection enhanced.

Our proposed PoSA strategy achieves the desirable design characterictis (enhancing detection rate and allows for logistic constraints) by maintaining a comparable efficiency at the estimation stage with respect to the traditional banchmark design. At the same time the randomness of sample size, its tendency to large variability with highly likely small and very small samples, appear as the main issues in urgent need of further developments. Therefore it seems reasonable to pursue the development of an integrated strategy, though controlling the variability in final sample size. In the next chapter we will thus address the variability in the sample size and we will propose a strategy able to control the final sample size.

Elementary Statistics	Population 1				Population 2			
	UPCS	ACS	SCPS	PoSA	UPCS	ACS	SCPS	PoSA
Min.	19923	13822	19913	7941	19918	13870	20012	10709
1st Qu.	20343	23202	20371	22062	20350	59085	20377	31014
Median	20440	30611	20468	25806	20443	63592	20468	36679
Mean	20442	29029	20469	25903	20443	62611	20469	36674
3rd Qu.	20541	33263	20566	29607	20536	66994	20562	42410
Max.	20954	41916	21006	46397	21056	75263	21066	66171

Table 2.2: Final sample size: elementary MC statitistics over 10000 runs

Chapter 3

Controlling the final sample size

In many sampling situations, such as in the case of TB prevalence surveys, a fixed sample size may be a desireable feature as, for instance, it helps the planning of survey costs. The proposed PoSA design enhances cases detection and deals with logistic constraints, as expected by the integration of adaptive and sequential designs, but it is characterised by a random sample size, with possibly large variability. In our inspirational example of TB prevalence surveys, as the budget needs to be fixed in advance, some sort of control over the final sample size is needed. In this chapter, we first present an adaptive strategy that controls the final sample size, highlighting its advantages and limitatations of application in the case of TB prevalence surveys, and, more in general of surveying a rare and clustered trait. Hence in the second part a proposal for controlling the final sample size in a PoSA design is illustrated.

3.1 Adaptive Web Sampling

One way to control the final sample size in adaptive designs is by using adaptive web sampling [23] (AWS), This strategy was proposed for sampling populations in network as well as in spatial settings. We are here interested in its spatial version.

As opposed to ACS, the encountered clusters (i.e. groups of units with a significant number of cases) do not need to be completely sampled, but may be sampled partially according to how deep into the high-values regions the sampler is interested in investigating. Differently from ACS, the AWS sampling procedure stops when the pre-fixed sample size is reached. Adaptive web sampling thus allows to concentrate the survey effort in the sub-areas considered of interest according to the observed survey values, while also continuing the selection in areas where no units of interest have yet been selected and controlling the final sample size.

In order to briefly illustrate AWS at both selection and estimation stage, we need to add some notation specifically for this sampling design. In addition to the set of $j = 1,...,M$ geographical areas and the associated values y_j, there is another variable of interest $w_{jj'}$ associated with any pair of areas j and j' $(j \neq j')$ that describes relationships between j and j'. We consider here the case in which $w_{jj'}$ is an indicator variable being equal to 1 if there is a link from area j to area j', e.g. if the two units are neighbours, and $w_{jj'} = 0$ otherwise. Moreover, let us define the active set α_t, as the set of sampled units for which the adaptive condition is satisfied together with their links to units that are not already in the sample at the t-step of the selection procedure.

The sampling procedure is as follows: first an initial sample $\mathbf{s}_0$ of size m_0 areas is selected by means of a traditional sampling design. If the variable of interest y_j for $j \in \mathbf{s}_0$ satsifies a chosen adaptive condition $y_j > y_{min}$, unit j together with y_j and all the positive links from j going to other units that are not already in the sample, are included in the initial active set α_0. The sampler can decide whether to continue by adding a set of units at the time or by adding one unit at the time. Here we give the details of the second choice, thus in order to conclude the sampling procedure there are $t = 1, ..., m - m_0$ additional steps, i.e. units to be selected. Once the initial sample is selected and the initial active set α_0 is defined, the next unit is selected from a mixture distribution, so that with a chosen probability p the unit is selected as one of the links out of the current active set α_0, and with probability $1 - p$ it is selected conventionally from the set of units that are not yet in the sample. The probability p is chosen large enough so that the probability of sampling links out from the active set is higher than the probability of sampling other population units. After each selection t the active set α_t is updated and the sampling procedure stops when all m units are selected. As mentioned, the adaptive selection can be made unit by unit or in waves. Selection can be said to occur in waves if the active set remains constant for several unit selections in a row, so that a whole group of selections is based on a given active set. Notice that the probability p can be used to seek a balance between spreading the sample out with placing it adaptively in the areas with high prevalence.

At step t, let the current sample at step be $\mathbf{s}_t = \cup_{j=0}^{t-1} \mathbf{s}_j$, let m_t be the number of units in the current sample and m_{α_t} be the number of units in the current active set α_t. Moreover let $w_{\alpha_{t+}}$ be the total number of links out from the active set α_t to units not in the current sample $\mathbf{s}_t$ and $w_{\alpha_t j}$ the number of links out from the active set to unit j. The probability that a unit j is selected in the t-step is defined as:

$$z_{tj} = p\frac{w_{\alpha_{tj}}}{w_{\alpha_{t+}}} + (1-p)\frac{1}{(N - m_{s_t})}$$

where $0 < p < 1$. This means that with probability p one of the links out from the current active set is selected at random, thus the area connected to it is included in the sample, and with probability $1-p$ the new sample unit is selected from the units not already in the sample.

Thus the AWS sampling design is defined as:

$$p(\mathbf{s}) = p_0 \prod_{j=1}^{M} \prod_{t=1}^{m-m_0} z_{tj}$$

where p_0 is the selection proability for the initial sample.

For the sampling setting above illustrated, different estimators have been suggested [23]. Among these, the first and simplest one is the estimator based on initial sample mean. With initial sample size m_0, let $\hat{\bar{y}}_{01}$ be an unbiased estimator for the population mean based only on the initial sample: $\hat{\bar{y}}_{01} = 1/M \sum_{j \in \mathbf{s}_0} y_j/\pi_j$, where π_j is the probability of including unit j in the initial sample according to the chosen sampling design. This estimator may be improved via the Rao Blackwell method, finding the conditional expectation of the preliminary estimator given the minimal sufficient statistic. The improved estimator is thus given by:

$$\hat{\hat{y}}_1 = E(\hat{\hat{y}}_{01}|d_r) = \sum_{\mathbf{s}:r(\mathbf{s})=\mathbf{s}} \hat{\hat{y}}_{01}(\mathbf{s})p(\mathbf{s}|d_r) \tag{3.1}$$

where d_r is the set of reduced data that is the set $d_r = \{(j, y_j, w_{j+}, w_{jj'}), j \in \mathbf{s}, j' \in \mathbf{s}\}$.

For the first estimator, with the initial sample being a random sample without replacement of m_0 units, the variance estimator of the initial sample mean is:

$$v(\hat{\hat{y}}_0) = \frac{(N - m_0)v_0}{(Nm_0)} \tag{3.2}$$

where v_0 is the sample variance of the initial sample.

Other unbiased estimators were proposed by Thompson such as an estimator obtained by dividing observed values by conditional selection probabilities that depend on the step-by-step active sets, and two ratio estimators based on the ratio between an estimator for M and an estimator for $\bar{y}$ based on the conditional probability. However they loose the simplicity in calculations that is here seeked as an alternative to the traditional design suggested, for instance, into the WHO guidelines.

Developments of the described selection procedure and their estimators have been suggested [24]. For instance, as already mentioned, the selection may be made in waves, or the selection of links to follow may be made with unequal probabilities (see [23] and [24] for further details).

With regards to the spatial applications of AWS, a recent development is Spatially Balanced Adaptive Web sampling [13] (SBAWS).

The idea is to combine a spatially balanced sampling procedure that spreads the sample over space together with an adaptive sampling, so that the sample is first spread as much as possible throughout space and secondly, if regions of interst are detected, they are further investigated via AWS. The spatially balanced design suggested for drawing the initial sample $\mathbf{s}_0$ is a sort of spatially stratified sampling, where no stratification variables are available and the strata are built arbitrarily in order to divide the study area into several parts of similar size. From each strata $h = 1, ..., H$ a simple random sample of m_{0h} units is the drawn; the remaining $m - Hm_{0h}$ units are selected step by step following the procedure of AWS. A modification to the estimator in Equation 3.1 is then proposed to allow for stratification. With regards to our inspirational example of TB prevalence surveys, guidelines do mention the possibility of using a stratified design, however for simplicity we do not consider stratified sampling in this first proposal of integration between sequentiality and adaptivity.

AWS manages to control the final sample size, thus to reduce costs as compared to ACS. However it does not allow for explicitly controlling logistic constraints, as desireable due to the sampling setting of developing countries. Moreover the estimators proposed for AWS, may be complicated to implement, limiting their actual use unless the estimation is limited to the initial sample. The improved estimators, obtained by using the Rao-Blackwell version of the estimators, may be implemented although this would require the listing of all possible samples. The listing of all possible samples may be avoided by using Monte Carlo Markov Chain methods, thus computationally resources consuming.

The pursue to use an integrated strategy, such as the PoSA design, thus seems reasonable, being easy to implement and accounting for logistic constraints. In the next paragraph we thus continue on with a proposal for controlling the final sample size in a PoSA design that meets the requirment of being easy to implement, dealing with logistic constraints and enhancing cases detection.

