Journal of Official Statistics, Vol. 19, No. 1, 2003, pp. 45–57

Constrained Inverse Adaptive Cluster Sampling

Emilia Rocco[1]

Adaptive cluster sampling can be a useful design for sampling rare and clustered populations. In this article a new adaptive cluster sampling, which is an extension of the classical one, is suggested. It is denominated constrained inverse adaptive cluster sampling and its distinctive characteristic is to make sure that the initial sample contains at least one unit satisfying the condition for extra sampling. This is achieved by means of a sequential selection of the initial sample. This sort of selection of the initial units introduces a bias into the estimators of the mean of the population usually used in the adaptive cluster sampling. To overcome this difficulty two new unbiased estimators of the mean of the population are suggested in the article. The expressions of their variance and of their sample variance estimators are also proposed. To study the properties of the proposed strategies a simulation study is carried out.

Key words: Rare and clustered populations; sequential selection.

1. Introduction

Informative sampling (also known as adaptive sampling) designs are those in which the procedure for selecting units may depend on values of the variable of interest or on values of any other variable observed during the survey. For rare and clustered populations, such as populations examined in many environmental and natural resources studies, they can produce gains in precision compared to conventional designs. For studies of hidden human populations, such as injection drug users and others at risk of HIV transmission, adaptive link-tracing designs often provide the only practical way to obtain a large enough sample.

A particular informative design, which is applicable in either case is the adaptive cluster sampling first proposed by Thompson (1990, 1991a, 1991b, 1992).

In adaptive cluster sampling an initial sample of units is selected and, whenever the value of the variable of interest satisfies a specified condition, neighbouring units are added to the sample. The condition for extra sampling might be the presence of rare animal or plant species, detection of ''hot spots'' in an environmental pollution study, infection with a rare disease in an epidemiology study or observation of a rare characteristic of interest in a household or business survey. The neighbourhood of a unit may be defined by spatial proximity or, in the case of human populations, by social or genetic links or other connections.

Different types of adaptive cluster sampling have been proposed (Thompson 1990, 1991a, 1991b, 1992, 1993, 1996, 1997, 1998; Thompson and Seber 1994, 1996;

[1] University of Florence, Viale Morgagni, 59, I-50134 Firenze, Italy. Email: rocco@ds.unifi.it

Thompson and Frank 1998; Brown 1994; Brown and Manly 1998; Smith, Conroy and Brakhage 1995; Chao and Thompson 1997; Dryer and Thompson 1998; Roesh 1993; Salehi and Seber 1997a, 1997b) and their advantages have been pointed out. However, the possibility that no unit in the initial sample would satisfy the condition for extra sampling is, according to our research, a problem that is, for the most part, untouched. If this were to happen, the adaptive cluster sampling would coincide with the initial sample and no information on the distribution of the relevant values of the variable of interest would be gathered. Either taking a larger initial sample or setting a less restrictive condition for extra sampling (but sufficient to discover the relevant cluster in the population, for example to locate the areas in which a pollutant exceeds a dangerous threshold) would reduce this drawback but not always avoid it. In this article, a new adaptive cluster design, which is an extension of the adaptive cluster sampling of Thompson, is introduced. We denominated it constrained inverse adaptive cluster sampling (CIAC) and its aim is to ensure the presence in the initial sample of at least one unit satisfying the condition for extra sampling. This is achieved by a sequential selection of the initial sample. This kind of selection of the initial units, explained in detail in the next section, introduces a bias into the estimators of the population mean usually used in the adaptive cluster sampling. To overcome this difficulty two new unbiased estimators of the population mean are suggested in the article. In order to obtain unbiased estimators, however, it is not sufficient to include in the initial sample one unit satisfying the condition for extra sampling. Rather it is necessary to include at least two units. Thus, in practice, in the CIAC procedure the initial selection process does not stop until the second relevant unit is included in the sample.

The expressions of the variance of the two estimators and of their sample variance estimators are also proposed. Finally, the relative efficiency of the new strategy, compared to simple random sampling and to inverse sampling, is empirically evaluated.

