Balanced adaptive web sampling

E. Rocco

DISIA Dipartimento di Statistica, Informatica, Applicazioni
"Giuseppe Parenti" - University of Florence

Kick-Off Meeting, PRIN-2014, 2012F42NS8

Roma, 12-13 Febbraio 2014

Outline

Adaptive sampling designs

Adaptive sampling designs are those in which the procedure for selecting units may depend on values of the variable of interest observed during the survey.

They are generally suggested for sampling hidden to reach populations that have a spatial clustered or a network structure.

Adaptive sampling in a spatial setting

In a spatial setting, the population units are usually plots or cells of a grid overlapping an area of interest.

The use of the adaptive design is motivated when the spatial distribution of the study is highly uneven and above all when few units of the population exhibit relevant values of the study variable and these units present a clustered pattern.

Adaptive sampling is exemplified by a survey in which whenever a unit in the sample is observed to have a value of the study variable that satisfies a certain condition, specified in advance, nearby units have high probability to be added to the sample.

Adaptive sampling in a network setting

Usually population with network structure are conceptualized as graphs with nodes of the graph representing the units of the population and the edge or arcs of the graph representing the relationships or links between the units of the population.

In a network setting, adaptive (or link-tracing) sampling is exemplified by a survey in which whenever a unit in the sample satisfies a specified condition, or belongs to a particular subpopulation, social links from it are followed to locate and add additional members of the subpopulation to the sample.

In network setting s the developments of adaptive designs has been motivated by problems in sampling people with rare diseases and in sampling hidden populations such as those at high risk for HIV/AIDS or other epidemics.

Adaptive web sampling

The study of populations with the characteristics above mentioned has motivated the development over the time of several sampling designs.

Non-probabilistic designs, among which the snowball sampling for example.

Probabilistic designs that allows randomization-based finite population estimation and inference; among them network sampling and several variants of adaptive cluster sampling.

One of the more recently suggested adaptive Sampling designs is the **adaptive web sampling** (Thompson, 2006). It has an important advantage, compared with previously existing adaptive and link tracing designs: the control over the sample size.

Adaptive web sampling

Actually the expression "adaptive web sampling" is used to identify a class of adaptive sampling designs in which several components may be varied.

Only a particular, and perhaps the simpler, version of it has been effectively investigated by Thompson (2006).
Hereafter we call this variant of the adaptive web sampling as simple adaptive web sampling (SAWS)

Adaptive web sampling - notation

- The area of study is divided into N units or plots labelled $1, 2, ..., N$.
- With each unit i is associated a variable of interest, denoted as y_i.
- In addition for each plot a set of neighboring plots is defined, this may be exemplified defining for any pair of plots an indicator variable w_{ij} with $w_{ij} = 1$ if i and j are neighbors and zero otherwise.

Adaptive web sampling - selection scheme (1)

An adaptive web sampling is selected in steps, the selection scheme of its simplest version (SAWS) may be exemplified as follows:

- An initial sample s_0 of n_0 units is selected by simple random sampling.
- For each of these units the variable of interest is observed and if for a unit i the observed value y_i, $(i \in s_0)$, satisfies a condition of interest specified a priori, the unit and its associated variables, that is y_i and the values of all the links out from it, are enclosed in a set denominated *active set*.
- Then the remaining $n - n_0$ units (n is the final sample size fixed in advance), are selected one at a time in successive steps.

Adaptive web sampling - selection scheme (2)

- At each step k with $k = 1, ..., n - n_0$ the next unit is selected from a *mixture distribution*, so that:
  - with probability d one of the links in the current active set that do not go to other units already in the sample is selected at random and followed to bring a new unit in the sample
  - with probability $1 - d$ (usually small) a new unit is selected at random from the total set of units not yet in the sample.
- Moreover, at each step if there are no links out from the current active set to any unsampled unit the next unit is select at random from the collection of unsampled units.