3.2 CPoSA sampling design

In real sampling situations, a fixed sample size is a desireable feature as, for instance, it helps the planning of survey costs. The PoSA design proposed in chapter 2 is able to over detect cases, naturally accomodates for a predefined route and it is easy to implement, but it is characterised by a large variability in the final sample size. As the proposed design meets most of the desirable features, in this section we consider again a Poisson-type design [26] but apply some sort of control over the final sample size. More specifically we found that often the sampling proposed PoSA procedure samples only few units. Thus in this section, as a first proposal to control the final sample size, we fix a minimum sample size in order to avoid too small sample sizes. Notice that in this context, not having an upper bound may be acceptable as long as the added units are only additional cases.

We call the proposed strategy Conditional Poisson Sequential Adaptive Sampling, CPoSA for short. CPoSA is still composed of a sequential component for dealing with logistic and cost constraints and of an adaptive component for enhancing cases detection.

Let us consider again a dichotomous study variable y being equal to 1 if the unit is a case and 0 otherwise. Given the population U ordered according to some auxiliary variable, units are visited following the chosen route. At the t-th step of the sequential selection, unit $i = t$ is/is not selected in the sample upon a Bernoulli trial. The inclusion probabilities of the remaining units are thus updated according to a chosen adaptive rule and also according to the number of units so far selected.

More formally, given the adaptive condition $y_i > 0$, i.e. $y_i = 1$, and a set of initial inclusion probabilities $\pi_1, ...\pi_i, ..., \pi_N$ with $\sum_{i=1}^{N} \pi_i = n$, at t-th step of the sequential selection, unit $i = t$ is certainly selected (i.e. is selected with probability 1) if the previous unit $i-1$ was selected and the adaptive condition on $i-1$ holds, otherwise it is selected with the latest updated probability $\pi_i^{(i-1)}$. Set, $\pi_i^{(0)} = \pi_i$ for $i = 1, ..., N$, the updated value of the inclusion probability for unit $i = t+1, ..., N$ at the generic step $t = 1, ..., N$ can be written as:

$$\pi_i^{(t)} = \begin{cases} 1 & \text{if} \quad i = t+1 \text{ and } s_{i-1} y_{i-1} = 1 \\ \pi_i^{(t-1)} - (S_{t-1} - E(S_{t-1})) w_{i-t}^{(t)} & \text{otherwise} \end{cases} \tag{3.3}$$

where in this first proposal $w_{i-t}^{(t)}$ is chosen so that $\sum_{i=1}^{N-t} w_{i-t}^{(t)} = 1$. In Section 2.1 we stressed that different choices for the weights leads sampling designs with different properties. In this setting, we chose to control the final sample size thus we chose to use weights with unitary sum. The difference with the updating procedure discussed in Section 2.1 is thus given by a different choice for the weights. For the sake of its simplicity, we considered the case in which $w_{i-t}^{(t)} = 1/(N-t)$. Notice that, with this specific choise, weights depend only on the row, meaning that on each row the inclusion probabilities are updated with the same $w_{i-t}^{(t)}$ for $i = t+1, ..., N$. The chosen updating procedure assures that the sample size given by $\sum_{i=1}^{N} \pi_i = n$ can only be exceeded if a cluster, i.e. a group of units geographically close and satisfying the given adaptive condition, is encountered. This means that, differently from PoSA, we expect the additional units to be cases only.

Analogously to Section 2.1, the distribution of the inclusion membership indicator S_i is $S_1 \sim Bernoulli(\pi_1) = Bernoulli(\pi_1^{(0)})$ for $i = 1$, while for $i = 2, \ldots, N$

$$S_i \sim Bernoulli(\pi_i^{(i-1)})(1 - y_{i-1}P(S_{i-1} = 1)) + y_{i-1}P(S_{i-1} = 1) \qquad (3.4)$$

where $\pi_i^{(i-1)}$ is given by the updating procedure as given in Equation 3.3.

The proposed CPoSA sampling procedure is synthetized in Algorithm 3.

```
Input: Ordered sequence of $N$ units & a vector of
  inclusion probs $\pi$.
Return: vector of sample membership indicators of size $N$
Procedure: Visit unit $i = 1$ and select with probability $\pi_1$.
  If $S_1 = 1$, collect $y_1$. Update inclusion probabilities for
  units 2-N.
for $t$ in $2 : N$ do
    point unit $i = t$ and select with probability $\pi_i^{(t-1)}$ if
      $S_t = 1$ then
        collect $y_t$
    end
    for $i$ in $(t+1) : N$ do
        if $y_t s_t = 1$ and $i = t+1$ then
            $\pi_i^{(t)}$=1
        else
            $\pi_i^{(t)} = \pi_i^{(t-1)} - (S_t - E(S_t)) w_{i-t}^{(t)}$
        end
    end
end
```

Algorithm 3: CPoSA Algorithm

The expectation of the conditional sample membership indicator is, for $i = 1$, $E(S_1) = \pi_1$ and for $i = 2, \ldots, N$:

$$E(S_i) = \begin{cases} 1 & \text{if} \quad s_{i-1}y_{i-1} = 1 \\ \pi_i^{(i-1)} & \textit{otherwise} \end{cases} \tag{3.5}$$

Notice that $\pi_i^{(i-1)}$ coincides with the value of the initial inclusion probability updated for individual i at the step before being selected, thus when $t = i - 1$.

The expectation of S_i under the CPoSA design can also be written as follows:

$$
\begin{aligned}
E(S_i) &= E(Bernoulli(\pi_i^{(t-1}))(1 - S_{i-1}y_{i-1}) + S_{i-1}y_{i-1}) \\
&= \pi_i^{(t-1)}E(1 - S_{i-1}y_{i-1}) + E(y_{i-1}S_{i-1}) \\
&= \pi_i^{(t-1)} - \pi_i^{(t-1)}E(S_{i-1})y_{i-1} + E(S_{i-1})y_{i-1}
\end{aligned}
$$

Estimation under CPoSA design

Similarly to PoSA, a trivially unbiased esimator for the population mean/prevalence $\bar{Y}$ can be easily derived:

$$\hat{\bar{Y}}_{CPoSA} = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i S_i}{E(S_i)} \tag{3.6}$$

In practice, the design weights to be used in Equation 3.6 can easily be identified and may be calculated according to the sample selected for all the population units (although it is not necessary). Table 3.1 represent the values to be used as design weights according to the selected sample.

Notice that $\pi_i^{(i-1)}$ is the value of the inclusion probability updated at step previously to making a decision about the inclusion the i-th individual. Thus the CPoSA design can be represented with the updating matrix given in chapter 1.4 and the updating prodecure as given in Equation 3.3. Notice that for $i < t = 1,...,N$, $E(S_i) = \pi_i^{(i-1)}$, while for $i = 1$, $E(S_1) = \pi_1$. It follows that, a ready-to implement

Table 3.1: deign weights in CPoSA estimator

s_{i-1}	y_{i-1}	design weight for unit i
0	0	$\pi_i^{(i-1)}$
0	1	$\pi_i^{(i-1)}$
1	0	$\pi_i^{(i-1)}$
1	1	1

formula for calculating an estimate for the population prevalence is given by:

$$\hat{\bar{Y}}_{CPoSA} = \frac{1}{N} \sum_{i \in \mathbf{s}} \frac{y_i}{\pi_i^{(i-1)}}$$

The variance for the CPoSA estimator has the same form of that given in Equation 2.13, however the computation of the second order inclusion probabilities needs further investigation.

3.3 CPoSA in the example of TB prevalence surveys

In the case of our inspirational example of TB prealence surveys, the sampling units are geographical areas and, once a geographical area is included in the sample, all the individuals that belong to that area are included in the final sample. In this section we present the extention of our proposed CPoSA design to the case of cluster sampling, where the sampling units are geographical areas. As noticed in chapter 2, the word "cluster" is here used as in traditional sampling, for indicating a group of individuals that are naturally grouped into the same primary unit. In order not to generate confusion, from now on we refer to areas or primary units, meaning the unit j, $j = 1,...,M$ that partition the population of N individulas $i = 1,...,N$. The geographical areas j are thus the primary sampling units.

Using the notation as given in Section 1.1, we consider the usual simplification, also inspired by the our inspirational example of TB prevalence surveys, in which $y_{ij} = 1$ if individual i in area j is a TB case and 0 otherwise. Thus the number of cases in unit j is $y_j = \sum_{i \in j} y_{ij}$. Let $y_{min} \in N$ be a threshold chosen such that the following (adaptive) condition qualifies j as a unit with a significant number of cases:

$$y_j = \sum_{i=1}^{N_j} y_{ij} > y_{min}$$

Notice that if we look at the prevalence instead of the total number of cases, we can refer to the prevalence for unit j as $\bar{y}_j = \sum_{i \in j} y_{ij}/N_j$. The threshold that qualifies unit j as a unit with a *significant* number of cases is thus $\bar{y}_{min} \in (0,1)$ and the adaptive condition is the following:

$$\bar{y}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} y_{ij} > \bar{y}_{min} \qquad (3.7)$$

In the case of geographical areas, we need to define an indicator I_j taking value 1 if the adaptive condition (3.7) holds in unit j, and 0 otherwise.

Moreover a set of initial inclusion probabilities $\pi_j^{(0)} = \pi_j$ is given for the ordered sequence of all units $j = 1,\ldots,M$, that may be proportional to the area size N_j.