2. Sampling Scheme

The cluster adaptive sampling design proposed by Thompson can be briefly described as follows. An initial probability sample of fixed size n is selected. For each selected unit the variable of interest y is observed and if the observed value y_i ($i = 1, 2, \ldots, n$) satisfies a condition of interest C_0 (specified a priori), additional units in the neighbourhood of the ith unit are sampled. If the condition is met in any units of the ith neighbourhood, then their neighbourhoods will be also sampled. This is repeated until the condition is not met for any adaptively sampled units. The result is a sample of n clusters. Each cluster has a core of units satisfying the condition C_0 called network and a boundary of units called ''edge units'' which do not satisfy C_0. The units of the initial sample which do not satisfy the condition C_0 are size-one clusters. Finally, any unit that does not satisfy C_0 is defined as a network of size one (this means that any cluster of size one is a network of size one and that any edge unit is also a network of size one).

The proposed sampling scheme is different from Thompson's in the sampling design used for the selection of the initial sample. It can be described as follows. Let y be the study variable, let l denote a minimal size for the initial sample and let C_0 be the condition for extra sampling. Assume l units are selected by simple random sampling. If among these units at least two satisfy C_0, then the procedure for selecting the initial sampling is stopped

and what follows is identical to what happens in the adaptive cluster strategy proposed by Thompson. Otherwise, the sampling is carried on in a sequential way – that is, one unit is added to the initial sample at each step, 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 larger than l, we can have two possible samples: the sample s^n which does not include the last selected unit and the sample s^{n+1} which includes all the selected units. 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}.

3. Estimators

Let us consider the possible CIAC samples:

i) s_F^n obtained from the initial sample s^n
ii) s_F^{n+1} obtained from the initial sample s^{n+1}

For each of them we shall define an unbiased estimator of the population mean. These will be described respectively in Subsections 3.1 and 3.2.

3.1. Estimator related to sample s^n

The initial sample s^n can be thought of as a simple random sample without replacement of l units (the first l selected units), with the possible addition of other units sequentially selected. No additional units are selected if two or more units among the first l satisfy C_0. Otherwise, the additional units are all the units selected before the second one that satisfies C_0.

Let s^l denote the first l selected units and s_F^l the adaptive cluster sample with initial sample s^l. If T_0 is an unbiased estimator of the population mean for an adaptive cluster sample obtained from an initial simple random sample, then

$$T'(s_F^n) = T_0(s_F^l) \tag{1}$$

is an unbiased estimator of the mean of the population for our sample s_F^n. However, when the size of s^n is larger than l, T_0 (being calculated on the final sample obtained from the initial sample formed by only the first l units selected) does not take into account the only nonunitary network in s_F^n if this network is not intersected from one of the first l selected units. Therefore, we propose a new estimator that depends on the complete information in s_F^n. This new estimator is obtained taking the expected value of T_0 conditional on a suitable sufficient statistic. In other words, it is obtained by an application of the Rao-Blackwell theorem to T_0 (Rao 1945, Blackwell 1947).

Let

$$d^n = \{(i, y_i), i \in s^n\} \tag{2}$$

be the set of distinct data in the initial sample. This set is a function of the set of reduced data associated to s_F^n and since this last set is a minimal sufficient statistic for $\theta = (y_1, y_2, \ldots, y_N)$, d^n is a sufficient statistic for the same parameter. The new unbiased

estimator is

$$T(s_F^n) = E\left\lfloor T_0(s_F^l) \mid d^n \right\rfloor \tag{3}$$

Let n denote the size of the initial sample, s^n, and $T_0(s_{F\pi}^l)$ the value of T_0 when the initial sample consists of the permutation π of the n units in s^n (and then T_0 is applied to the first l units of this permutation). The number of permutations is $n!$ and conditionally on d^n each of these is equally likely. Hence an explicit expression of $T(s_F^n)$ is

$$T(s_F^n) = \frac{1}{n!} \sum_{\pi \in \Pi} T_0(s_{F\pi}^l) \tag{4}$$

where Π is the set of all the permutations of s^n. If T_0 is invariant for the permutation (i.e., its value does not depend on the position of the initial sample units) the Expression (4) can be written as follows:

$$T(s_F^n) = \frac{1}{\binom{n}{l}} \sum_{c \in C} T_0(s_{Fc}^l) \tag{5}$$

where C is the set of all possible combinations of l units from the n in the initial sample, c is any of the possible combinations and $T_0(s_{Fc}^l)$ the value of T_0 when the initial sample is the combination c. If we define

$$I_{(n=l)} = \begin{cases} 1 & \text{if } n = l \\ 0 & \text{if } n > l \end{cases} \quad \text{and} \quad I_{(n>l)} = 1 - I_{(n=l)} \tag{6}$$

Expression (5) becomes

$$T(s_F^n) = I_{(n=l)} T_0(s_F^n) + I_{(n>l)} \frac{\sum_C T_0(s_{Fc}^l)}{\binom{n}{l}} \tag{7}$$

