Published in Proceedings of the 47th Scientific Meeting of the Italian Statistical Society (Cagliari - June 11/13, 2014)

Spatially balanced adaptive web sampling

Il campionamento spaziale bilanciato e adattivo

Emilia Rocco

Abstract The paper deals with sampling from a finite population that is distributed over space and has an highly uneven spatial distribution. It suggests a sampling design that allocates a portion of the sample units well-spread over the population and selects sequentially the remaining units in sub-areas that appear of more interest according to the study variable values observed during the survey. In order to estimate the population mean while using this sampling design, a computational intense estimator, obtained via the Rao-Blackwell approach, is proposed and a resampling method that makes the inference computationally feasible is used. The whole sampling strategy is evaluated through several Monte Carlo experiments.

Abstract *Nell'articolo si affronta il problema di campionare una popolazione che ha una distribuzione spaziale molto irregolare. Si suggerisce un disegno di campionamento con il quale una porzione delle unità campionarie sono equidistribuite nell'intera area di studio e le rimanenti unità sono selezionate in sottoaree considerate di maggiore interesse in base ai valori della variabile di studio osservati durante la stessa indagine. Per stimare la media della popolazione viene proposto uno stimatore basato sul criterio di Rao-Blackwell per l'inferenza da popopolazioni finite. Poichè tale stimatore è computazionalmente intenso, si suggerisce di appross-simarlo utilizzando un metodo di ricampionamento. La metodologia proposta viene valutata mediante diversi esperimenti Monte Carlo.*

Key words: Adaptive sampling, Rao-Blackwell, Uneven distribution

1 Introduction

Usually the goal of sampling in a spatial setting is the efficient estimation of a parameter of a population, such as density or abundance, in a certain region of study.

Emilia Rocco
Dept of Statistics, Informatics, Application ”G. Parenti” - Uviversity of Firenze
e-mail: rocco@disia.unifi.it

Difficulty in achieving this goal may depend on the fact that, often the spatial distribution of the study variable is highly uneven and sometimes only few units of the population exhibit relevant values of the study variable and these units also present an extreme clustered pattern. Spatial rare and clustered populations are common in many fields of application and in particular in many environmental and natural resources studies. For example, many populations of plants and animals are made up of a few specimens and have aggregative tendencies, as well as different kinds of pollution have a negligible concentration in most sites and a very high concentration in a few scattered subareas. Likewise, in a different contest, is quite common the case of human populations and populations consisting of socio-economic units that have a network structure and also an inherent geographical or spatial structure. The study of this kind of populations has motivated the development over the time of several adaptive sampling designs in which the procedure for selecting units may depend on values of the variable of interest observed during the survey. Among these designs, one of the more recently suggested is the adaptive web sampling (Thompson, 2006) which has the advantage, compared with most of the previously existing adaptive and link tracing designs, to control the sample size and the proportion of effort allocated to adaptive selection. An adaptive web sample is selected in steps. First an initial sample is selected by some design. Then, in the simplest case, in each of the subsequent steps, a link out from units selected in the previous steps is selected at random and the unit to which it connects is added to the sample. More generally a set of links can be selected at each step. The links in a spatial setting are commonly defined on the base of some relationship of geographical proximity. Therefore, in steps subsequent to the first, the web adaptive selection allows to concentrate the survey effort in the subareas considered of great interest, on the base of the survey values previously observed. But in order for this to happen, it is necessary that the units selected in the first steps, and in particular the initial sample, provide as much information as possible on which are the possible subareas of interest. In this paper, in order to reach this objective the selection of an initial sample well-spread over the population is suggested. A well-spread sample is usually said to be spatially balanced. Different types of spatially balanced sampling designs exist in literature and are commonly used, for example different types of systematic designs. However 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 and therefore only a spatially balanced sampling design that is compatible with the estimation procedure associated to the adaptive web selection can be used. The rest of the paper is structured as follows. Section 2 sets out the notation and describes the adaptive web sampling selection scheme and the associated inferential procedure. Then, in Section 3, our variations to the original adaptive web design are pointed out. Finally Section 4 describes some Monte Carlo experiments performed in order to give a first evaluation of the suggested method and concludes with some final remarks and ongoing questions.