At the t-th step of the sequential selection, unit $j = t$ is certainly selected (i.e. is selected with probability 1) if the adaptive condition (3.7) holds for the previous unit $j-1$, otherwise it is selected with probability $\pi_j^{(j-1)}$. The probability $\pi_j^{(j-1)}$ is the probability of selecting unit j at step $t = j$ and it is the initial inclusion probability opportunily updated at step $j-1$. More specifically, similarly to he previous section, the initial inclusion probabilities for unit $j = t+1,...,M$ are updated at each step t as follows:

$$\pi_j^{(t)} = \begin{cases} 1 \quad \textit{if} \quad j = t+1 \; \textit{and} \; s_t I_{t-1} = 1 \\ \pi_j^{(t-1)} - (S_t - E(S_t)) w_{j-t}^{(t)} \quad \textit{otherwise} \end{cases} \tag{3.8}$$

Algorithm 3 is thus slightly modified as showed in algorithm 4.

```
Input: Ordered sequence of $M$ units & a vector of
  inclusion probs $\pi$.
Return: vector of sample membership indicators of size $M$
Procedure: Visit unit $j = 1$ and select with probability $\pi_1$.
  If $S_1 = 1$, collect $y_1$ and $I_1$. Update inclusion probabilities
  for units 2-M.
for $t$ in $2:M$ do
    point unit $j = t$ and select with probabilty $\pi_j^{j-1}$ if
     $S_j = 1$ then
        collect $y_j$ and $I_j$
    end
    for $j$ in $(t+1):M$ do
        if $I_t s_t = 1$ and $j = t+1$ then
            $\pi_j^{(t)}$=1
        else
            $\pi_j^{(t)} = \pi_j^{(t-1)} - (S_t - E(S_t))w_{j-t}^{(t)}$
        end
    end
end
```

Algorithm 4: Cluster PoSA Algorithm

An unbiased estimator for the population mean/prevalence is then given by:

$$\hat{\hat{Y}}_{CPoSA} = \frac{1}{N}\sum_{j=1}^{M}\sum_{i=1}^{N_j} y_{ij}\frac{S_j}{E(S_j)} = \frac{1}{N}\sum_{j=1}^{M} y_j\frac{S_j}{E(S_j)} \tag{3.9}$$

where the design weights are given by:

$$E(S_j) = \begin{cases} 1 & \text{if} \quad s_{j-1}y_{j-1} = 1 \\ \pi_j^{(j-1)} & otherwise \end{cases} \tag{3.10}$$

where $\pi_j^{(j-1)}$ is the updated inclusion probability at step $t = j-1$ as given in Equation 3.3.

3.4 Preliminary empirical results on the CPoSA design

Following the structure of the previous chapters, we again show here some empirical evidence in order to compare the proposed strategy with the existing designs and stress its advantages and limitations. In particular we focus here on comparing UPCS, AWS, PoSA and CPoSA with respect to *(i)* total survey costs, *(ii)* ability to detect positive cases, *(iii)* sample size and *(iv)* estimators properties.

For the sake of illustration we show a small simulation study with the two populations presented in chapter 1 (Figure 1.2). Recall that the two populations considered share the same size $N = 200000$ individuals that are evenly distributed over a two dimensional space. A 15×15 grid was overimposed and produced a population of $M = 15^2 = 225$ areas to be sampled as primary sampling uinits, in a two stage design with inclusion into the final sample of all population (secondary) units inlcuded into every selected area. The study variable y is the usual binary variable indicating presence/absence of TB. The actual population TB prevalence $\bar{Y} \approx 0.01$ and we assumed to have a perfect guess of the true prevalence. The sample size is

$n = 19729$ individuals (see Section 1.2 for details on the calculations), leading to a sample of $m = 23$ areas. The adaptive condition for all adaptive designs was set to $y_j > N_j p_g$.

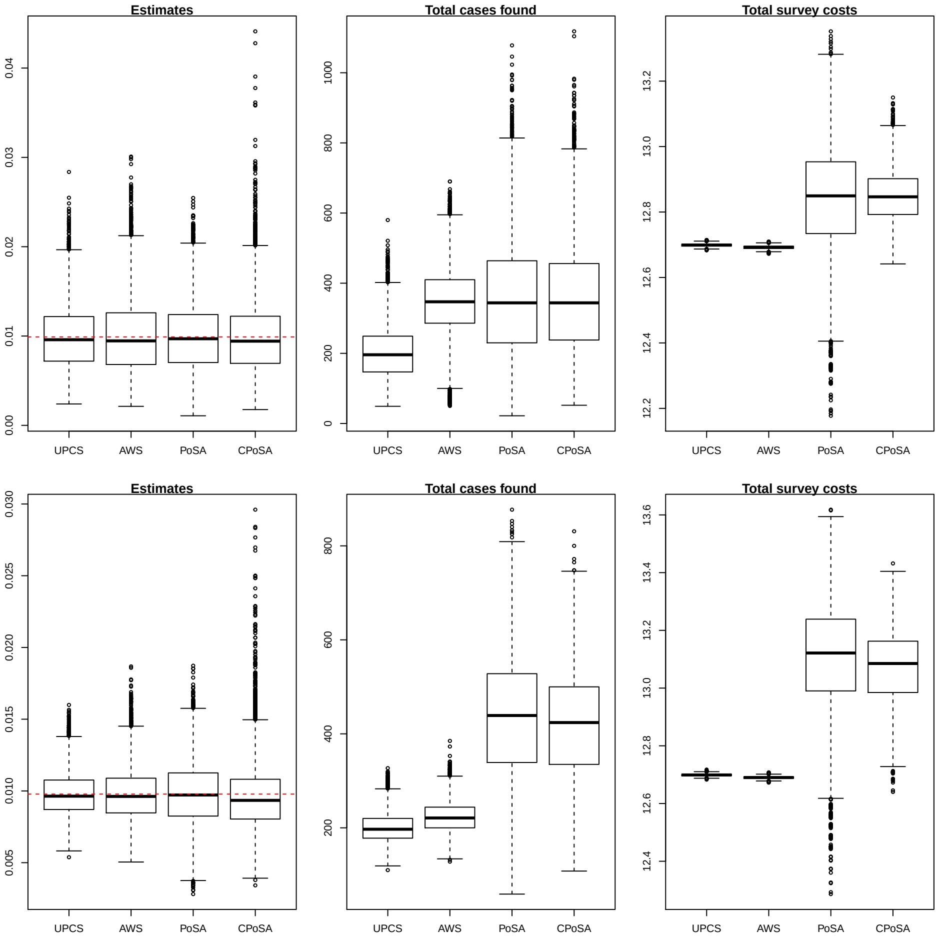

Figure 3.1: MC distribution of estimators (red line indicates the true population prevalence), of number of cases found and of total survey costs for UPCS, AWS, PoSA and CPoSA. Upper panel refers to populaion 1, while lower panel refers to population 2

Figure 3.1 synthetises the MC distribution of the estimators, number of cases and of the total survey costs (as calculated following cost function given in Equation 1.4). Top panel refers to population1 while bottom panel refers to population 2. Notice that, the proposed PoSA estimator seems not to loose efficiency as compared to the traditional estimator in population 1, although when the population is less clustered, as expected the proposed estimators seems to perform worst than the traditional ones. This is expected as the PoSA strategies are designed to work well for rare and clustered populations. It is remarkable to notice that the CPoSA estimator is characterised by a large number of outliers in the right end of its MC distribution. This is due to the fact that large samples in CPoSA are associated only with situations of large oversampling of cases. In fact, while the variability in the final sample size given by the PoSA procedure is due to both the fact that is based on Poisson sampling and that it allows for a flexible sample size, in CPoSA, if no cases are sampled, then the sample size is fixed. Moreover, notice that the proposed CPoSA procedure avoids the risk of sampling only few units as showed in both Table 3.2 and in Figure 3.1 by looking at the left end of the MC distribution of the estimators.

In terms of cases detection, although the sample size in CPoSA has been controlled (Table 3.2) the proposed procedure does not loose its ability of overdetecting cases as compared to PoSA. The CPoSA design is still a sampling design with random sample size but by introducing a minimum sample size, it manages to lower the variability in the final survey costs and in the final sample size as compared to PoSA. As compared to AWS, CPoSA is simpler to implement, is able to spot more cases and to account for logistic constraints, but AWS performs better in terms of costs control, being a design with fixed sample size.

Elementary Statistics	Population 1				Population 2			
	UPCS	ACS	SCPS	CPoSA	UPCS	ACS	SCPS	CPoSA
Min.	19923	19928	7941	19241	19918	19936	10709	19227
1st Qu.	20343	20456	22062	23990	20350	20403	31014	31269
Median	20440	20568	25806	25879	20443	20497	36679	35621
Mean	20442	20569	25903	26125	20443	20501	36674	35345
3rd Qu.	20541	20680	29607	27936	20536	20595	42410	39318
Max.	20954	21143	46397	38641	21056	21088	66171	54520

Table 3.2: Final sample size: elementary MC statitistics over 10000 runs

Chapter 4

Empirical evidence

In the previous chapters we discussed the main reasons for the development of a strategy that combines sequentiality and adaptivity, and showed its performance in some preliminary simulation studies. We compared them with ACS, AWS, SCPS and the traditional UPCS design, highlighting advantages and limitations of all strategies. This chapter aims at comparing the two proposed strategies, namely PoSA and CPoSA, with the traditional UPCS design currently in use by WHO by means of an extensive simulation study. The main features of the designs as well as the behaviour of the estimators are compared under different scenarios aiming at reproducing realistic TB prevalence survey situations. As noticed in the preliminary simulative results shown at the end of chapters 1, 2 and 3, the design properties as well as estimators' heavily depend on the distribution of the study variable y. Thus the different simulated scenarios are first described in details and secondly relevant simulation results are showed.

4.1 Objective of the simulation study

The main objective of the simulation study presented in this chapter is to provide empirical evidence of the advantages of the integration of adaptive and sequential procedures. Using the design implemented by WHO as a benchmark, we thus compare the proposed sampling strategies with UPCS with respect to:

(i) *ability to detect cases*: enhanced detection rate is a desirable feature in the inspirational example of TB prevalence surveys, since every found case, will be treated. Notice that enhanced case detection may also allow to better study the risk factors of TB, hence having a better understanding of its epiedmiology, and it may allow for subnational estimates.