Also the expression of $\operatorname{var}\lfloor T(s_F^n) \rfloor$ can be obtained by the Rao-Blackwell theorem. If $\operatorname{var}[T_0]$ is the variance of T_0, we have

$$\operatorname{var}[T(s_F^n)] = \operatorname{var}[T_0] - \mathrm{E}\left[\frac{\sum_\Pi [T(s_F^n) - T_0(s_{F\pi}^l)]^2}{n!} \right] \tag{8}$$

and if T_0 is invariant for the permutation

$$\operatorname{var}[T(s_F^n)] = \operatorname{var}[T_0] - \mathrm{E}\left[\frac{\sum_C [T(s_F^n) - T_0(s_{Fc}^l)]^2}{\binom{n}{l}} \right] \tag{9}$$

To find an unbiased estimator of $\operatorname{var}\lfloor T(s_F^n) \rfloor$, we need unbiased estimators of the two terms in (9). If $\hat{\operatorname{var}}[T_0]$ denotes an unbiased estimator of $\operatorname{var}[T_0]$, an unbiased estimator of $\operatorname{var}\lfloor T(s_F^n) \rfloor$ is

$$\hat{\operatorname{var}}[T(s_F^n)] = \hat{\operatorname{var}}[T_0] - \frac{\sum_\Pi [T(s_F^n) - T_0(s_{F\pi}^l)]^2}{n!} \tag{10}$$

and if T_0 is invariant for the permutation

$$\hat{\text{var}}[T(s_F^n)] = \hat{\text{var}}[T_0] - \frac{\sum_C [T(s_F^n) - T_0(s_{Fc}^l)]^2}{\binom{n}{l}} \tag{11}$$

Since $\text{var}[T_0]$ is also a function of θ and $\hat{\text{var}}[T_0]$ denotes an unbiased estimator, we can apply the Rao-Blackwell theorem once again to obtain another unbiased estimator of $\text{var}\lfloor T(s_F^n)\rfloor$ with a smaller variance, namely

$$\hat{\text{var}}'[T(s_F^n)] = \frac{\sum_\Pi (\hat{\text{var}}[T_0(s_{F\pi}^l)] + [T(s_F^n) - T_0(s_{F\pi}^l)]^2)}{n!} \tag{12}$$

and if T_0 is invariant for the permutation

$$\hat{\text{var}}'[T(s_F^n)] = \frac{\sum_C (\hat{\text{var}}[T_0(s_{Fc}^l)] + [T(s_F^n) - T_0(s_{Fc}^l)]^2)}{\binom{n}{l}} \tag{13}$$

It should be noted that both the estimators of the variance could produce negative estimates with some samples in some data sets.

T can be calculated from any unbiased estimator of the population mean in the adaptive cluster sampling with initial random sample without replacement. Two possible estimators of this type are the modified version of the Horvitz-Thompson estimator and the modified version of the Hansen-Hurvitz estimator proposed by S. K. Thompson (1990). For both these estimators Thompson has proposed an expression of their variance and an unbiased estimator of their sample estimate. The estimator T_0 used in the applications of Section 4 is the modified version of the Horvitz-Thompson estimator proposed by S. K. Thompson.

3.2. Estimator related to the sample s_F^{n+1}

In contrast to s^n, which does not include the last unit selected in the initial sample when these are more than l, s^{n+1} always includes all the selected units. Note that when the units in s^{n+1} are more than l, two and only two units satisfy the condition C_0 and one of the two is necessarily the last unit. So, in contrast to the units in s^n that are permutable in all ways, the units in s^{n+1} are not permutable in all ways. The last unit can be changed only if the other satisfies C_0. But, after having made this change, we can again consider all the possible permutations of the first n units. This observation can be used to define another estimator of the population mean which is based on all the information in s^{n+1} and takes into account the nonunitary network intersected by the second selected unit satisfying C_0 even when this unit is the last of the initial sample and is selected after the first l. This estimator, as well as T, is obtained by taking the expected value of T_0 conditional on a sufficient statistic. The subset of units in s^{n+1} satisfying C_0 is denoted by $s_{C_0}^{n+1}$ and the statistic considered is

$$d^{n+1} = \begin{cases} \{(i, y_i),\, i \in s^{n+1}\} & \text{if } n+1 = l \\ \{\{(i_1, y_{i_1}), \ldots, (i_n, y_{i_{1_n}})\}, (i_{n+1}, y_{i_{1_{n+1}}}) \text{ with } i_{n+1} \in s_{C_0}^{n+1}\} & \text{if } n+1 > l \end{cases} \tag{14}$$

