Methodological perspectives for surveying rare and clustered population: towards a sequentially adaptive approach

Federico Andreis

Abstract Sampling a rare and clustered trait in a finite population is challenging: traditional sampling designs usually require a large sample size in order to obtain reasonably accurate estimates, resulting in a considerable investment of resources in front of the detection of a small number of cases. A notable example is the case of WHO's tubercoulosis (TB) prevalence surveys, crucial for countries that bear a high TB burden, the prevalence of cases being still less than 1%. In the latest WHO guidelines, spatial patterns are not explicitly accounted for, with the risk of missing a large number of cases; moreover, cost and logistic constraints can pose further problems. After reviewing the methodology in use by WHO, the use of adaptive and sequential approaches is discussed as natural alternatives to overcome the limitations of the current practice. A small simulation study is presented to highlight possible advantages and limitations of these alternatives, and an integrated approach, combining both adaptive and sequential features in a single sampling strategy is discussed as a promising methodological perspective.

Key words: spatial pattern, prevalence surveys, logistic constraints, sampling strategies

Federico Andreis
Dipartimento di Analisi delle Politiche e Management Pubblico
Università Bocconi, Via Roentgen 1, 20136 Milano
e-mail: federico.andreis@unibocconi.it

1 Introduction and motivational example

When knowledge pertaining a trait that is particularly rare in the population is of interest, for example the estimation of its prevalence, the collection of survey data presents various challenging aspects. For instance, in order to obtain a reasonably accurate estimate, 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 tend to perform poorly [11]. The inspirational example for this paper is an epidemiological undertaking of international relevance: the estimation of tuberculosis (TB) prevalence in countries considered to bear a high burden, and where notification data obtained through routine surveillance are incomplete or of unproven accuracy [3]. In these countries, typically developing countries in 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 the World Health Organisation (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 presence of medical doctors on field during these surveys, every TB case that can be found can and will be cured. This paper discusses the most critical aspects of the current practice and possible research lines to overcome its limitations.

2 Sampling strategies

2.1 Current practice in TB prevalence surveys

The sampling strategy currently implemented by WHO is a multistage procedure where at the first stage a probability-proportional-to-size (π-ps) design is implemented. The population is divided into a certain number of areas, defined as geographical regions of as homogenous population size as possible; this working hypothesis allows keeping the final sample size in control thus helping, to some extent, the planning of the survey. All eligible individuals within the sampled areas are invited to show up at a moving lab, where a medical examination takes place, and spotted TB positives can be treated immediately. The number of areas to be sampled is chosen according to the required sample size as function of *(i)* a prior guess of the true prevalence, *(ii)* the desired estimation precision (usually around 25%), and *(iii)* an estimate of the variability existing between the areas' prevalences [12]. A classic Horvitz-Thompson (HT) approach is employed to estimate the prevalence based on sample evidence.

The suggested sampling strategy, Unequal Probability Cluster Sampling (UPCS, from now on), is easy to implement and understand, as expected of an approach suggested in official guidelines, however it has limitations. The rarity of TB positives

and their uneven distribution over the inspected areas, in particular, lead to the need for a very large sample size to obtain an accurate estimate of the true prevalence, and possible information on between areas variability is only accounted for in the sample size determination, i.e., it is in some sense suffered. It is well understood that traditional designs tend to miss cases when they are clustered: it is speculated that information concerning between areas variability should be exploited to inform unit selection itself, in order to be able to concentrate surveying efforts in areas where spotting a case is more likely. WHO's practice for TB prevalence surveys may then draw benefit from a more refined sampling strategy that are able, for instance, to lead to an oversampling of cases, the sample size being equal, and explicitely allow for controlling variable survey costs and possible logistic constraints.