(ii) *survey costs*: reducing survey costs means being able to find as many cases as possible with the same budget

(iii) *estimators properties*: all proposed estimators are unbiased, hence the focus is on evaluating their efficiency as compared to the traditional ones

In the following sections we investigate different scenarios in order to show how the population and the design characteristics combine and influence the three features under discussion.

4.2 Preliminary simulation results

In adaptive sampling, results on the estimators' properties as well as on the ability to over detect the trait of interest, heavily depend on the distribution of the study variable y, more specifically on its homogeneity throughout the geographical areas. In order to describe the homogeneity/dishomogeneity of the distribution of the study variable, we can use the coefficient of between areas variation k as defined in Section 1.2. When the study variable is unevenly distributed over space (k is high), adaptive designs tend to perform well with any kind of adaptive condition. For instance, in the extreme situation in which cases are located only within clusters and there are no positive cases outside (that is the case in which $y_j > 0$ in the clusters and $y_j = 0$ anywhere else), sampling procedures based on any adaptive condition $y_j > y_{min}$ with $y_{min} > 0$ are able to discriminate whether the sampler is visiting a cluster or not. The sample size would thus increase as compared to non adaptive designs, but in all added areas the prevalence is likely to be high and, unless clusters are very large, the sample size only moderately increases. On the other hand, in a situation in which the area specific prevalences are very similar, yielding very small values of k, only a deep knoweldge of the distribution of the study variable and thus an accurate choice of the the threshold y_{min} would make the procedure able to discriminate between clusters and non-clusters. If the threshold for the adaptive condition is chosen accurately, only few areas may be added, otherwise adaptive designs without a fixed sample size may include too many additional units.

It is thus reasonable to combine parameters that influence the population characteristics (population parameters) and those that influence the way the sampling design itself is carried out, such as the threshold y_{min} for adaptivity (design parameters). These are:

1. the population size (N)

2. the population prevalence ($\bar{Y}$)

3. the number of clusters (B), that is the number of groups of areas with a higher prevalence of cases compared to the rest of the population

4. the size of the grographical areas (N_j) that is defined by the refinement of the grid

5. the clusters' prevalence, that is defined by the % of cases within clusters, i.e. $q = \sum_{b=1}^{B} \sum_{j \in b} \sum_{i=1}^{N_j} y_{ij} / \sum_{j=1}^{M} \sum_{i=1}^{N_j} y_{ij}$ (notice that the higher q, the higher $\bar{y}_j, j \in b$)

6. the threshold y_{min}

The population size was found to be uninfluential, thus it was arbitrarily set to $N = 250000$. Parameters *(2)-(5)* combine together and may be summarised in the coefficient of between areas variation k. Specifically, their combinations influence the homogeneity/dishomogeneity of the study variable y over the population, thus they contribute to generating populations with different k as shown in Figure 4.1. We fixed 3 prevalence levels ($\bar{Y} = 0.01$, $\bar{Y} = 0.005$, $\bar{Y} = 0.001$), 4 levels of clusters ($B = 3$, $B = 6$, $B = 10$ or $B = 20$ clusters), 3 levels of refinement of the grid (10×10, 15×15, 25×25 grid) and 5 levels of clusters' prevalence ($q = 0\%$, $q = 20\%$, $q = 40\%$, $q = 60\%$, $q = 80\%$). Fixing the prevalence at 0.005 for instance, the coefficient of between areas variation increases as the overimposed grid gets thinner, i.e. as the geographical areas are smaller. Fixing the percentage of cases within clusters, the prevalence between the geographical areas is less variable as the number of clusters generated gets higher. In general, clusters' prevalence, positively influences k, while the other parameters negatively influence k.

We found that the coefficient of between areas variation k is in the range $0.1876 - 3.4$, reaching as low as 0.1876 when the geographical areas are large ($N_j \approx 2500$ individuals $j = 1, ..., M$, $M = 100$) and when the population prevalence is high (third panel of Figure 4.1). The maximum is reached with a very thin overimposed grid, i.e. geographical areas of size about $N_j \approx 400$ individuals $j = 1, ..., M$ ($M = 625$), with low prevalence and with few clusters ($B = 3$) with $q = 80\%$.

Parameter *(6)* is a design parameter, meaning that it only influences the way the design is carried out and thus does not contribute to generating different populations. In adaptive designs, the choice of a good threshold for the adaptive condition is essential for being able to discriminate when a cluster is encountered and when it is not, especially when k is low. If the chosen threshold is too low, the number of additional units added by adaptive designs is very large, inflating survey costs and possibly detecting only few additional cases. On the other hand the risk of choosing a too high threshold y_{min} is that clusters may be skipped thus not being able to oversample cases. We may expect that *(6)* affects the ability to detect cases as well as the final survey costs and the efficiency of the estimators.

The following two different thresholds y_{min} for the adaptive condition are chosen:

(a) $y_{min} = \bar{Y}$

(b) $y_{min} = max\{\bar{Y}/2, min_j\{\bar{y}_j\}\}$

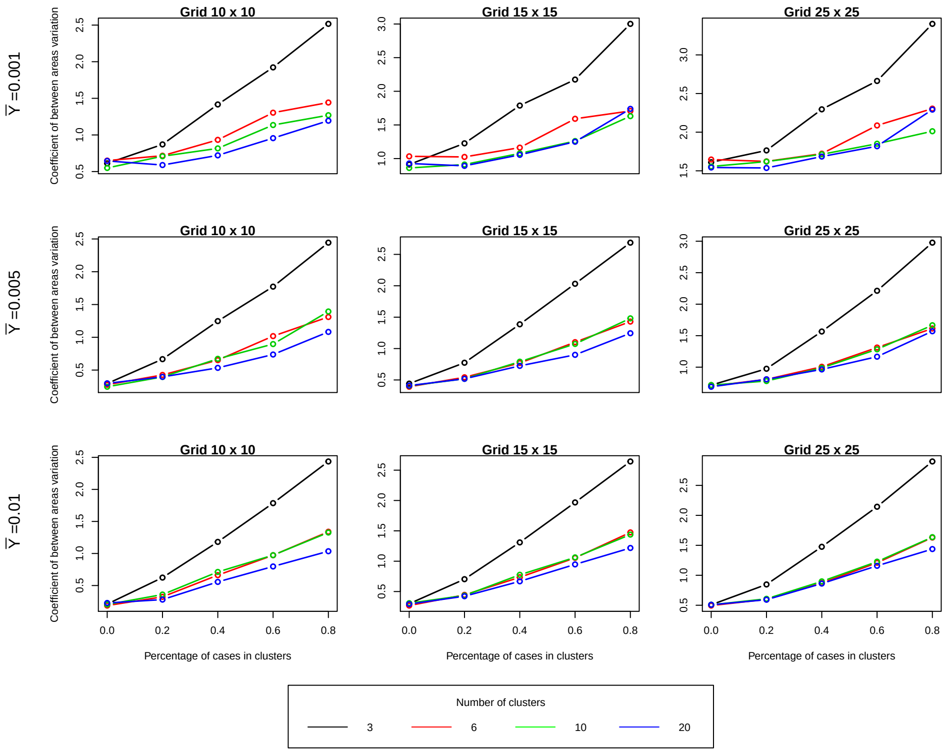

Figure 4.1: The coefficient of between areas variation with different levels of population prevalence $\bar{Y}$, number of clusters, percentage of cases within clusters and refinement of the grid

As mentioned above, we expect that when the between areas variation is low, there is a large difference between the sample size given by *(a)* and that given by *(b)*, while when the between areas variation is high, the choice of any of the above thresholds is equivalent.

The combinations of the described parameters yield 180×2 scenarios. We carried out some preliminary simulations in order to select the parameters that are actually influential and those that are not. A major simulation parameter is the coefficient of between areas variation k, as well as the threshold y_{min} for the adaptive condition. In the next paragraph we give details about the choice of the scenarios and of the population parameters combination.

4.3 Simulation scenarios

Out of all the possible scenarios we fixed $N = 250000$ individuals, $\bar{Y} = 0.005$ and $B = 3$, that means that cases are concentrated in three main clusters. The value for the prevalence and for the number of clusters were chosen because they seem the combination that better allows for variations in k as shown in Figure 4.1. We focused on 6 levels of coefficient of variation, specifically $k = \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$. The six populations are represented in Figure 4.2 and details about them are summarised in Table 4.1. All populations have $B = 3$ clusters except for population 1 where there are no clusters. Populations 2-6 have an increasing percentage of cases within clusters up to population 6 where 80% of cases are located in the three clusters.

The populations were entirely built in the R environment using the library spatstat. Once the desired prevalence $\bar{Y}$ was set, the number of positive cases $\bar{Y}N$ to be allocated in the two dimensional space was defined and positioned in space as follows: first, the centers of the B clusters were assigned coordinates, i.e. were randomly positioned in the two-dimensional space; secondly, the chosen percentage of positive cases was equally assigned within B circles with center assigned in the previous step and with a predefined radius; the remaining $\bar{Y}N(1-q)$ cases were then randomly assigned other coordinates. Last, a regular grid was overimposed over the popula-

Population	$\bar{N}$	**q**	**k**
Population 1	1111	0%	0.5
Population 2	400	20%	1.0
Population 3	400	40%	1.5
Population 4	1111	60%	2.0
Population 5	2500	80%	2.5
Population 6	400	80%	3.0

Table 4.1: The six populations considered in the simulation study, listed together with their main features: average area size $\bar{N}$, percentage of cases located inside the clusters q and coefficient of between areas' prevalence variation k

tion yielding M geographical areas of average size equal to $\bar{N}$ individuals.