that is, the set of distinct and unordered data in s^{n+1} if the size of s^{n+1} is l, and the set of

the distinct and unordered data in s^{n+1} but so that the last unit satisfies C_0 if the size of s^{n+1} is larger than l. Furthermore d^{n+1} is a function of the reduced data associated with s_F^{n+1}. So it is also a sufficient statistic for θ as well as being d^n. Then, the new estimator is

$$T_M(s_F^{n+1}) = \mathrm{E}[T_0(s_F^l)\,|\,d^{n+1}] \tag{15}$$

When the size of s^{n+1} is larger than l, the number of samples compatible with d^{n+1} is $2(n!)$, that is, all the permutations of the first n values including in turn one of the two units satisfying C_0. When the size of s^{n+1} is equal to l, s^{n+1} is equal to s^n and d^{n+1} is equal to d^n, and thus the number of possible permutations of data is $l!$. In either case each possible permutation is equally likely. It follows that an explicit expression for T_M is

$$T_M(s_F^{n+1}) = I_{(n+1=l)}\frac{\sum_{\Pi} T_0(s^l_{F\pi})}{l!} + I_{(n+1>l)}\frac{\sum_{\Pi^1} T_0(s^l_{F\pi^1}) + \sum_{\Pi^2} T_0(s^l_{F\pi^2})}{2n!} \tag{16}$$

where, if $n+1$ is equal to l, Π is the set of all permutations of s^{n+1}. If, instead, $n+1$ is larger 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 unit satisfying C_0.

If T_0 is invariant for the permutation, then

$$T_M(s_F^{n+1}) = I_{(n+1=l)}T_0(s^l_{Fc}) + I_{(n+1>l)}\frac{\sum_{C^1} T_0(s^l_{Fc^1}) + \sum_{C^2} T_0(s^l_{FC^2})}{2\binom{n}{l}} \tag{17}$$

where C^1 and C^2 are the sets of combinations corresponding to those of permutations Π^1 and Π^2. T_M is obviously unbiased and its variance is

$$\mathrm{var}[T_M(s_F^{n+1})] = \mathrm{var}[T_0] - \frac{1}{\#(\Pi^*)}E\left[\sum_{\Pi^*}(T_M(s_F^{n+1}) - T_0(s^l_{F\pi^*}))^2\right] \tag{18}$$

where Π^* denotes the set of possible permutations,

$$\Pi^* = \begin{cases} \Pi & \text{if } n+1 = l \\ \Pi^1 \cup \Pi^2 & \text{if } n+1 > l \end{cases}$$

If T_0 is invariant for the permutation

$$\mathrm{var}[T_M(s_F^{n+1})] = \mathrm{var}[T_0] - \frac{1}{\#(C^*)}E\left[\sum_{C^*}(T_M(s_F^{n+1}) - T_0(s^l_{Fc^*}))^2\right] \tag{19}$$

where C^* is the set of possible combinations.

Two unbiased estimators of $\mathrm{var}[T_M]$ are

$$\hat{\mathrm{var}}[T_M(s_F^{n+1})] = \hat{\mathrm{var}}[T_0] - \frac{1}{\#(\Pi^*)}\sum_{\Pi^*}(T_M(s_F^{n+1}) - T_0(s^l_{F\pi^*}))^2 \tag{20}$$

$$\hat{\mathrm{var}}[T_M(s_F^{n+1})] = \frac{1}{\#(\Pi^*)}\sum_{\Pi^*}(\hat{\mathrm{var}}[T_0(s^l_{F\pi^*})] - (T_M(s_F^{n+1}) - T_0(s^l_{F\pi^*}))^2) \tag{21}$$

To obtain their expressions in the case in which T_0 is unchangeable for permutation, it is sufficient to consider C^* instead of Π^*.

It is easy to verify that T_M is only the unweighted mean of the two estimators that we denote with T_1 and T_2 and that they are equal to T if the units selected in the initial sample

are l. Otherwise, if the units selected in the initial phase are more than l, T_1 is the estimator T applied to the part of s_F^{n+1} obtained from the first n selected units and T_2 is the estimator T applied to the part of s_F^{n+1} obtained from the first n units of the initial sample after the last unit (the second unit selected satisfying C_0) has been substituted with the first unit selected satisfying C_0.

