

A Critical Look at Adaptive Sampling

Uno Sguardo Critico al Campionamento Adattivo

Stefania Naddeo

Università di Siena, Dipartimento di Metodi Quantitativi, naddeo@unisi.it

Caterina Pisani

Università di Siena, Dipartimento di Metodi Quantitativi, pisani4@unisi.it

Riassunto: Il campionamento adattivo è stato proposto per la stima del totale o della densità di popolazioni rare e con comportamento spaziale di tipo contagioso. Finora le sue caratteristiche sono state analizzate essenzialmente con riferimento alla sua efficienza rispetto al campionamento casuale semplice. In questo lavoro il confronto, esteso ad altri disegni campionari comunemente utilizzati, viene effettuato non solo in termini di efficienza, ma anche per quanto riguarda l'accuratezza degli stimatori.

Keywords: adaptive sampling, spatial autocorrelation, accuracy, simulation study.

1. Introduction

The main purpose of adaptive sampling methods is to take advantage of population characteristics to increase the effectiveness of sampling effort and to obtain more precise estimates of population parameters. Thompson (1992) suggests the use of adaptive cluster sampling as a suitable strategy for both obtaining estimates of the abundance or density of rare, clustered populations and increasing the number of interesting observations. This aggregation tendency may be peculiar of animal or plant populations as well as minerals, fossils fuels and some human populations.

It is worth noting that the relative advantage of the adaptive cluster design is usually measured by means of comparative efficiency with respect to simple random sampling. The aim of this paper is to investigate the performance of adaptive cluster sampling with respect to other spatial conventional or more innovative designs not only by means of efficiency, but also computing the accuracy of the estimates. Indeed, it must be pointed out that the performance of an estimator should be evaluated not only in terms of unbiasedness and efficiency but also in terms of accuracy, that is the probability of obtaining estimates close to the real unknown parameter.

2. Some Outlines of Adaptive Cluster Sampling

“Adaptive cluster sampling was motivated by the problem of sampling rare, clustered populations” (Thompson and Seber, 1996). The basic idea of this design is selecting an initial sample of units and subsequently adding neighbouring units whenever the variable of interest satisfies a given criterion. For example, in epidemiology studies this criterion might be the presence of individuals infected by a rare disease in such a way that the neighbourhood of a unit may be defined in terms of social or genetic links. On

the other hand, in environmental surveys the condition for extra sampling may be the presence of rare biological species or high concentration of fossil fuels, so that the neighbourhood is usually defined by spatial proximity.

Consider a study region partitioned in N spatial units, denote by $y_1, y_2, \dots, y_N$ the values of the variable of interest and by $T = \sum_{i=1}^N y_i$ the parameter to be estimated. An initial sample of m units is selected using a conventional sampling design. If the y -value of a sampled unit i satisfies a given condition C (e.g. $y_i > k$), additional units in the neighbourhood of the i -th unit are added to the sample. If any other unit in that neighbourhood satisfies C , then its neighbouring units are also surveyed, and so on until a final sample S is obtained. It is worth noting that adaptive sampling procedures give rise to selection or inclusion probabilities which may not be quantified for each unit in S in such a way that conventional estimators can not be adopted. Therefore, Thompson (1992) suggests the use of estimators which retain unbiasedness and computes their efficient with respect to simple random sampling by means of simulation studies performed on artificial populations.

3. Short Outlines of Alternative Sampling Designs

It is worth noting that simple random sampling may not be a suitable design for estimating the abundance or density of spatially clustered population. Indeed, neighbouring units tend to present similar values of the variable of interest and this phenomenon, usually referred to as *positive spatial autocorrelation*, seriously deteriorates the performance of simple random sampling, since this design does not avoid the selection of adjacent units. To this purpose, the population units can be partitioned into blocks of neighbouring units and then a unit is selected in the first block, while the remaining units are selected according to the well-known systematic sampling (SS) scheme. Obviously, caution should be used if there is a possibility of a cyclical pattern in the units that might match the sampling frequencies. In order to avoid the loss of efficiency due to spatial regularity, the units can be randomly selected from each block according to a stratified sampling (STS). However, stratified sampling reduces but does not eliminate the chance of selecting neighbouring units and the same drawback also holds for cluster sampling (CS), where only a given number of blocks are selected and a unit is subsequently randomly selected from each block. An alternative procedure which seems to be more attractive is proposed by Fattorini and Ridolfi (1997). The authors propose a modification of sequential sampling designs in such a way that, in presence of positive spatial autocorrelation, at each drawing the probabilities of selecting units surrounding the previously selected units are reduced according to a prefixed factor.

4. Some Monte Carlo simulations

In this section, simulation results are illustrated using two clustered populations, both of $N=20 \times 20=400$ squared units, described by Thompson (1992, p.285 and p.301). The conventional designs compared with adaptive sampling (AS) are simple random sampling without replacement (SRSWOR), SS, STS and CS, where 100 clusters has

been formed by grouping 4 neighbouring units. Moreover, the modified sequential sampling for positive spatial autocorrelation (MSSPSA) proposed by Fattorini and Ridolfi (1997) is also considered. In this design at the first drawing each unit has the same probability of being selected, while at the subsequent drawings the probabilities of selecting units surrounding those already included in the sample are reduced to zero. For each alternative conventional sampling design the Horvitz-Thompson estimator of the population total T is considered. Since the computation of inclusion probabilities is cumbersome when adopting MSSPSA, they have been approximated by using a Monte Carlo simulation based on the selection of M independent samples from both populations. It must be pointed out that, for a sufficiently large M , the Horvitz-Thompson estimator computed using the estimated inclusion probabilities is proven to be unbiased, so that M has been chosen equal to 100,000.