2.2 Enhancement of detection power: adaptive cluster sampling

Adaptive strategies have been developed to deal with populations presenting spatial patterns of the trait of interest, as well as to deal with hidden and hard to sample populations [10]. Different approaches are possible under the general idea of adaptive sampling: among these, we deem most suitable for our epidemiological example the so-called adaptive cluster sampling (ACS, [9]). Under this approach, the country at study is divided into a regular grid of M non-overlapping areas called quadrats; ACS then requires *(i)* a proximity measure between quadrats so that a neighbourhood can be defined for each one, *(ii)* a condition, typically of the form $y_j > c, c \in \mathbb{N}$ where y_j is the number of TB positives in area $j = 1,...,M$, and *(iii)* an initial sampling step with given inclusion probabilities π_j to draw a first sample of areas. The initial sample is drawn using a simple π-ps design and for those areas that satisfy the condition, i.e. for which a certain prescribed number of cases has been found, the neighbourhood is chosen to be included in the final sample as well. The sampling procedure stops when no more neighbouring areas satisfy the condition. This method has proven to be more efficient than traditional non-adaptive sampling strategies when the population is rare and clustered and when the within quadrats variability in terms of prevalence is lower than the between quadrats variability ([10]). As compared to traditional designs, ACS would provide unbiased estimation of the population prevalence while most likely returning a larger amount of cases. However, in its basic form the final sample size is random, thus making it difficult to plan survey costs. Moreover, the route that the moving lab will follow is mainly determined by the sample that has been drawn at the first stage, thus preventing researchers from effectively dealing with possible logistic constraints at the design stage of the survey.

2.3 Dealing with logistic constraints: list sequential sampling

Renewed interest has arisen recently on sequential methods (a notable example can be found in [2]), that apply when units can be ordered in some way, leading to interesting developments in the field of spatial sampling. Such methods allow the user to specify a flexible weighting system used to sequentially adjust inclusion probabilities on the basis of auxiliary information (not necessarily available beforehand). As opposed to adaptive designs, these methods offer some way to control the final sample size but do not possess the specific feature of being able to oversample study cases. Bondesson and Thorburn ([2]) developed a general sequential method for obtaining a π-ps sample that well suits the example at hand: this very broad approach considers populations whose units can be ordered somehow and visited sequentially. Moving from some preliminary choice, the inclusion probabilities are revised after each unit has been visited by means of an updating procedure that allows correlation to be purposively introduced between successive sample membership indicators (we refer the reader to the seminal paper [2] and to [4], [5] for details on the updating algorithm); we focus in particular on Spatially Correlated Poisson Sampling (SCPS, [6]), that provides a spatial extension to the method.
The sequential approach can naturally accomodate for a predefined route: this might be particularly relevant when planning TB prevalence surveys, in that specific knowledge might lead to individuate a path along which transportation costs are minimized and logistic constraints can be taken into account beforehand. An HT approach to unbiased estimation can be undertaken with this sequential strategy as well, where sample cases are weighted according to the conditioning mechanism induced by the updating rule of the inclusion probabilities (cfr. [2]). Differently from ACS, however, the list-sequential setting does not allow, in its current formulation, for over-detection of cases nor to adaptively incorporate sample evidence on the response on the run.

3 Some empirical evidence

A small simulation study inspired by the TB prevalence surveys example is presented. The main aim is to compare UPCS, ACS and SCPS with respect to the following two key aspects:

1. *cases detection rate*: a crucial challenge in TB prevalence surveys (as well as in surveying any rare and clustered trait) is the enhancement of the detection power of the sampling strategy, since every found case, is a case that can potentially be treated
2. *costs*: from an operational point of view, reducing costs could lead to the opportunity of furthering the survey and thus managing to spot (and treat) more cases, as well as to improve accuracy in estimating the true prevalence.

The simulation has been carried out completely in the `R` environment ([8]), and the packages `spatstat` ([1]) and `BalancedSampling` ([7]) have been used to implement spatial patterns generation and SPCS, respectively.
For the sake of illustration and to highlight limitations and advantages of each of the considered strategies, an artificial population assumed as a possible TB prevalence survey scenario has been simulated. The population is composed by $N = 100000$ units evenly spread over a two-dimensional space. The study variable has value 1 for population units that are TB cases and value 0 otherwise, with a popuplation prevalence $p \approx 0.004$. In the simulated scenario, the cases (423) are mostly clustered in 3 groups (two of which slightly overlapping), homogeneous in terms of prevalence, the coefficient of variation of the cluster-specific prevalences being equal to 0.2. Figure 1 depicts this simple scenario.