4. A Simulation Study

A simulation was used to study the properties of the CIAC. Ten patchy populations were simulated using a Poisson cluster process model (Neyman 1939; Neyman and Scott 1958; Cressie 1991) within a defined study site divided into 30×30 equal sized quadrants. The number of clusters in the study site was a random variable from a Poisson distribution with a mean equal to 4. Cluster centres were randomly located in the study site. The number of individuals per cluster was a random variable from a Poisson distribution with a mean equal to 90. Each individual was located at a radial unitary distance from the centre of the cluster selected from a normal distribution uniformly distributed between 0° and 360°.

For each population, three Monte Carlo experiments were carried out in order to compare the estimators related to the CIAC sampling with the sample mean related to the simple random sampling and two other estimators related to the inverse sampling. What we did is described in detail in the following three items:

i) Each population was sampled 10,000 times using constrained inverse adaptive sampling. The condition for extra sampling was C_0: $y > 0[\dots]$ and the minimal size of the initial sample was 50. In detail, from each population 10,000 initial samples were selected using the design described in Section 2. Not including in the initial sample the last selected unit when the selected units were more than 50, we obtained the samples s^n from which we had the final samples s_F^n. From s_F^n we estimated the mean using the estimator T. Including in the initial sample all the selected units also when these were more than l we obtained the samples s^{n+1} from which we had the final samples s_F^{n+1}. From s_F^{n+1} we estimated the mean using the estimator T_M. The modified version of the Horvitz-Thompson estimator was used as the starting estimator to calculate T_M as well as T. At the end we had 10,000 samples s_F^n and 10,000 samples s_F^{n+1}. The sample size and the number of sampled units satisfying the condition for extra sampling were also calculated for each sample s_F^n and for each sample s_F^{n+1}.

ii) Each population was sampled 20,000 times using simple random sampling without replacement. 10,000 simple random samples equal in size to the expected size (empirically evaluated from the data described in the previous item and obviously including all the units selected, not only those satisfying C_0 but also the edge units) of the samples s_F^n, and a further 10,000 simple random samples equal in size to the expected size of the samples s_F^{n+1} were selected. The sample mean of the first 10,000 simple random samples was compared to the estimator T, whilst the sample mean of the other 10,000 simple random samples was compared to the estimator T_M.

iii) Each population was sampled 20,000 times using the inverse sampling without

replacement. 10,000 inverse samples containing a number of units with relevant values of the variable of interest (values satisfying the condition for extra sampling) equal to the expected number of relevant values of the variable of interest in s_F^n (empirically evaluated from the data described in the first item). In addition 10,000 inverse samples containing a number of units with relevant values of the variable of interest equal to the expected number of relevant values of the variable of interest in s_F^{n+1} were selected. For each sample of the first 10,000 and for each sample of the additional 10,000 the following two unbiased estimators of the means were calculated:

$$T_{inv} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{22}$$

and

$$T_{invM} = \frac{1}{k} \sum_{k=1}^{k} T_{inv}(s_h) \tag{23}$$

where $n + 1$ denotes the number of units selected and n the number of units considered in order to estimate the mean (just as the constrained inverse adaptive sampling is not used to calculate T, the last selected unit is not used to calculate T_{inv}). T_{invM} works with the same logic as T_M. The last unit can be any of the units satisfying the condition selected. Let k denote the number of units satisfying the condition selected in the inverse sampling, and let s_h denote the inverse sample minus the hth unit from which satisfying the condition for extra sampling. Then T_{invM} is nothing else than the unweighted mean of the k possible values of T_{inv}.

Table 1 gives some properties of the ten populations, denoted by roman numerals from 1 to X. It gives the mean μ, the size N, the number of nonunitary networks (*n. net*), the number of units satisfying the condition for extra sampling ($n.\ y_i > 0$), the total variance (V_{TOT}), the ratio between the variance within and the total variance (V_W/V_{TOT}). Table 2 provides some results of the Monte Carlo experiments for the population IV which is less rare and that where the ratio between the variance within and the total variance is the highest. For the other populations, to avoid prolixity only the Monte Carlo-derived efficiency

Table 1. Some key characteristics of simulated populations

Populations	μ	N	*n. net*	$n.\ y_i > 0$	V_{TOT}	V_W/V_{TOT}
I	4	900	1	6	5116.4	0.5340
II	4	900	3	22	1030.8	0.3788
III	4	900	4	42	585.37	0.4383
IV	4	900	5	45	498.41	0.3884
V	4	900	3	31	762.20	0.4100
VI	4	900	6	40	595.75	0.4175
VII	4	900	3	24	1057.4	0.4472
VIII	4	900	3	24	1066.9	0.4490
IX	4	900	1	9	2900.0	0.5559
X	4	900	4	41	571.8	0.4104