Since all estimators are analitically unbiased, we controlled the Monte Carlo error with their empirical bias. We fixed the number of simulations runs to 5000 iterations, which guaranteed the MC bias to be $< 0.5\%$ for all the 3 estimators at any level of k.

Referring to the three main objectives given in Section 4.1, the performance of the proposed strategies was evaluated using the following MC measures of empirical performance:

(i) *ability to detect cases*: it is evaluated with the ratio of the number of cases found with the PoSA/CPoSA sampling strategy over the cases found with UPCS,

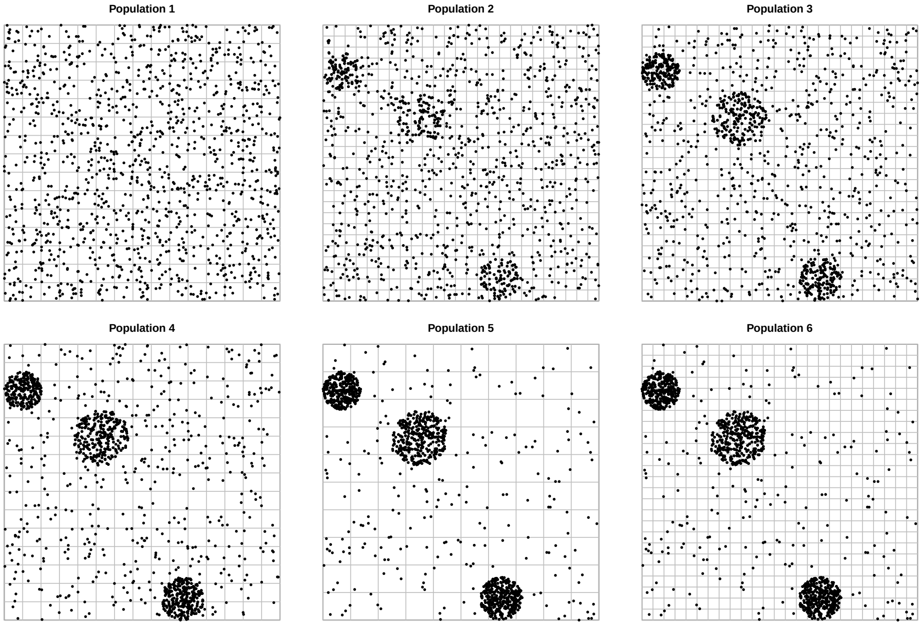

Figure 4.2: The six populations considered in the simulation study; $\bar{Y} = 0.005$, dots represent cases

$$\frac{\sum_{j \in \mathbf{s}_{PoSA,r}} y_j}{\sum_{j \in \mathbf{s}_{UPCS,r}} y_j}; \qquad \frac{\sum_{j \in \mathbf{s}_{CPoSA,r}} y_j}{\sum_{j \in \mathbf{s}_{UPCS,r}} y_j} \qquad r = 1, ..., 5000$$

Moreover, in order to account for the sample size, thus making sure that our method is able to add units only if they have a significant number of cases, we considered the *net ability to detect cases*:

$$\frac{\sum_{j \in \mathbf{s}_{PoSA,r}} y_j / n_{PoSA,r}}{\sum_{j \in \mathbf{s}_{UPCS,r}} y_j / n_{UPCS,r}}; \qquad \frac{\sum_{j \in \mathbf{s}_{CPoSA,r}} y_j / n_{CPoSA,r}}{\sum_{j \in \mathbf{s}_{UPCS,r}} y_j / n_{UPCS,r}} \qquad r = 1, ..., 5000$$

(ii) *survey costs*: as survey costs linearly depend on the sample size, we take into account both final costs and final sample size. Thus, we give three MC measures that are:

(a) *final survey costs*

$$\frac{C_{PoSA,r}}{C_{UPCS,r}}; \qquad \frac{C_{CPoSA,r}}{C_{UPCS,r}} \qquad r = 1, \ldots, 5000$$

where C is the final survey costs and is calculated as follows:

$$C = c_0 + c_1 m + \sum_{j \in \mathbf{s}} N_j c_2 \tag{4.1}$$

where c_0 is the fixed survey cost (c_0 = 100000 dollars), c_1 is the survey cost for each sampled geographical area (c_1 = 1000), m is the number of geographical areas sampled, and c_2 is the survey cost for each individual within a sampled geographical area (c_2 = 10 dollars). Moreover, a 20% discount was applied to sequential designs (PoSA and CPoSA) for the planning of the route.

(b) *final sample size*

$$\frac{\sum_{r=1}^{5000} n_{PoSA,r}}{\sum_{r=1}^{5000} n_{UPCS,r}}; \qquad \frac{\sum_{r=1}^{5000} n_{CPoSA,r}}{\sum_{r=1}^{5000} n_{UPCS,r}} \qquad r = 1, \ldots, 5000$$

(c) *cost per detected case*, that is the cost for spotting one single case

$$\frac{C_{PoSA,r}/\sum_{j \in \mathbf{s}_{PoSA,r}} y_j}{C_{UPCS,r}/\sum_{j \in \mathbf{s}_{UPCS,r}} y_j}; \qquad \frac{C_{CPoSA,r}/\sum_{j \in \mathbf{s}_{CPoSA,r}} y_j}{C_{UPCS,r}/\sum_{j \in \mathbf{s}_{UPCS,r}} y_j} \qquad r = 1, \ldots, 5000$$

(iii) *stability*: it is evaluated through an efficiency comparison, that is the ratio of the root mean squared error (RMSE) of the PoSA/CPoSA sampling strategy over the root mean squared error (RMSE) given by UPCS

$$\frac{\sqrt{\sum_{r=1}^{5000}[\hat{\bar{Y}}_{PoSA,r}-\bar{Y}]^2}}{\sqrt{\sum_{r=1}^{5000}[\hat{\bar{Y}}_{HT,r}-\bar{Y}]^2}}; \qquad \frac{\sqrt{\sum_{r=1}^{5000}[\hat{\bar{Y}}_{CPoSA,r}-\bar{Y}]^2}}{\sqrt{\sum_{r=1}^{5000}[\hat{\bar{Y}}_{HT,r}-\bar{Y}]^2}}$$

4.4 Results

According to WHO we assumed a 0.5 coefficient of between areas' prevalences variation thus determined the required sample sizes in all populations (see Chapter 1.2 for further details), hence the number of areas to be sampled is calculated and the chosen sample size according to the two scenarios.

Figures 4.3-4.5 refer to the design features and show the MC first quartile, median and third quartile of the above described quantities.

The ability to detect cases (Figure 4.3) in PoSA and CPoSA is nearly equal, meaning that the constraint on the sample size does not reduce detectability. In both proposed sampling strategies the detection power is increased as compared to the traditional WHO design. In fact they are able to detect, for all levels of between areas variation and for all considered scenarios, at least an additional 20% of cases. In all scenarios it seems that as the variability between the areas specific prevalence increases, the ability to detect cases as compared to UPCS is more variable. In fact, in populations 5 and 6, where the three clusters contain 80% of the population cases, the proposed designs either detect no additional cases compared to

UPCS (first quartile) or nearly all cases (third quartile). The peak in the number of cases detected by PoSA and CPoSA procedure in population 1 with threshold *(b)* (top right panel), is due to the fact that many areas satisfy the adaptive condition and thus many additonal areas are included in the final sample. However, when $k < 1$ the adaptive process with a too low threshold $\bar{y}$ fails at spotting areas with a significant number of cases, as the percentage of cases found in the sample is very similar to that given by UPCS. Notice that, as previously mentioned, the ability to over detect cases is essential not only for treating as many cases as possible, but also for allowing subregional estimates.

CPoSA was introduced to reduce the variability in the final sample size that seemed the largest limitation of PoSA. More specifically with PoSA design, we highlighted the risk of sampling only few units, thus we concentrated on setting a lower bound for the final sample size. In Figure 4.4, a comparison between the MC distribution of the sample size of CPoSA and PoSA is shown as well as their comparison with UPCS. Although the medians of the final sample size in PoSA and CPoSA seem to overlap, the distribution of the sample size in CPoSA appears to be less variable, with the third and first quartile being closer to the central part of the distribution compared to PoSA. By using the CPoSA design we manage not only to avoid the problem of an extremely low sample size, but its use also seems to reduce the chance of samping too many areas. Moreover, by construction, every time the minimum sample size is exceeded, the sampling procedure adds geographical areas with a large num-

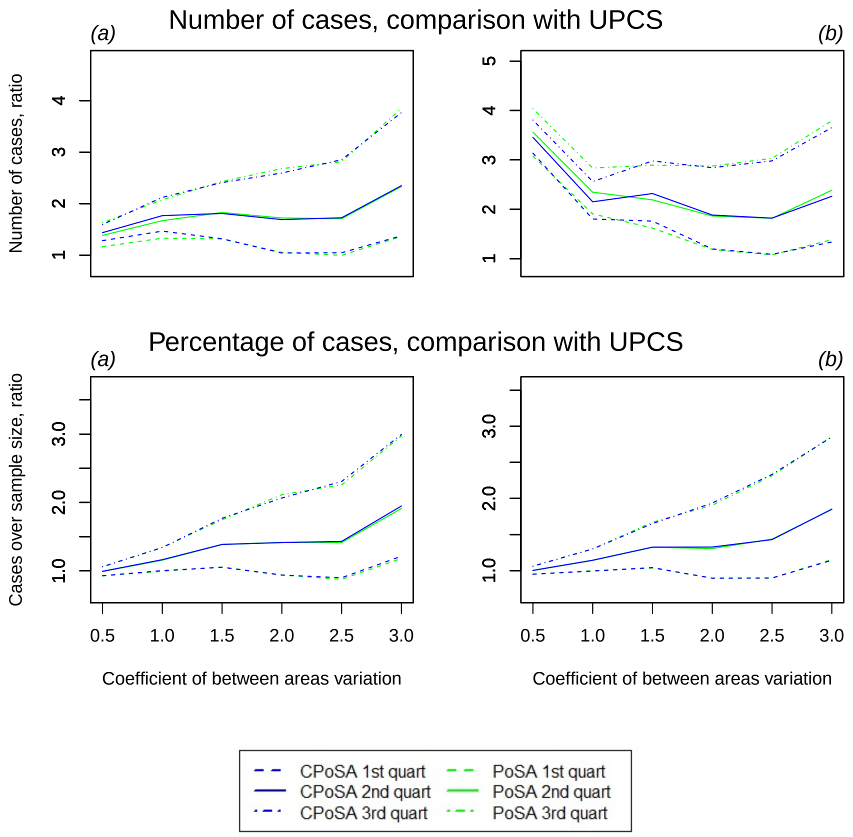

Figure 4.3: Measures of cases detection. Left panels refer to scenario *(a)* and right panels refer to scenario *(b)*. Top two panels show the MC distribution of the number of cases ratio and the bottom two panels refer to the the MC distribution of the net detected cases ratio.

ber of cases. In fact, whereas the sample size in PoSA is variable even when clusters are not encountered, CPoSA allows for increasing the sample size only if a cluster is encountered.