In order to check the performance of the considered procedures, 100,000 samples are selected with sampling size equal to the empirical expectations of the sample size n under the adaptive design. To this purpose, when adopting SS or STS the populations have been divided into n blocks or strata respectively and one unit has been selected from each of them. It is worth noting that the considered partition is far from being optimal, since nothing is supposed to be known about the population structure. Useful suggestions about forming blocks can be found in literature (see *e.g.* D'Orazio, 1997). The performed strategies are listed in the first column of the following tables, where the efficiency is computed with respect to the variance of adaptive sampling, while the accuracy is the quote of estimates falling in the interval $(T-0.5T, T+0.5T)$, in such a way that it represents the probability of obtaining a relative estimate error lower or equal to 0.5.

Table 1: Empirical comparison of various sampling design. Population 1 ($\beta=0.2829$).

sampling design	$n=5$		$n=10$		$n=20$		$n=25$	
	efficiency	accuracy	efficiency	accuracy	efficiency	accuracy	efficiency	accuracy
AS	1.0000	0.0000	1.0000	0.0000	1.0000	0.1050	1.0000	0.1161
SRSWOR	1.1794	0.0612	1.2021	0.1410	1.1459	0.2852	1.2020	0.2847
SS	1.0493	0.0748	0.9117	0.1987	0.6731	0.4996	1.3816	0.3128
STS	1.1426	0.0621	1.2196	0.1569	1.0682	0.3062	1.0542	0.3198
CS	1.1605	0.0617	1.1510	0.1436	1.0640	0.3046	1.1042	0.3071
MSSPSA	1.1907	0.0637	1.1160	0.1494	0.9916	0.3023	1.0219	0.3246

Table 2: Empirical comparison of various sampling design. Population 2 ($\beta=0.4486$).

sampling design	$n=5$		$n=10$		$n=20$		$n=25$	
	efficiency	accuracy	efficiency	accuracy	efficiency	accuracy	efficiency	accuracy
AS	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000	0.3112
SRSWOR	0.8715	0.0728	0.9018	0.1516	0.7995	0.2355	0.8344	0.3292
SS	0.7363	0.0876	0.5807	0.2018	0.5417	0.3502	0.6314	0.4359
STS	0.8562	0.0758	0.8481	0.1598	0.7641	0.2400	0.7645	0.3434
CS	0.8617	0.0727	0.8728	0.1542	0.7452	0.2416	0.7581	0.3469
MSSPSA	0.8560	0.0927	0.8367	0.1599	0.6762	0.2550	0.6711	0.3758

Moreover, the Moran (1950) correlation coefficient β is computed for both populations as an index of spatial homogeneity among contiguous units and it is reported for each table. It must be noted that the results for $n=10$ are really compared with those obtained from adaptive sampling of expected sample size equal to 11.

From the analysis of the results it is at once apparent that the performance in terms of

efficiency of the different sampling strategies depends on the spatial structure of the populations, even if both of them show a highly clustered pattern. Indeed, the adaptive sampling variance in the first population is generally lower than the variance of the alternative strategies, with the exception of systematic sampling, whose behaviour however reveals all the well-known limits of this sampling scheme, in presence of spatial regular patterns. On the other hand, in the second population the adaptive sampling efficiency is always lower than those of the remaining designs.

As to the comparison in terms of accuracy, the performance of adaptive sampling is rather disappointing, since for any sample size the probability of obtaining an estimate in the interval $(T-0.5T, T+0.5T)$ is the lowest and equals zero till n increases to 20 or 25. Moreover, it is worth noting that the estimated probability distribution is quite anomalous showing a U-shape form which reveals a very high probability of obtaining estimates equal or close to zero or rather higher than the true parameter.

5. Concluding remarks

On the basis of the simulation study, whose results are far from being general, it is at once apparent that the use of adaptive sampling is not always advisable, although in presence of rare and clustered populations. Indeed, it may not produce gains in terms of efficiency or accuracy, even if compared to very simple well-known sampling designs. Moreover, the final sample size for adaptive sampling heavily depends on the patchiness of the population and can be considerably larger than the initial sample size. Thus, when sampling is very expensive and it may be necessary to predict the overall number of units to be selected fairly accurately, some methods for limiting the total sampling effort have to be introduced. However, these techniques (Brown and Manly, 1998, Christman and Lan, 2001) generally introduce bias in the estimation of both the parameter of interest and its variance as well as they reduce the potential gains of efficiency of the adaptive strategy. All the previous considerations may be helpful in explaining the reasons why, even if theoretically deeply investigated, adaptive sampling has been rarely adopted by applied researchers for planning spatial surveys. However, adaptive sampling surely offers a suitable solution in presence of neighbourhood defined not in terms of spatial proximity and whenever the aim of the survey is not only obtaining parameter estimates but also observing as many individuals as possible.

References

- Brown J.A., Manly, B.J.F. (1998) Restricted adaptive cluster sampling, *Environmental and Ecological Statistics*, 5, 49-63.
- Christman M.C., Lan, F. (2001) Inverse adaptive sampling, *Biometrics*, 57, 1096-1105.
- D'Orazio M. (1997) A sequential sampling procedure for clusters of territorial units, *Metron*, Vol. LV n. 1-2, 73-91.
- Fattorini L., Ridolfi, G. (1997) A sampling design for areal units based on spatial variability, *Metron*, Vol. LV n. 1-2, 59-72.
- Moran P.A.P. (1950) Notes on continuous stochastic phenomena, *Biometrika*, 37, 17-23.
- Thompson S. K. (1992) *Sampling*, John Wiley and Sons.
- Thompson S. K., Seber, G. A. F. (1996) *Adaptive Sampling*, John Wiley and Sons.