

Adaptive Designs for Soil Contamination

Disegni Informativi per lo Studio dell’Inquinamento del Suolo

Alessandra Petrucci, Emilia Rocco

Dipartimento di Statistica “G. Parenti”, Viale Morgagni, 59 Firenze

alex@ds.unifi.it - rocco@ds.unifi.it

Riassunto: Lo scopo di questo lavoro è di valutare la strategia di campionamento “adattiva-inversa condizionata” rispetto al campionamento inverso e a quello casuale semplice in un’applicazione a dati “pseudo reali”. La popolazione oggetto di indagine è costituita da una griglia di celle quadrate sovrapposta alla mappa di concentrazione del cadmio in una zona della Svizzera nord-orientale. Come molti altri fenomeni ambientali la distribuzione della concentrazione di cadmio nel suolo presenta le caratteristiche spaziali di rarità e aggregazione. Si vuole quindi mettere in evidenza come per tali fenomeni i campionamenti informativi sono da preferire a quelli convenzionali sia per la stima dell’intensità del fenomeno sia per la localizzazione degli hot-spots.

Keywords: Adaptive cluster sampling, constrained inverse adaptive cluster sampling, inverse sampling

1. Introduction

In environmental pollution analysis a common situation is designing an area sample in order to estimate the pollution level in a study area, where the pollution is usually very little or negligible in most sampled sites and it is highly concentrated in a few scattered small areas. Adaptive sampling, introduced by Thompson (1990), could provide a good approach to the problem. The basic idea of this sampling procedure is that the direction at any stage is determined by the information that comes from the sampling. To achieve this objective an initial sample of units is selected and, whenever the value of the variable of interest satisfies a specified condition, neighboring units are added to the sample. The adaptive cluster sampling of Thompson, however, doesn’t consider the possibility that no unit in the initial sample would satisfy the condition for extra sampling and so no information would be gathered on the variable of interest. Otherwise the constrained inverse adaptive cluster sampling (Rocco 1999), through a sequential selection of the initial sample, ensure the presence in the initial sample of at least one unit satisfying the condition for extra sampling.

The aim of this paper is to provide an evaluation of constrained inverse adaptive cluster sampling respect to the inverse sampling and respect to the simple random sampling to estimate the mean concentration of cadmium in the soil, through a simulation experiment where samples are drawn from a grid of 18x30 equal sized quadrants.

The structure of the study population are described in the section 2, some definitions concepts of constrained inverse adaptive sampling procedure are recalled in paragraph 3, finally, the paragraph 4 is devoted to the description of the simulation procedure used and on reporting the results and the comments.

2. Target Population

A variety of adaptive cluster sampling has been proposed but a few applications are carried out on continuous environmental population (i.e. heavy metal soil contamination). The reason is the practical difficulty in organizing survey that use adaptive selection. Although we don't conduct a direct survey of the cadmium concentration we select and observe the concentration of cadmium in the cells of a grid of 18x30 equal sized quadrants (400x400 m) superimposed to a map of cadmium concentration made by disjunctive kriging (Von Steiger *et al.*, 1996) from some soil sampled points in the area of interest around a town in the north-east of Switzerland. These points were sampled to find the areas in which the cadmium concentration exceeded the federal guide value (0.8 mg kg^{-1}).

3. Sampling Strategies

In constrained inverse adaptive cluster sampling, an initial simple random sample of size l is selected and the variable of interest y_i is observed. If at least two of the observed values y_i , satisfy a specified condition of interest C_0 , the procedure for selecting the initial sample is stopped and what follows is identical to what happened in the adaptive strategy proposed by Thompson (1990). Otherwise, the sampling is carried on in a sequential way until at least two units satisfying C_0 are selected. The last unit, that is the second satisfying C_0 , may either be retained or rejected from the sample. Therefore, if the number of units selected in the initial sample is greater than l , we have two possible initial samples, s^n and s^{n+1} . From these initial samples, using the same mechanism of adaptive addition of units used by Thompson we can obtain respectively the final samples s_F^n and s_F^{n+1} . If we denote with T_0 the modified Horvitz-Thompson estimator proposed by Thompson (1990) to estimate the mean of a population using an adaptive cluster sample obtained from an initial simple random sampling, it is easy to note that it is unbiased only for the final sample s_F^l . So for the sample s_F^n and s_F^{n+1} we consider two other estimators (one for each of them). These estimators are both obtained using the Rao Blackwell theorem, conditioning T_0 on two different statistics that are sufficient but not minimal. The two estimators are:

$$1. \quad T(s_F^n) = \frac{1}{n!} \sum_{p \in \Pi} T_0(s_{Fp}^l)$$

where n is the size of the initial sample s^n , Π is the set of all permutations of s^n , $T_0(s_{Fp}^l)$ is the value of T_0 when the initial sample consist of the permutation p of the n units in s^n and then T_0 is applied to the first l units of this permutation.

$$2. \quad T(s_F^{n+1}) = I_{(n+1=l)} \frac{\sum_{\Pi} T_0(s_{Fp}^l)}{l!} + I_{(n+1>l)} \frac{\sum_{\Pi^1} T_0(s_{Fp^1}^l) + \sum_{\Pi^2} T_0(s_{Fp^2}^l)}{2n!}$$

where, if $n+1$ is equal to l , Π is the set of all permutations of s^{n+1} ; if instead, $n+1$ is greater than l , Π^1 and Π^2 are the sets of the permutations of the first n terms of s^{n+1} among which there are, respectively, either the first or the second units satisfying C_0 .