Table 2. Population IV: empirical and theoretical results

	$\sqrt{\text{MSE}[.]}$	$\overline{n}$	$\sqrt{\text{MSE}[.]} * \overline{n}$	$\overline{n}.\, y_i > 0$	% irrelevant	eff$[T]$	eff$[T_M]$
T	2.0874	94.37	196.99	20.18	0	1	0.9627
T_M	2.0822	98.27	204.62	21.98	0	1.0387	1
$\overline{y}$	2.1824	94	205.15	4.72	0.65	1.0414	–
	2.1791	*94*	*204.84*	*4.70*	*0.61*		
$\overline{y}^*$	2.1351	98	209.24	4.92	0.44	–	1.0226
	2.1289	*98*	*208.63*	*4.90*	*0.49*		
T_{inv}	0.9027	392.53	354.34	20	0	1.7989	–
		391.30		*20*	*0*		
T_{invM}	0.8864	392.53	347.94	20	0	1.7662	–
		391.30		*20*	*0*		
T^*_{inv}	0.8071	431.17	348.00	22	0	–	1.7007
		430.43		*22*	*0*		
T^*_{invM}	0.8953	431.17	342.91	22	0	–	1.6758
		430.43		*22*	*0*		

indexes for the two proposed estimators T and T_M are provided in Table 3. Before studying these tables, let us explain the notations used in them:

T: mean estimator related to the sample s_F^n
T_M: mean estimator related to the sample s_F^{n+1}
$\overline{y}$: sample mean of the simple random sample of size equal to the expected size of s_F^n
$\overline{y}^*$: sample mean of the simple random sample of size equal to the expected size of s_F^{n+1}
T_{inv}: mean estimator that does not take into account the last unit selected related to the inverse sample containing a number of relevant units equal to the expected number of relevant units in s_F^n
T_{invM}: mean estimator that takes into account all the units selected related to the inverse sample containing a number of relevant units equal to the expected number of relevant units in s_F^n
T^*_{inv}: mean estimator that does not take into account the last unit selected related to the inverse sample containing a number of relevant units equal to the expected number of relevant units in s_F^{n+1}
T^*_{invM}: mean estimator that takes into account all the units selected related to the inverse sample containing a number of relevant units equal to the expected number of relevant units in s_F^{n+1}
$\sqrt{\text{MSE}[.]}$: root square of the mean squared error of the corresponding estimator
$\overline{n}$: expected size of the sample to which the estimator is related
$\sqrt{\text{MSE}[.]} * \overline{n}$: product of the mean squared error and expected size of the corresponding sample
$\overline{n}.\, y_i > 0$: expected number of units satisfying the condition for extra sampling present in the sample
% [...] irrelevant sample: percent of sample that do not contain any unit satisfying the condition for extra sampling
eff$[T]$: efficiency of the estimator T evaluated as a ratio between $\sqrt{\text{MSE}[T]} * \overline{n}$ and $\sqrt{\text{MSE}[.]} * \overline{n}$, where [.] is in turn one of the estimators to which T is compared

Table 3. Efficiency indexes for the two estimators, T and T_M, empirically evaluated and compared with all the estimators used in the experiments

Populations		T	T_M	$\overline{y}$	$\overline{y}^*$	T_{inv}	T_{invM}	T^*_{inv}	T^*_{invM}
I	eff$[T]$	1	1	1.1994	–	1.1098	0.7750	–	–
	eff$[T_M]$	1	1	–	1.1994	–	–	1.1098	0.7750
II	eff$[T]$	1	1.0719	1.0450	–	1.8086	1.7529	–	–
	eff$[T_M]$	0.9329	1	–	1.0258	–	–	1.5864	1.5837
III	eff$[T]$	1	1.0503	1.1107	–	1.8563	1.8112	–	–
	eff$[T_M]$	0.9521	1	–	1.0780	–	–	1.7480	1.7147
V	eff$[T]$	1	1.0699	1.1190	–	1.8759	1.8268	–	–
	eff$[T_M]$	0.9346	1	–	1.0784	–	–	1.6800	1.6253
VI	eff$[T]$	1	1.0538	1.1189	–	1.8119	1.7734	–	–
	eff$[T_M]$	0.9489	1	–	1.0902	–	–	1.6735	1.6324
VII	eff$[T]$	1	1.0666	1.1251	–	1.9421	1.8809	–	–
	eff$[T_M]$	0.9376	1	–	1.0836	–	–	1.8086	1.7419
VIII	eff$[T]$	1	1.0706	1.0975	–	1.8984	1.8425	–	–
	eff$[T_M]$	0.9341	1	–	1.0597	–	–	1.7194	1.6513
IX	eff$[T]$	1	1	1.2955	–	1.0549	0.7718	–	–
	eff$[T_M]$	1	1	–	1.2955	–	–	1.0549	0.7718
X	eff$[T]$	1	1.0534	1.0805	–	1.8229	1.7807	–	–
	eff$[T_M]$	0.9493	1	–	1.0501	–	–	1.7049	1.6642