Adaptive web sampling - selection scheme (3)

Formally the probability that unit i is selected in the k_{th} step is:

$$q_{ki} = p\frac{w_{\alpha_k i}}{w_{\alpha_{k+}}} + (1-p)\frac{1}{N - n_{s_{ck}}} \tag{1}$$

where:

- $s_{ck} = \bigcup_{i=0}^{k-1} s_i$ denotes the current sample at step k,
- α_k ($\alpha_k \subset s_{ck}$) denotes the current active set at step k,
- $n_{s_{ck}}$ is the number of units in the current sample,
- $w_{\alpha_{k+}} = \sum_{i \in \alpha_k, j \in \bar{s}_{ck}} w_{ij}$ is the total number of links out, from the active set to units not in the current sample,
- $w_{\alpha_k i} = \sum_{j \in \alpha_k} w_{ij}$ is the number of the links out, from the active set to unit i.

If there are no links at all out from the current active set, then:

$$q_{ki} = \frac{1}{N - n_{s_{ck}}} \tag{2}$$

The overall sample selection probability for the ordered sample $\mathbf{s} = \{s_0, s_1, ..., s_{n-n_0}\}$ is:

$$p(\mathbf{s}) = \left(\begin{array}{c} N \\ n_0 \end{array} \right)^{-1} \prod_{k=1}^{n-n_0} q_{ki}$$

Rao-Blackwellized estimator

The simplest and more efficacy unbiased mean estimator suggested by Thompson (2006), for a sample selected by the scheme described above, is obtained, finding, via the Rao-Blackwell approach, the conditional expectation of sample mean of the initial sample, $\bar{y}(s_0)$ given the reduced set of data.

Reduced set of data:

$$d_r = \{(i, y_i, w_{i+}, w_{ij}), i \in \mathbf{s}, j \in \mathbf{s}\}$$

Estimator:

$$\hat{\mu} = \sum_{\mathbf{s}:r(\mathbf{s})=s} \bar{y}(\mathbf{s}) p(\mathbf{s}|d_r)$$

Markov Chain reesampling approach

Computation of the estimator $\hat{\mu}$ and of its variance estimator requires enumerating all the reorderings of the sample units. For each reordering the probability of that reordering need to be computed along with the value of the estimator and of its variance estimator.

Enumerative calculus is prohibitive even for relatively small values of n.

A Markov chain resampling procedure (corresponding to the Markov chain accept/reject procedure of Hastings (1970)) is suggested in Thompson (2006) for making inference computationally feasible.

Variations on the simple adaptive web sample

- After the selection of the initial sample, in successive steps the adaptive web selection allows to concentrate the survey effort in the subareas considered of great interest, on the base of the survey values previously observed;
- therefore,it is important that the units selected in the first steps and in particular the initial sample provides as much as possible information on which are the possible subareas of interest.
- For achieving this goal it is important the proportion of total units allocated to the initial sample;
- but above all, it is important the sampling design used for the initial sample.
- These considerations are especially true in situations where the population size is large and the sampling fraction is small.

We suggest to select the initial sample by using a spatial balanced sampling design.

A possible balanced sampling design for the initial sample

- Not all the spatially balanced designs present in literature can be used;
- the choice should be made taking into account that the sample selected with this design is only the first step of the more complex web adaptive selection procedure.
- For example we cannot use a systematic design that selects n_0 units equally spaced in the study area because it is not compatible with the estimation procedure.

Here we suggest to divide the population into strata as small as possible (consistent with the size of the initial sample) and then select an initial stratified sample.