2 Adaptive web sampling: basic setup, design and estimation

Usually, in a spatial setting, the population units are plots or cells of a grid overlapping an area of interest. With each unit i ($i = 1,..,N$) of the population is associated a value, y_i, of a variable of interest. Moreover, using a neighborhood relationship based on geographic proximity, to each pair of units, (i, j), is associated a variable, w_{ij}, usually called link variable, that describe the neighborhood relationship between them. In many cases w_{ij} is an indicator variable with $w_{ij} = 1$ if the units i and j are neighbors and $w_{ij} = 0$ otherwise. More generally w_{ij} is a weight of the relationship between the two units and can be function of their distance.
Actually, in literature, the expression ”adaptive web sampling” is used to identify a class of adaptive sampling designs in which several components may be varied. However only a sampling design of this class has been effectively investigated (Thompson, 2006). This design, hereafter denoted simple adaptive web sampling (SAWS), selects a sample of size n through the following steps. First an initial sample s_0 of n_0 units is selected by simple random sampling. The study variable values for these units are observed. If the value y_i associated to unit i satisfies a condition of interest specified a priori, then this unit, together with the associated y value and the values of all the links out from it, are enclosed in a set denominated *active set*. The remaining $n - n_0$ units are selected one at a time in successive steps and after each selection the active set is updated since the added unit in turn may satisfy the condition of interest. Moreover 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 connect 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.
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:

$$p(\mathbf{s}) = \binom{N}{n_0}^{-1} \prod_{k=1}^{n-n_0} q_{ki}$$

Among the mean estimators suggested by Thompson (2006) while using SAWS, the simplest and more accurate is that obtained finding, via the Rao-Blackwell approach, the conditional expectation of sample mean of the initial sample, $\bar{y}_0(\mathbf{s})$, given the reduced set of data $d_r = \{(i, y_i, w_{i+}, w_{ij}), i \in \mathbf{s}, j \in \mathbf{s}\}$. This estimator is:

$$\hat{\mu} = \sum_{\mathbf{S}:r(\mathbf{S})=s} \bar{y}_0(\mathbf{s}) p(\mathbf{s}|d_r) \tag{3}$$

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 needs 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 based on the Markov chain accept/reject procedure of Hastings (1970) is suggested in Thompson (2006) for making inference computationally feasible. Denoting with n_r the number of permutations resampled with this procedure and with $\overline{y}_j$ the mean of the first n_0 units for the permutation j, the resampling estimator used to replace $\hat{\mu}$ is:

$$\tilde{\mu} = \frac{1}{n_r} \sum_{j=1}^{n_r-1} \overline{y}_j$$

3 Spatially balanced adaptive web sampling

The SAWS, as well as many other adaptive or link tracing designs, has been suggested for sampling spatial populations in which the relevant values of the study variable are concentrated in a few small sub-areas and no prior information on which are these sub-areas are available. Therefore the idea underlying all these designs is to focus the survey effort in the sub-areas considered of great interest on the base of the relevant values observed during the same survey. To this end, in SAWS first a portion of sampling units is selected through the simple random sampling and then the next selections are made in steps and depend on the information collected in the initial sample and in the other previous selection steps. But, if nearby units are more similar than units further apart, then in order to identify, with higher probability, the more relevant sub-areas, it is advantageous to make sure that the initial sample is well spread over the population. For this reason here we suggest the *spatial balanced adaptive web sampling* (SBAWS) in which first an initial spatial balanced sampling of n_0 units is selected, then the remaining $n-n_0$ units are adaptively selected in steps following the same scheme described in Section 2 for the SAWS. Different types of spatially balanced sampling designs exist in literature and are commonly used. Among them, the one that we suggest here is a sort of ”spatial stratified sampling”. It has been chosen for its simplicity and for the fact that it allows the straight use of the inferential approach suggested by Thompson for the SAWS. Moreover it is an unusual stratified design because no stratification variables are available and the strata are built arbitrarily only in order to divide the area of study into several parts. They have all the same size, N/H (where H denotes the number of strata), and are as small as possible consistently with the initial sample size. Beyond the selection of the initial sample the stratification of the population is not taken into account anymore, the remaining units are selected following the same scheme described in Section 2. Therefore the expression of the probability that the unit i is selected in step $k(k=1,...,n-n_0)$, are the same as those defined for SAWS. On the contrary the overall sample selection probability for the ordered sample change and become:

$$p(\mathbf{s}) = \binom{N/H}{n_0/H}^{-H} \prod_{k=1}^{n-n_0} q_{ki}$$

Regarding the estimation process both the the Rao-Blakwell estimator defined in (3) and the Markov chain resampling approach presented in Thompson (2006) for making its computation feasible are readily applicable.