eff$[T_M]$: efficiency of the estimator T_M evaluated as a ratio between $\sqrt{\text{MSE}\ [T_M]} * \overline{n}$ and $\sqrt{\text{MSE}\ [\,.\,]} * \overline{n}$, where $[\,.\,]$ is in turn one of the estimators to which T_M is compared.

Notice that Tables 2 and 3 do not include any reference to the expected values of the estimators because they are all unbiased such that their expected values are approximately equal to the mean of the population, which is given for each population in Table 1.

Two other important considerations regarding Tables 2 and 3 are the following:

i) for quantities for which it was possible we have considered the theoretical calculus apart from the empirical. This can be found in italics under the corresponding theoretical one;
ii) the estimator T is compared only to $\overline{y}$ which shares its expected size, to T_{inv} and T_{invM} which share its expected number of units satisfying the condition for extra sampling, and T_M which shares its initial sample. Equivalently, T_M is compared only to $\overline{y}^*$ which shares its expected size, to T^*_{inv} and T^*_{invM} which share its expected number of units satisfying the condition for extra sampling, and to T which shares its initial sample.

Table 4. Some characteristics of a population for which T_M is more efficient than T

	μ	Size	Variability
Population	4.4444	900	644.874
Network 1	26.6667	6	291.222
Network 2	133.714	28	3031.00

5. Discussion

From Tables 2 and 3 it should be noted that, apart from populations I and IX, the most efficient estimator is T. The relative efficiency of an estimator in comparison to another is evaluated using the ratio between the relative mean squared error multiplied by the expected size of the corresponding sample (through variability for observation cost).

T is also more efficient than T_M despite the fact that it is based on less information: it does not consider the last unit selected in the initial sample and the corresponding network. But the increase of the sample size as a result of their consideration is larger than the decrease in the estimator variability obtained by using more information. It should be noted, however, that T is not more efficient than T_M for all possible populations. Table 4 gives some characteristics of a population for which T_M is more efficient than T. For this population the Monte Carlo-derived efficiency index of T_M compared to T is 1.018.

A factor which could always make T_M more efficient than T is the introduction of a cost function which assigns a lower cost of observation to the units belonging to the same network.

In relation to the comparison between T and $\bar{y}$ it is clear that T is more efficient than $\bar{y}$ for all the populations considered. This result is more evident for some populations than for others because it is strictly related to the patchy structure of the considered population. Apart from the variability by unit of observation another factor which makes the constrained inverse adaptive cluster sampling preferable to the simple random sampling is the larger expected number of the relevant units selected (see Table 2).

In relation to the comparison between the CIAC sampling and the inverse sampling, it should be noted that, apart from populations I and IX, the first is more efficient. Populations I and IX include only one network: since in the inverse sampling we want to select the same number of relevant units as in the CIAC sampling, we are selecting almost all the populations in order to capture the units of this network. As a consequence, the variability of the two mean estimators is almost zero.

The relative efficiency of the estimators related to the CIAC sampling in comparison to the sample mean related to the simple random sampling and to the estimators related to the inverse sampling increases if we consider a cost function which assigns a lower cost of observation to the units belonging to the same network.

We do not have to consider the comparison between the estimators associated to the CIAC sampling and the estimators of the mean associated to the adaptive cluster sampling of Thompson. The last are of course more efficient because the same elements work on the variability of both with the exception that also the variability of the size of the initial sample works on the estimators related to the CIAC sampling. But the aim of this work was not to find a strategy more efficient than the adaptive one proposed by Thompson, but to find a strategy that permits us in every case to say something about the study variable. We deem that the method proposed here fulfils this objective.