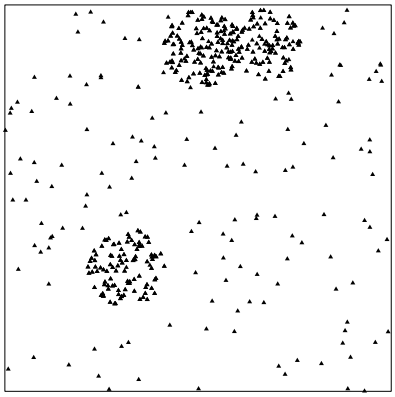

Fig. 1 Simulated population. $N = 100000$, $p \approx 0.004$, 3 clusters. Black triangles indicate cases

The sample size has been computed, as suggested in the latest WHO guideline ([12], chapter 5) for a desired precision level in the estimation of the prevalence, to be $n \approx 27983$, that in turn leads to approximatley 34 areas to be sampled to reach the desired number of selected units. Note that this translates into a fairly large sample fraction of 28%. In order to compare sampling strategies with respect to costs, we consider the popular linear cost function. The actual total cost C for the survey is computed as a simple linear function of the total number of selected population units n, the total number of sampled areas n_a (implying that the cost for moving the lab from one area to another is a significant component of the total survey cost) and of some fixed costs c_0 as follows:

$$C = c_0 + c_1 n + c_2 n_a. \tag{1}$$

We set $c_0 = 100000$ (essential staff, equipment, advertising, ...), $c_1 = 10$ (cost per unit), and $c_2 = 1000$ (cost for transportation and installation of the moving lab in the new location). We ignore, for simplicity, other sources of cost such as unexpected events or area-specific issues that can be expected in an actual implementation. Moreover, we encode the advantage of a careful route planning, easily accomodated by the sequential approach, by reducing by 50% the area sampling cost, i.e. $c_2 = 0.5 \cdot 1000 = 500$, for SCPS.
The area sampling frame of the country is divided into $M = 100$ quadrats by means of a regular grid; each of these areas is a primary sampling unit $j = 1, ..., M$ and all the individuals living in a selected area will be invited to participate into the final sample. The ACS adaptive condition is set so to include nearby areas in the sample if the number of cases y_j in a selected area j exceeds the arbitrarily chosen threshold $\lfloor p_g N/M \rfloor = 11$, for a maximum of two subsequent steps (i.e., no more than two additional neighbourhoods can be included). We assume for simplicity a second-stage units 100% participation rate to the survey (WHO guidelines report an actual participation rate usually in the range $85\% - 90\%$ [12]).

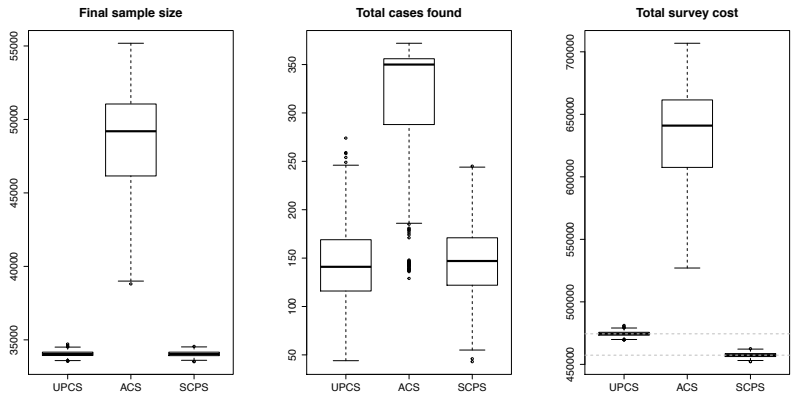

Fig. 2 Final sample size, detected cases and total costs distributions over 1000 runs. The dashed lines indicate median costs for UPCS and SCPS for ease of comparison.

Figure 2 summarizes the results of 1000 Monte Carlo runs on the proposed scenario. The characteristic feature of adaptive sampling (ACS), i.e. the ability to oversample cases, is immediately evident (center panel in Figure 2), as it is the out of control impact this may have both on the final sample size (left panel in Figure 2) and on the total costs (both in terms of location and variability, see right panel of Figure 2). On the other hand, the sequential approach (SCPS) presents a very stable behaviour in terms of detection power and costs, also highlighting the gain in planning a cost-minimizing route beforehand, but is unable, as expected, to spot more cases than the current practice (UPCS).

ACS was simulated by initially sampling the prescribed 34 areas, as in UPCS, which explains the out of control final sample size; we note that choosing a smaller amount of primary sampling units may help reducing sample size variability. However, a smaller initial sample size that would guarantee both more control in sample size variability and oversampling of cases would depend on the spatial structure of the population: without a priori specific knowledge on such aspect, a univoque determination of how many areas to select in the first stage would not be obtainable.