4 Simulation results and concluding remarks

In order to evaluate the SBAWS, simulations of sampling were carried out from three empirical populations: (A) the Blue-Winged teal population considered in Thompson (2006) in which the area of study is divided in 50 plots; (B) the same Blue-Winged teal population but with the area of study divided in 200 plots (Smith et al. 1995); (C) the Caste Hill buttercups population described in Brown et al. (2012). Due to space limitations, for the description of these three populations please refer to the cited works. 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 Grafstr*ö*m et al. (2012). To this end, for each design the mean square error (MSE) of the associated mean estimator is evaluated and, only for a few of the performed experiments, is given in Table 1 (MSEs in the table are standardized by dividing by that of a simple random sample of equal size). The number of simulation runs for each experiment is fixed at 1,000 and for both SAWS and SBAWS the number of Markov chain resampling used in the estimation procedure is set to 10,000. The results clearly show that when the sample fraction is high SBAWS is equivalent to the SAWS. On the other hand, 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. Regarding the comparison between the SBAWS and the SBSTPM, the greater efficiency of one or the other depends on the spatial structure of the population.

Table 1 Mean square errors (standardized by dividing by that for a random sample of equal size) of mean estimators for different populations and different sampling designs

			SAWS			*SBAWS*				*SBSTPM*
Population	*N*	*n*	n_0	*d*	*MSE*	n_0	*d*	*H*	*MSE*	*MSE*
A	50	20	15	0.9	**0.62**	10	0.9	10	**0.62**	**1.00**
B	200	40	30	0.9	**1.31**	25	0.7	25	**0.79**	**0.83**
C	300	40	30	0.9	**1.31**	30	0.9	30	**0.71**	**0.66**

The SBAWS represents a first compromise between two opposite approaches suggested in literature for sampling uneven spatial populations, the one that recommends to spread the survey effort over the whole population and the one that suggests to concentrate the survey effort in sub-areas of great interest. Further investigation is still necessary, because, on one hand, we find that for sampling an uneven spatial population (in absence of a priori information) it is advantageous to allocate at least a part of the sample units well spread over the population. On the other hand, our results clearly show that the choice whether allocate a portion of sample units adaptively in sub-areas of great interest may depend on the unknown distribution of the population. The investigation of the possibility to define a sampling scheme and a corresponding inference procedure for whom the information gathered with the initial sample can also be used to decide whether to allocate the remaining resources adaptively in some sub-areas or well spread over all the population, would certainly be our next step. Moreover our research is already devoted to investigate the possibility of selecting the initial sample by using different spatially balanced designs.

Acknowledgements This work has been developed under the support of the project PRIN-2012F42NS8 ”Household wealth and youth unemployment: new survey methods to meet current challenges”.

References

1. Brown, J.A., Salehi, M.M,, Moradi, M., Panahbehagh, B., Smith, D.R.: Adaptive survey designs for sampling rare and clustered populations. Mathematics and Computers in Simulation (2012) doi:10.1016/j.matcom.2012.09.008 (2012)
2. Grafström A., Lundström, N.L.P., Schelin, L.: Spatially balanced sampling through the pivotal method. Biometrics **68**, 514–520 (2012)
3. Hastings, W.K.: MOnte-Carlo sampling methods using MarKov chainsand their applications. Biometrika **57** 97–109 (1970)
4. Thompson, S.K.: (2006) Adaptive web sampling. Biometrics **628**, 1224–1234 (2006)
5. Smith, D.R., Conroy, M.J., Brakhage, D.H.: Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl. Biometrics **51**, 777–788 (1995)