In all proposed scenarios, the final sample size is always greater than that used by WHO. The choice of a good adaptive condition and

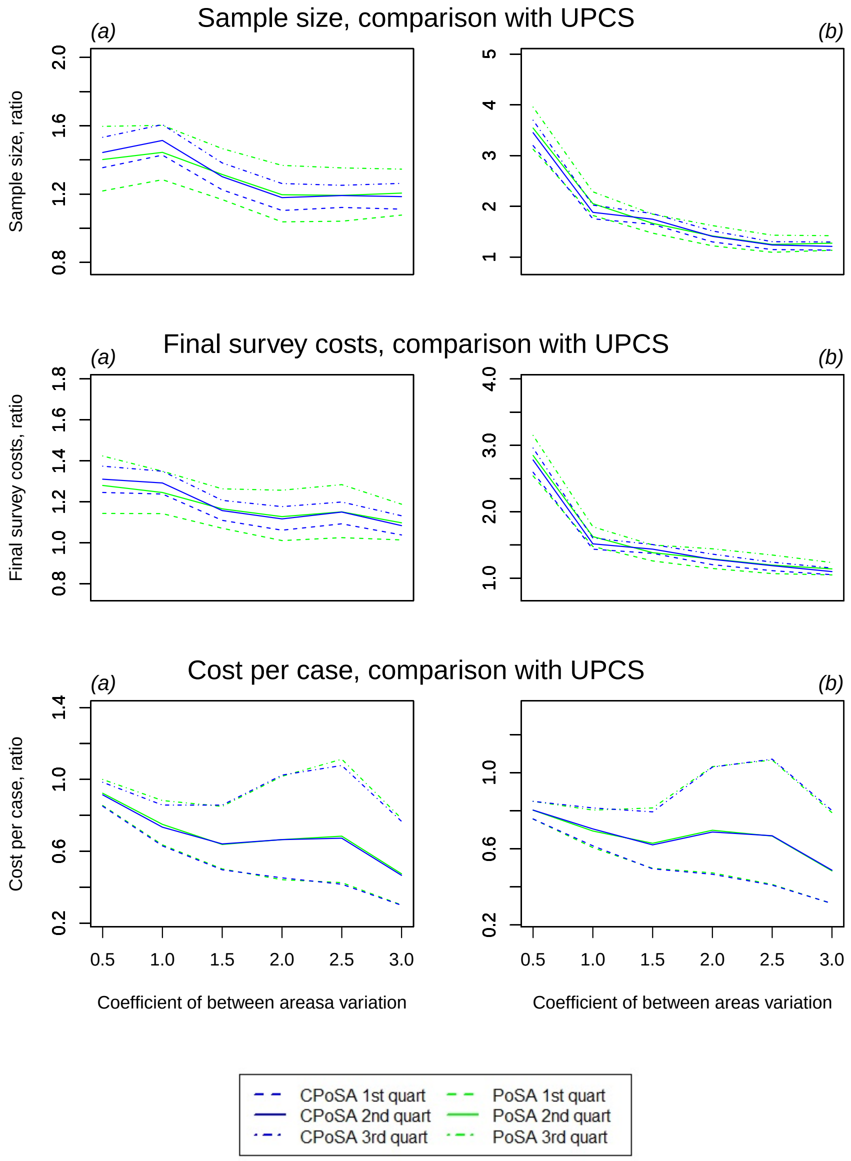

Figure 4.4: Measures of cost. Left panels refer to scenario *(a)* and right panels refer to scenario *(b)*. Top two panels show the MC distribution of sample size ratio, the central two panels refer to the the MC distribution of total survey costs ratio, the bottom two panels refer to the MC distribution of the cost per spotted case ratio.

the planning of a route that minimises survey costs is thus essential in order to improve the probability that the additional sampled units are cases. The peak in the sample size produced by the proposed designs in population 1 given a lower threshold y_{min} (top left panel), is due to the fact that many areas satisfy the given adaptive condition. However, as noticed for the ability to detect cases, the adaptive procedure fails at spotting areas with a significant number of cases.

Final survey costs linearly depend on the final sample size as shown in equation 4.1. The variability in the final survey costs given by the CPoSA design is smaller than the variability given with the PoSA design, due to a smaller variability in the final sample size. As compared to UPCS, if k is very low, there is no costs reduction, while as the between area variation increases there is a larger reduction in final survey costs. The increase in survey costs is however compensated by a gain in spotted cases, as shown in Figure 4.3. The cost reduction should thus be commented together with a measure that takes into account the number of spotted cases. In the bottom panel of Figure 4.4, we notice that the cost required to find one case is always lower than or equal to that of UPCS. From the top panel of Figure 4.4 we noticed that the final sample size found with the proposed designs is always higher than that found with the traditional WHO design, although only when $k \geq 1$ the added areas have a large number of cases (Figure 4.3). The cost to spot one case is, however, equal or lower than that required when the sampling design used is the traditional UPCS. Notice that, again, the cost to spot a case with our proposed strategies is more variable as the variability between areas increases. Last, notice that although variability increases as k gets higher, the peak coincides with the situation where high k is combined with a large size of the geographical areas ($k = 2.5$ coincides with $\bar{N} = 2500$). This suggests that the choice of the areas' size should be taken into account in the survey planning. More specifically, the choice of smaller areas' size may help reducing the variability in the final costs as well as helping to spot more cases with the same budget. Further research may thus address the choice of the areas' size.

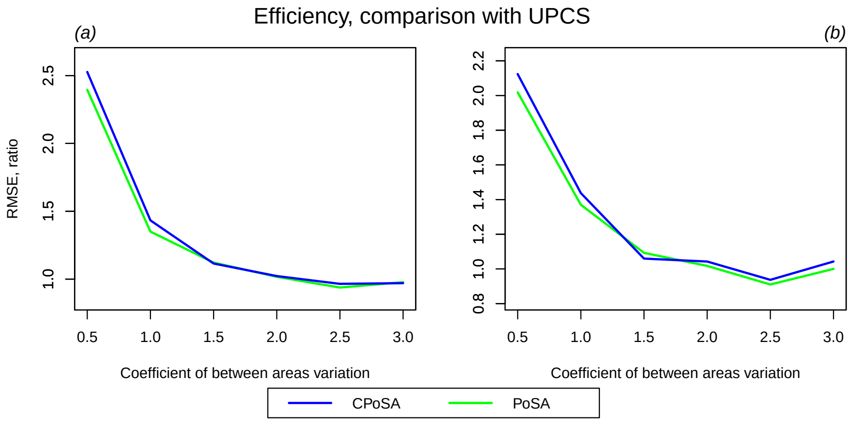

Figure 4.5: MC root mean squared error of PoSA and CPoSA estimates over MC root mean squared error of UPCS. Left panel refers to scenario *(a)* and right panel refers to scenario *(b)*

Last, Figure 4.5 contains the comparison of the MC root mean squared error of the proposed estimators is compared with the traditional HT estimator. Results are presented as the ratio of the MC root mean squared error of PoSA/CPoSA design over that of UPCS. For $k \leq 1$ there seems to be a loss in stability in all scenarios. However as the variability in the between areas prevalences increases the final estimator seems to become more stable as compared to the traditional HT estimator and there are no losses of efficiency in any of the considered scenarios. This suggests that in populations for which $k \geq 1.5$, PoSA and CPoSA meet the features described as desirable (enhanced case detection, lower costs and ability to account for logistic constraints), the produced estimators do not lose in stability and they are also easy to implement.

Choosing a different population prevalence or a different number of clusters does not change the overall results on cases, costs and efficiency, but it may affect the variability of the results. Figure 4.6 represents the discussed MC measures for the threshold *(a)*, $\bar{Y} = 0.005$ and $B = 6$, while Figure 4.7 shows the MC measures for the threshold *(a)*, $\bar{Y} = 0.01$ and $B = 3$. With a larger number of clusters and equal prevalence, the behaviour of the performance measures is the same as discussed above, but the variability in all the MC measures of performance seems lower. If there are more clusters in fact, the distribution of the MC measures are more likely to be close to the central part of the distribution, lowering the variability. For instance, the number of cases found is less variable when cases are concentrated within 6 clusters other than when they are only in 3 clusters. On the other hand, with a higher prevalence and the same number of clusters there seems to be no difference in terms of performance of the proposed strategies.