4 Final remarks: a first step towards an integrated strategy

Sampling a rare and clustered trait is a challenging task that can be tackled in many different ways. A field in which this is particularly relevant is that of WHO's TB prevalence surveys, that could greatly benefit from advances in the sampling strategy currently suggested in the latest guidelines. In this paper we compare the application to the TB inspirational example of two existing approaches, namely Adaptive Cluster Sampling and Spatially Correlated Cluster Sampling, to the current practice of Unequal Probability Cluster Sampling. ACS and SPCS are viable alternatives and can address specific limitations that are known to affect UPCS: for instance, the adaptive setting can improve the number of cases detected, whereas the sequential approach allows to account for logistic constraints and might help to plan the survey, while not neglecting the spatial pattern which usually characterizes national TB prevalence surveys. The current practice, as suggested by latest WHO guidelines, relies on some working assumptions such as (i) choosing areas of approximately equal size, and (ii) computing the required sample size without taking full advantage of spatial information; both assumptions aim at simplifying the survey setup and subsequent estimation of the true prevalence, but may lead to a large dissipation of resources (both in monetary and cases-missing terms). By means of a simulation study we stress these specific features with respect to final sample size, detection power and survey costs in a an artificial simplified scenario where a rare and clustered trait in a finite population is to be investigated. The results highlight the contribution of ACS to increasing the number of cases detected and of SCPS to controlling survey costs by accounting for logistic constraints and exploiting a carefully chosen survey route, while also pointing out their limitations. The choice of a simple linear cost function as a mean of comparison of survey costs under the different strategies is in line with the current practice for large scale surveys; however, more complex and ad-hoc solutions could be devised with reference to specific applications to better account for various sources of expenditure. Drawing on theoretical considerations and empirical evidence, we outline some research perspectives towards an integration of the adaptive and the sequential approach: we claim that an integrated approach could pursue both the desirable features of oversampling cases and controlling costs simultaneously, which is not possible using either of UPCS, ACS or SPCS alone. It also is stressed that the issue of estimation should receive particular attention as well, given the importance of obtaining accurate estimates of

the true TB prevalence needed, for example, to inform public health policies and plan interventions. Finally, it is worth noting that the development of a sampling strategy that allows to enhance cases-detection while controlling survey costs as well as accounting for logistic constraints, has the potential to effectively apply in a wider range of practical fields besides the epidemiological example that motivated this paper.

References

1. Baddeley, A. and Turner, R.: spatstat: An R Package for Analyzing Spatial Point Patterns. Journal of Statistical Software **12(6)**, 1-42. http://www.jstatsoft.org/v12/i06/ (2005)
2. Bondesson, L., Thorburn, D.: A list sequential sampling method suitable for real-time sampling. Scandinavian Journal of Statistics. **35**, 466–483 (2008)
3. Glaziou, P., van der Werf, M.J., Onozaki, I. Dye, C.: Tubercolosis prevalence surveys: rationale and cost. International Tubercolosis Lung Disease. **12**(**9**), 1003–1008 (2008)
4. Grafström, A.: On a generalization of Poisson sampling. Journal of Statistical Planning and Inference. **140**, 4, 982–991 (2010)
5. Grafström, A., Lundström, N.L.P. and Schellin, L.: Spatially Balanced Sampling through the Pivotal Method. Biometrics. **68**, 2, 514–520 (2011)
6. Grafström, A.: Spatially Correlated Poisson sampling. Journal of Statistical Planning and Inference. **142**, 1, 139–147 (2012)
7. Grafström, A. and Lisic, J.: BalancedSampling: Balanced and Spatially Balanced Sampling. R package version 1.5.1. http://CRAN.R-project.org/package=BalancedSampling (2016)
8. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/ (2015)
9. Thompson, S.K.: Adaptive cluster sampling. Journal of the American Statistical Association. **85**, 1050–1059 (1990)
10. Thompson, S.K., Seber, G.A.F.: Adaptive Sampling. John Wiley & Sons, Inc. (1996)
11. Thompson, W.L.: Sampling rare or elusive species. Island Press, New York (2004)
12. The World Health Organisation: Tubercoulosis PREVALENCE SURVEYS: a handbook. WHO Press, Geneva (2011)