Attention: The strata are arbitrarily built only in order to divide the area of study into several parts

Spatial Balanced adaptive web sampling

- First an initial spatial stratified sample is selected.
- Then the remaining units are selected one at time in steps following the identical procedure described for SAWS.
- Beyond the selection of the initial sample the stratification of the population is not taken anymore into account.
- When a sample is selected with this scheme, the expressions for the probability that unit i is selected in step k ($k = 1, ..., n - n_0$), are the same as those defined for SAWS.
- The overall sample selection probability for the ordered sample become:

$$p(\mathbf{s}) = \prod_{k=1}^{n_0} \left(\begin{array}{c} N_h \\ n_{h0} \end{array} \right)^{-1} \prod_{k=1}^{n-n_0} q_{ki}$$

- Regarding the inference both the Rao-Blakwell estimation procedure and the Markov chain resampling approach for making its computation feasible are readily applicable.

Simulations

The properties of spatial balanced adaptive web sampling (SBAWS) have been evaluated via several simulation studies using three empirical population with different spatial distribution.

For each population SBAWS is compared with simple random sampling (SRS), with SAWS and with the "spatially balanced sampling through the pivotal method" (SBSTPM) suggested by Grafström et al. (2012).

For every pair of sampling designs the comparison is made by means of the ratio between the mean square errors of the two corresponding estimators. populations.

Results of simulations, some remarks and open questions (1)

The simulation results are clearly favorable to SBAWS.

- When the sample fraction is high SBAWS is equivalent to the SAWS.
- When the size of the population is high and the sample fraction is small the gain in efficiency of the SBAWS compared to the SAWS, may be significant.
- Moreover the results obtained for the two sample parameters settings in the case of population B seem to suggest that more the initial sample is spread, greater is the efficiency gain.

Results of simulations, some remarks and open questions (2)

What is better? A sample as much as possible well-spread in the study area or a sample that concentrate the survey effort in some subareas?

- SBAWS may be seen as a first compromise between the two approaches.
- However much work remains to be done:
  - From the results it is clear that for sampling an uneven spatial population, it is advantageous first of all to collect information from units well spread over the population
  - however if it is appropriate (and in which proportion) to set aside some survey resources in order to subsequently allocate them to subareas of greatest interest depends on the unknown distribution of the population.
  - ....The next goal of our research is to find a sampling design in which the initial sample is used even in order to indicate if the remaining survey resources should be allocated in some subareas or well-distributed over all the population.

Balanced adaptive web sampling vs PRIN-2014 (2)

Population A

0	0	3	5	0	0	0	0	0	0
0	0	0	24	14	0	0	10	103	0
0	0	0	0	2	3	2	0	13639	1
0	0	0	0	0	0	0	0	14	122
0	0	0	0	0	0	2	0	0	177

Figure: Blue-Winged teal population. Study area divided in 50 plots. (Smith et al. 1995)

Population B

0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	20	4	2	12	0	0	0	0	0	10	103	0	0	0
0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	150	7144	0	0
0	0	0	0	0	0	0	0	2	0	0	0	0	2	0	0	6	6339	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	122	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	114	60
0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	3	0

Figure: Blue-Winged teal population. Study area divided in 200 plots. (Smith et al. 1995)

Population C

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	2	0	0	0	3	0	0
0	0	0	0	0	0	5	2	2	2	1	0	0	0	0
0	0	0	0	0	0	0	1	0	1	1	0	0	0	0
0	0	0	0	0	0	0	3	2	3	1	0	1	1	0
0	0	0	0	0	0	6	7	1	0	0	0	0	1	0
0	0	0	0	0	0	1	3	0	0	0	2	0	0	0
0	0	0	0	0	0	0	2	8	0	0	2	0	0	0
0	0	0	0	0	0	3	5	20	2	2	1	0	0	0
0	0	0	0	0	0	3	5	6	0	0	2	0	0	0
0	0	0	0	0	0	5	13	11	0	0	0	0	0	0
0	0	0	0	0	0	0	16	2	0	1	0	0	0	0
0	0	0	0	0	0	0	4	1	0	0	0	0	0	0
0	0	0	0	0	0	1	2	2	2	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure: Caste Hill buttercups population. Study area divided in 300 plots.