For the inverse sampling we have considered the following estimators:

$$T_{invM} = \frac{1}{k} \sum_{h=1}^k T_{inv}(s_h)$$

where $T_{inv} = \sum_{i=1}^n y_i / n$, $n+1$ denotes the number of units selected and n the number of units considered in T_{inv} in order to estimate the mean (all without the last). T_{invM} works with the same logic as T_M . The last unit can be every one of the units satisfying the condition selected in the inverse sampling, let s_h denote the inverse sample minus the h -th unit from which satisfying the condition for extra sampling.

Each of the estimator considered is unbiased and for each are available an expression for the variance and an unbiased estimator of the variance.

4. Results and Remarks

Five Monte Carlo experiments were carried out in order to compare the different estimators: 1) the population was sampled 1.000 times using constrained inverse adaptive cluster sampling. The condition for extra sampling was a concentration of cadmium greater then $0.8 mg kg^{-1}$ and the minimal size of the initial sample was $l=36$ (exactly the expected number of units necessary to select at least two with a concentration of cadmium greater then threshold), for each time the two type of sample s_F^n and s_F^{n+1} were constructed and the corresponding estimators were calculated; 2) 1000 simple random sampling of size equal to mean size of s_F^n were selected and for each one the sample mean was calculated; 3) 1000 simple random sampling of size equal to mean size of s_F^{n+1} were selected and for each one the sample mean was calculated; 4) 1000 inverse samples containing a number of units with a level of cadmium's contamination greater then $0.8 mg kg^{-1}$ equal to the expected number of this type of units in s_F^n were selected and for each one the estimator T_{invM} was calculated; 5) 1000 inverse samples containing a number of units with a level of cadmium's contamination greater then $0.8 mg kg^{-1}$ equal to the expected number of this type of units in s_F^{n+1} were selected and for each one the estimator T_{invM} was calculated. In each Monte Carlo experiment for each sample also the total size of the sample and the number of relevant units selected is calculated.

In the following table some results of the Monte Carlo experiments are given. Before to comment the results we have to underline that the relative efficiency of an estimator in comparison to another is evaluated using the ratio between the relative mean square error multiplied by the expected size of corresponding sample (through variability for observation cost).

Table1: Results of Monte Carlo experiments

	$\sqrt{MSE[.]}$	$\bar{n}$	$\sqrt{MSE[.] \cdot \bar{n}}$	$\bar{n} \cdot y_i > c$	% camp. vuoti	$eff(T)$	$eff(T_M)$
T	0.0396	77.0650	3.0537	18.1170	0	1	1.0357
T_M	0.0394	80.3180	3.1629	19.6320	0	0.9655	1
$\bar{y}$	0.0336	77	2.5901	4.0380	1.4	1.1790	-
$\bar{y}^*$	0.0332	80	2.6573	4.2230	0.7	-	1.1903
T_{invM}	0.0124	323.9890	4.0106	18	0	0.7614	-
T_{invM}^*	0.0107	361.3270	3.8653	20	0	-	0.8183

The estimator T is more efficient than the estimator associated to the inverse sampling containing the same number of relevant values but it is also more efficient than T_M despite the fact that it is based on less information. In terms of relative efficiency the constrained inverse adaptive cluster sampling doesn't bear comparison with simple random sampling, both T and T_M are less efficient of the sample's mean of the simple random samples having the same expected size. However the simple random samples have a significant drawback respect to all the other considered strategies: they include less then 15% of the 29 relevant units (units exceeding the federal value) of the population whereas the others sampling procedures more than 62% and the number of observed relevant units is extremely important in environmental pollution analysis to know the spatial distribution of a pollutant. In contrast to the other considered strategies, the simple random samples have also a non zero probability of including at least one relevant units. Even the Thompson's adaptive cluster sampling is more efficient than the constrained inverse one being not affected by the variability of the initial sample size. On the other hand, however, the possibility for the Thompson's adaptive cluster sampling of gathering the final sample – which could coincide with the initial sample – without information on the variable of interest is a serious problem. To show this problem, another Monte Carlo experiment has been carried out selecting 1000 adaptive cluster samples with an initial simple random sample of 36 units and, in this case, 130 of them do not contain any significant value of the variable of interest.

While the Thompson's adaptive cluster sampling is ever more efficient than the constrained inverse one, the higher efficiency of the simple random sampling depends on the pattern of the population.

Usually the expected number of units necessary to select at least two relevant values is unknown. In a Monte Carlo experiment with a fewer initial sample size ($l=20$) we achieved equivalent results.

References

- Rocco, E. (1999), *Constrained Inverse Adaptive Cluster Sampling*, Quaderni del Dipartimento di Statistica "G. Parenti", n.47, Università degli studi di Firenze.
- Thompson, S. K. (1990), Adaptive cluster sampling, *Journal of the American Statistical Association*, 85, 1050-1059.
- Von Steiger B., Webster R., Schulin R. and Lehmann R. (1996), Mapping Heavy Metals in Polluted Soil by Disjunctive Kriging, *Environmental Pollution*, 94, 2, 205-215.