So far we presented results on 6 populations combined with 2 dif-

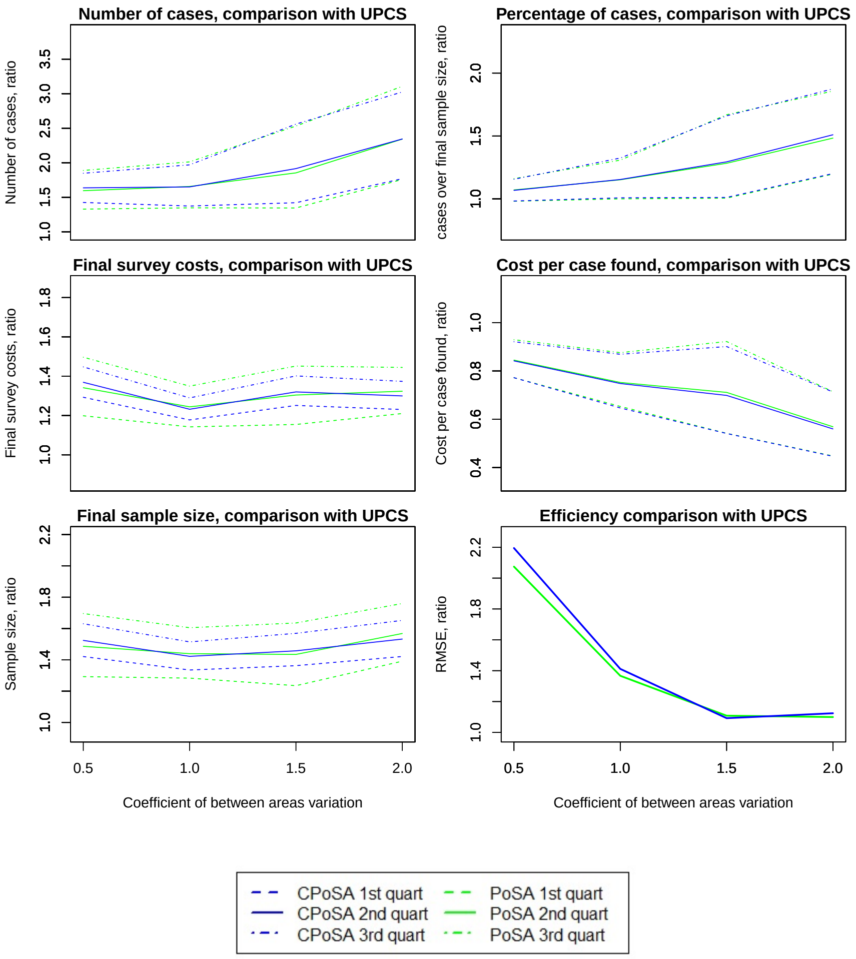

Figure 4.6: MC performance measures for 4 populations with $N = 250000$, $\bar{Y} = 0.005$, $B = 6$, $k = \{0.5, 1.0, 1.5, 2.0\}$ and scenario *(a)*. From top left to bottom right, MC distribution of: ratio of the number of detected cases, ratio of net detected cases, ratio of sample size, ratio of the total survey costs, ratio of the cost per spotted case, root mean squared error of PoSA and CPoSA estimates over MC root mean squared error of UPCS

ferent design parameters *(a)* and *(b)* (Figures 4.3-4.5), 4 populations combined with only one design parameter *(a)* (Figure 4.6) and then

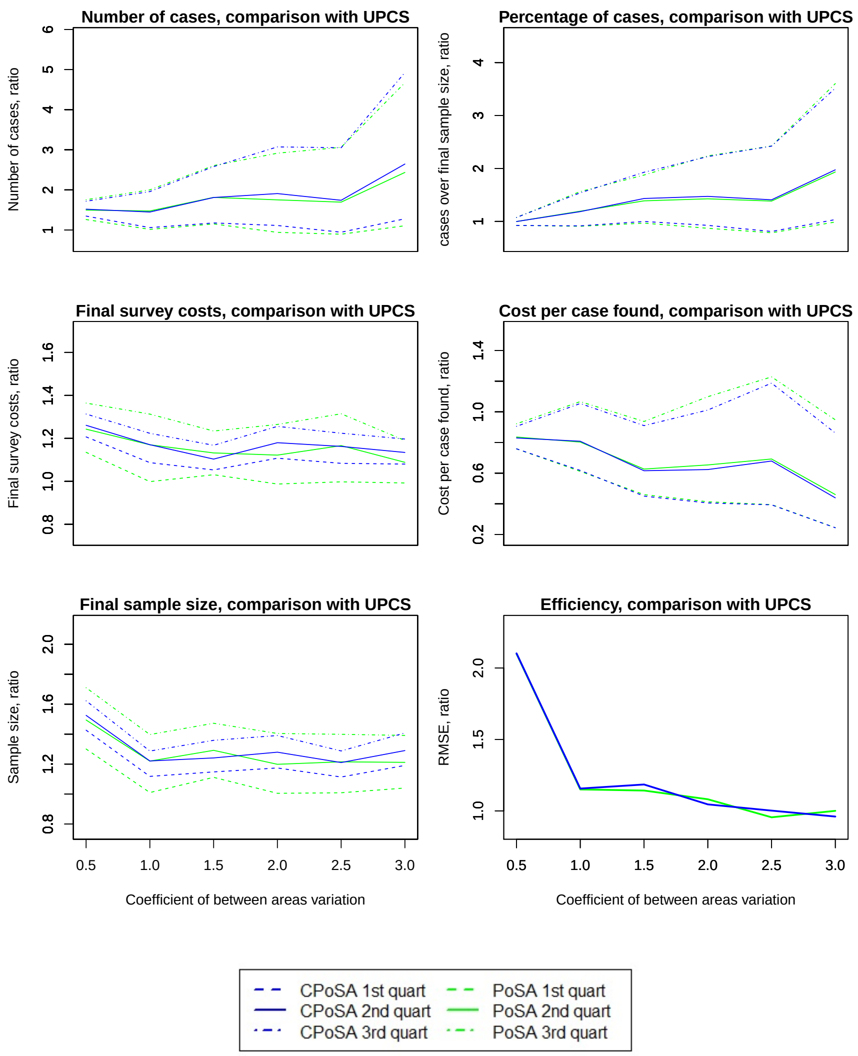

Figure 4.7: MC performance measures for 6 populations with $N = 250000$, $\bar{Y} = 0.01$, $B = 3$, $k = \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$ and scenario *(a)*. From top left to bottom right, MC distribution of: ratio of the number of detected cases, ratio of net detected cases, ratio of sample size, ratio of the total survey costs, ratio of the cost per spotted case, root mean squared error of PoSA and CPoSA estimates over MC root mean squared error of UPCS.

on 6 populations combined with *(a)* (Figure 4.7). Moreover simulations on the initial 6 populations combined with four additional design parameters (relative to the prevalence guess) were ran but are not presnted as results are very much similar to those in Figures 4.3-4.5. The MC performance measures are calculated on the 46 different scenarios and summary statistics are presented in Tables 4.2 and 4.3.

The final number of cases found with PoSA design is often larger than that found with the traditonal UPCS design, although it is fairly variable. The quartiles for the final number of cases found with the PoSA design over UPCS are very similar to those of the ratio of the number of cases found with CPoSA over the traditonal design. This means that although controlling the final sample size, the CPoSA design does not lose the ability to over-detect cases. In fact the minimum sample size is only exceeded when areas with a large number of cases are encountered.

In the preliminary simulation results given in Chapter 2, we highlighted the risk of sampling only few units when using the PoSA strategy. The aim of CPoSA of ensuring a minimum number of areas to be sampled is perfectly attained as shown in the first panel of Figure 4.4. Moreover we notice here that in general the CPoSA variability in the final sample size, as measured with a coefficient of variability, is reduced as compared to PoSA. In terms of efficiency, averaging the results on all populations shows that the average loss of stability is of about 35%. Thus together with the large loss of stability registered at $k = 0.5$ there seems to be a gain in stability when $k \geq 2$ while for the intermediate levels the PoSA estimator performs

Statistics	Detected cases	Survey costs	Sample size	Efficiency
1st quart.	1.27	1.08	1.14	1
median	1.78	1.21	1.34	1.09
3rd quart.	2.56	1.39	1.59	1.37
average	2.16	1.30	1.45	1.27
cv	0.85	0.31	0.37	0.34

Table 4.2: PoSA: Detected cases ratio, total costs ratio, sample size ratio and effiiciency ratio. Elementary statistics over 46 simulated populations

similarly to UPCS.

Statistics	Detected cases	Survey costs	Sample size	Efficiency
1st quart.	1.31	1.11	1.18	1.02
median	1.77	1.21	1.33	1.08
3rd quart.	2.53	1.34	1.52	1.43
average	2.16	1.29	1.44	1.35
cv	0.69	0.27	0.32	0.45

Table 4.3: CPoSA: Detected cases ratio, total costs ratio, sample size ratio and effiiciency ratio. Elementary statistics over 46 simulated populations

Chapter 5

Conclusions and research perspectives

In this thesis a new sampling strategy is proposed for sampling a rare and clustered population under both cost and logistic constraints. It is motivated by the example of national TB prevalence surveys, promoted by WHO in countries with a high TB-burden. The characteristics of the trait to sample (rare and clustered) in combination with the pecularities of the countries (mainly located in developing areas, hence not all areas are equally accessible) motivate the need for a non-traditional sampling design. In particular it may be desirable to use a sampling strategy that is able to over detect TB cases, as every found case is a treated one, while taking into account cost and logistic constraints.