6. References

Blackwell, D. (1947). Conditional Expectation and Unbiased Sequential Estimation. Annals of Mathematical Statistics, 18, 105–110.

Brown, J.A. (1994). The Application of Adaptive Cluster Sampling to Ecological Studies. In Statistics in Ecology and Environmental Monitoring, 86–97, D.J. Fletcher and B.F.J. Manly (eds). Dunedin: University of Otago Press.

Brown, J.A. and Manly, B.J.F. (1998). Restricted Adaptive Cluster Sampling. Environmental and Ecological Statistics, 5, 49–63.

Chao, C. and Thompson, S.K. (1997). Optimal Sampling Design Under a Spatial Model. Technical Report 97-11. Department of Statistics, Pennsylvania State University.

Cressie, N.A.C. (1991). Statistics for Spatial Data. New York: Wiley.

Dryver, A. and Thompson, S.K. (1998). Improved Unbiased Estimators in Adaptive Cluster Sampling. Technical Report 98-06. Department of Statistics, Pennsylvania State University.

Neyman, J. (1939). On a New Class of ''Contagious'' Distribution, Applicable in Entomology and Bacteriology. Annals of Mathematical Statistics, 10, 35–57.

Neyman, J. and Scott, E.L. (1958). Statistical Approach to Problems of Cosmology. Journal of the Royal Statistical Society, 20, 1–29.

Rao, C.R. (1945). Information and Accuracy Attainable in Estimation of Statistical Parameters. Bulletin of the Calcutta Mathematical Society, 37, 81–91.

Rocco, E. (1999). Constrained Inverse Adaptive Cluster Sampling. Bulletin of the International Statistical Institute, 52nd Session, Contributed Papers, Tome LVII, Book 3, 191–192.

Roesch, F.A. Jr. (1993). Adaptive Cluster Sampling for Forest Inventories. Forest Science, 39, 655–669.

Salehi, M.M. and Seber, G.A.F. (1997a). Adaptive Cluster Sampling with Networks Selected Without Replacement. Biometrika, 84, 209–219.

Salehi, M.M. and Seber, G.A.F. (1997b). Two-Stage Adaptive Cluster Sampling. Biometrics, 53, 959–970.

Sampford, M.R. (1962). Methods of Cluster Sampling With and Without Replacement of Clusters of Unequal Size. Biometrika, 42, 27–40.

Smith, D.R., Conroy, M.J., and Brakhage, D.H. (1995). Efficiency of Adaptive Cluster Sampling for Estimating Density of Wintering Waterfowl. Biometrics, 51, 777–788.

Thompson, S.K. (1990). Adaptive Cluster Sampling. Journal of the American Statistical Association, 85, 1050–1059.

Thompson, S.K. (1991a). Adaptive Cluster Sampling: Designs with Primary and Secondary Units. Biometrics, 47, 1103–1115.

Thompson, S.K. (1991b). Stratified Adaptive Cluster Sampling. Biometrika, 78, 389–397.

Thompson, S.K. (1992). Sampling. New York: Wiley.

Thompson, S.K. (1993). Multivariate Aspects of Adaptive Cluster Sampling. In Multivariate Environmental Statistics, 561–572, G.P. Patil and C.R. Rao (eds). New York: North Holland/Elsevier Science Publishers.

Thompson, S.K. (1996). Adaptive Cluster Sampling Based on Order Statistics. Environmetrics, 7, 123–133.

Thompson, S.K. (1997). Adaptive Sampling in Behavioural Surveys. In The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates, 296–319,

L. Harrison and A. Hughes (eds). NIDA Research Monograph 167, Rockville, MD: National Institute of Drug Abuse.

Thompson, S.K. (1998). Design-based Adaptive Sampling in Graphs. Technical Report 98-01. Department of Statistics, Pennsylvania State University.

Thompson, S.K. and Frank, O. (1998). Model-based Estimation with Link-Tracing Sampling Design. Technical Report 98-01. Department of Statistics, Pennsylvania State University.

Thompson, S.K. and Ramsey, F.L. (1992). An Adaptive Procedure for Sampling Animal Populations. Biometrics, 48, 712–724.

Thompson, S.K. and Seber, G.A.F. (1994). Delectability in Conventional and Adaptive Sampling. Biometrics, 50, 1195–1199.

Thompson, S.K. and Seber, G.A.F. (1996). Adaptive Sampling. New York: Wiley.

Received March 2000
Revised June 2002