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 is a simple strategy composed by both an adaptive component, able to over detect cases 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 design weights that adjust for both the over-selection bias and for the conditional structure induced by the sequential selection. We also proposed an exactly unbiased estimator for its variance in a closed form and ready to implement for actual computation. We showed some preliminary simulation results which highlighted the potentials of PoSA for enhancing cases while considering logistic issues. Our preliminary simulations also highlighted the need for some sort of control over the final sample size. In fact the PoSA design as thought of in this first proposal, allows for a random sample size, being, sometimes extremely low.

Motivated by the achievements obtained with PoSA, we considered again a Poisson-type design but we fixed a minimum sample size, by updating, at every step of the sequential selection, the inclusion probabilities of all the units left to sample by using the procedure discussed in Chapter 2. In fact this flexible updating system allows to modify step by step the inclusion probabilities of units not yet in the sample and to obtain a sequential sampling design with predefined characteristics, such as fixed sample size. We called this method Conditional Poisson Sequential Adaptive design (CPoSA) and an HT-type estimator was developed under this design. The mothod allows for relaxing the assumption of a fixed sample size only if clusters are encountered.

By means of an extensive simulation study, we showed the behaviour of the proposed strategies under different scenarios aiming at representing possible TB prevalence sampling settings. More specifically, following WHO guidelines, we considered the survey area divided into goegraphical areas of population units. We simulated populations with different levels of cases clusterisation by modifying the between areas prevalence variability. The two proposed strategies were then compared with the traditional strategy reccommended in the WHO guidelines with regards to ability to detect cases and cost control. Both PoSA and CPoSA procedures, thanks to their adaptive component, are able to spot more cases than the traditional procedure suggested by WHO, for any level of between areas variability. In particular, as expected by adaptive designs, as the variability between areas prevalence increases, the ability to over detect cases increases. It is remarkable that, although the sample size in the proposed CPoSA design increases as compared to the traditonal design, the added units are always cases, meaning that the additional cost is used only to spot additional cases. In other words the difference in the sample size given by the CPoSA design is due to the addition of positive cases, meaning that the cost to spot one case is lower in the proposed design as compared to traditional designs.

The proposed estimators are very easy to implement as compared to other estimators proposed for adaptive strategies. Moreover when the variation between the areas prevalences is high, the proposed estimators do not loose in efficiency. However when cases are not clustered or only slightly clustered, our proposed estimators loose in efficiency as compared to the traditional HT estimator. The loss in efficiency is due to the fact that, when cases are only sligthly clustered, the sample size yielded by our proposed designs is more variable than when cases are highly clustered. Notice however that the proposed design is intended to be used in situations of high clusterisation. When a control over the final sample size is imposed with the CPoSA design, a minimum sample size is fixed, but as clusterization increases there are many occasions in which cases are over-sampled increasing the sample size. The MC distribution of the CPoSA estimator in fact presents many outliers on the upper tail as compared to traditional designs and to PoSA estimator. This suggests that *(i)* the loss in efficiency corresponds to the returning of a large number of cases and *(ii)* the proposed strategy may be finalised by adding a control over the maximum sample size as well.

5.1 Openings for future research

Interesting perspectives for future research are still opened. Notice that as a first proposal, we considered here a basic sampling design in combination with a simple estimator such as the HT-type estimator. Improvements may thus be made at both design level and estimation level. Moreover, for on field implementation, additional modifications may be needed when actually using the proposed sampling design.

Room for improving the selection design is offered by fully exploiting the potential of a probability updating system in a list-sequential design. In fact, the update of the inclusion probabilities could be finely tuned for example based on how well the adaptive condition is satisfied. For instance the probability of selecting a specific unit may increase more if the threshold is widely exceeded by nearby units, and be only slightly increased if the threshold is slightly exceeded. In other words, the updating probabilities may be proportional to the values of the study variable y observed on the selected sample. Moreover, instead of only refering to subsequent units, a different updating system may allow for the observed values of y to affect units differently according to spatial distance. Distance may not only be geographical, as considered in this first proposal, but may be based on socio-economic features hence be a similarity measure. In the case of TB, for example, a similarity measure may be based on characteristics that classify urban slums, so that if a high percentage of TB cases is found in an urban slam, the probability of including other urban slums is increased. Survey costs may be also integrated in the updating of inclusion probabilities. In fact, instead of setting a maximum/minimum number of units to be included in the sample, a maximum/minimum budget may be used for sequentially updating the inclusion probabilities of units yet to sample. For instance, the inclusion probabilities may be tuned on a combination of budget left and the value of the study variable y in nearby units, so that when the budget left is low, the survey efforts are concentrated mainly in promising areas.

The control over the final sample size needs to be further investigated. In fact, the choice of a lower bound seems reasonable for avoiding small sample sizes and for ensuring that added areas are cases, however the variability in the final sample size, as well as in the proposed estimator may still be large, possibly limiting its use.

In order to reduce the variability in the final sample size, further research may also address the choice of the areas' size. In fact, as discussed in Section 4.4, the variability in the final sample size, as well as in other measures such as detection rate, increases as the between areas variability increases, but can be lowered by choosing smaller areas' sizes.

With regards to the estimation, a more sophisticated estimator may improve the PoSA and CPoSA estimation. The implementation of an Hajék-type estimator, for instance, may be reasonable. Moreover the availability of auxiliary variables may be considered for improving the PoSA and CPoSA estimation via regression.

We found that the population ordering defining the sequential selection affects the probability of a unit to be included in the sample. In fact, units that in the ordered sequence are at the end of a cluster tend to be sampled more often than those that are at the beginning. A way to deal with this effect in our inspirational example, may be that of using information on TB prevalence that may be available beforehand. The population ordering may thus take into account not only costs and logistic constraints, but, based on possibly available information, units may also be ordered so that promising areas have a higher probability of being included in the final sample.

In a perspective of applying the proposed design in TB prevalence surveys, some additional changes in the design may also be discussed. In this thesis we have assumed that the route is chosen after negotiation with local authorities according to the standard practice. In fact, the feasibility of the design should be discussed with local authorities and the design should be tailored to each situation. Moreover, the assumption that added areas are areas with a large number of cases is reasonable and fully supported by simulation results, but in real applications knowledge about neighbourhoods may not be perfect, motivating the use of additional information for ensuring that added areas are mostly cases.

Finally, the choice of the adaptive condition is crucial in all adaptive designs. Possible available information may also be used on making the choice for the threshold defining a good adaptive condition. If the coefficient of between areas variation is expected to be high, because the study variable is highly clustered in the population, as might be the case for a infectious disease such as TB, any threshold may be good for adaptivity, while if the coefficient of between areas variation is low, a more refined threshold shuold be chosen.

Bibliography

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

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

[3] Brown, J. A., Salehi, M. M. (2008). An adaptive two-stage sequential design for sampling rare and clustered populations. *Population Ecology,* 50, 239–245.

[4] Dryver, A. L., Thompson S. K. (2005). Improved Unbiased Estimators in Adaptive Cluster Sampling. *Journal of the Royal Statistical Society. Series B (Statistical Methodology),* 67(1), 157-166.

[5] Floyd, S., Sismanidis, C., Yamada, N., Daniel, R., Lagahid, J., Mecatti, F., Vianzon, R., Bloss, E., Tiemersma, E., Onozaki, I., Glaziou. P., Floyd, K. (2013). Analysis of tuberculosis prevalence surveys: new guidance on best-practice methods. *Emerging Themes in Epidemiology,*10(10)

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

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

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

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

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

[11] Félix-Medina M.H., Thompson, S.K. (2004). Adaptive Cluster Double Sampling. *Biometrika*, 91, 877–891.

[12] Pontius, J.A. (1997). Strip Adaptive Cluster Sampling: Probability Proportional to Size Selection of Primary Units. *Biometrics* , 53, 1092–1096.

[13] Rocco, E. (2016). Spatially-balanced adaptive web sampling. *Environmental and Ecological Statistics*, 23, 219–231

[14] Roesch, F.A. Jr.: Adaptive Cluster Sampling for Forest Inventories. Forest Science **39**: 655-669 (1993)

[15] R Core Team: R (2015). A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

[16] Salehi, M.M., Seber, G. A. F. (1997). Adaptive Cluster Sampling with Networks Selected without Replacement. *Biometrika*, 84, 209–219.

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

[18] Smith, D.R., Conroy, M.J., Brakhage D.H. (1995). Efficiency of Adaptive Cluster Sampling for Estimating Density of Wintering Waterfowl. *Biometrics*, 51, 777–788.

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

[20] Thompson, S. K. (1991). Adaptive Cluster Sampling: Designs with Primary and Secondary Units. *Biometrics*, 47(3), 1103-1115.

[21] Thompson, S. K. (1991). Stratified Adaptive Cluster Sampling. *Biometrika*, 78(2), 389-397.

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

[23] Thompson S. K. (2006). Adaptive web sampling. *Biometrics*, 62, 1224–1234.

[24] Thompson S- K. (2011). Adaptive network and spatial sampling. *Survey Methodology*, 37, 183–196.

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

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

[27] Tourangeau, R., Michael Brick, J., Lohr, S., & Li, J. (2016). Adaptive and responsive survey designs: a review and assessment. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

[28] Turk, P., Borkowski, J.J. (2005). A Review of Adaptive Cluster Sampling: 1990–2003. *Environmental and Ecological Statistics*, 12, 55–94.

[29] Wang, Y. H. (1993). On the number of successes in independent trials. *Statistica Sinica* 3, 29–312.

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

[31] The World Health Organisation (2016). Global tuberculosis report 2015. WHO Press, Geneva.

[32] Yang, H., Magnussen, S., Fehrmann, L., Mundhenk, P., & Kleinn, C. (2016). Two neighborhood-free plot designs for adaptive sampling of forests. *Environmental and Ecological Statistics*, 23(2), 279-299.