Dipartimento di Statistica
"Giuseppe Parenti"

WORKING PAPER 2007/12

Two-Stage Restricted Adaptive Cluster Sampling

Emilia Rocco

Università degli Studi di Firenze

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

Two-Stage Restricted Adaptive Cluster Sampling

Emilia Rocco
Department of Statistics, University of Firenze, Viale Morgagni 59,
50134 Firenze, Italy - rocco@ds.unifi.it

ABSTRACT Adaptive cluster sampling can be a useful technique for parameter estimation when a population is highly clumped with clumps widely separated. In this design, however, the size of the final sample cannot be predicted prior to sampling, thus leading to design problems. In this paper a new version of adaptive cluster sampling which allows the sampler to know prior to sampling the exact upper limit of the final sample size and consequently the highest total sampling effort, is suggested. It is drawn from a combination of the restricted adaptive cluster sampling of Brown (1994) and a two-stage adaptive cluster sampling. For this new sampling design an unbiased estimator of the total and its sample variance are also suggested. The results of a simulation study, performed in order to provide a first evaluation of the method, are promising.

KEYWORDS: Clumped population, Murthy's estimator, Variable sample size.

1. INTRODUCTION

In adaptive cluster sampling, introduced by Thompson (1990), an initial sample of fixed size is selected and for each unit in the initial sample the neighbouring units are sampled if the variable of interest satisfies a condition, say C_0, specified a priori. If, in turn, any of the neighbouring units satisfy the condition, their neighbourhoods are sampled and so on, building up clusters of units. The condition for extra sampling might be, for example, the presence of rare animal or plant species, detection of "hot spots" in an environmental pollution study, infection with a rare disease in an epidemiological study or observation of a rare characteristic of interest in a household or firm 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. It is evident that the final sample size is unknown and can be quite variable depending on the patchiness of the population. In order to reduce the variability of the final sample size, Brown (1994) and Brown & Manly (1998) proposed an alternative method to select the sample. In their method, known as restricted adaptive cluster sampling, the units of the initial sample are drawn sequentially; for each unit its neighbouring units are sampled and a cluster is formed following the usual adaptive cluster method; the cumulating total sample size is then compared with a predefined limit on the number of units; if the total size is greater than or equal to the limit the process stops, otherwise it continues until such limit is reached or exceeded. Thus, restricted adaptive cluster sampling reduces the variability in the final sample size but does not remove it completely: the resultant final sample size may be much greater than the defined limit depending on the size of the last selected cluster. In order to gain further reduction in the variability of the final sample size and, above all, to be able to determine the exact upper limit of the final sample size, we propose a new method that combines the restricted adaptive cluster sampling with a two-stage adaptive cluster sampling method (Salehi & Seber, 1997) in which an initial two-stage sample is selected and then the related clusters of secondary units truncated at the primary unit boundaries are added to the sample.

The number of measurements of the variable of interest is controlled also with the adaptive cluster double sampling, recently suggested by Felix-Medina & Thompson (2004), in which a first-phase sample is selected using an adaptive cluster sampling design based on an inexpensive auxiliary variable associated with the survey variable and then the network structure of the adaptive cluster sample is used to select an ordinary one-phase or two-phase subsample of units and the values of the survey variable associated with those units are selected. However, this design has the following drawbacks: an auxiliary variable associated with the survey variable is required; the procedure does not control the number of measurements of auxiliary variable and so the sampler is not able to predict prior to sampling the whole sampling effort.

With the design introduced in this paper, it is possible to know the exact upper limit of the number of selected units before sampling.

The next section sets out the notation and describes the new sampling design denoted as two-stage restricted adaptive cluster sampling. In section 3 we derive an unbiased estimator of the total and its variance estimator by adopting a similar approach to Salehi & Seber (2002). In section 4, using a simulation, we give a first evaluation of the behaviour of the proposed method in terms of sample size variability. The unbiased estimator for the sample variance of the total is derived in the appendix.

2. Two-Stage Restricted Adaptive Cluster Sampling Design

Suppose that we have a population of N_T units partitioned into M primary units $(A_1,\dots,A_M)$ of size N_i $\left(i=1,\dots,M;\ \sum_{i=1}^{M} N_i = N_T\right)$. We assume N_i to be constant $(N_i = N\ \forall\ i)$. We first select a simple random sample of m $(m \geq 2)$ primary units without replacement and draw an initial simple random sample of n $(n \geq 2)$ secondary units from each of them. Then we adaptively add to each secondary unit of the sample its neighbourhoods and build up clusters. The clusters are truncated at the primary units boundaries. If the cumulating total sample size is below a predefined limit υ, another primary unit is sampled, n more secondary units are drawn from it and the corresponding clusters are picked out. This last step of the procedure is repeated on remaining primary units as long as, the cumulating sample size is less than υ. Thus, the final sample size υ_F will be equal to or greater than the defined limit υ, but the difference between υ_F and υ will be at most the known size of the last selected primary unit minus 1 and consequently the maximum sample size will be $\upsilon - 1 + N$. If υ is less than $m \times N$ and in the population under study there are clusters so large that they can cover a whole primary unit, the final sample size υ_F could be $m \times N$. Therefore, the upper limit of υ_F is:

$$UP(\upsilon_F) = \max(m \times N, \upsilon + N - 1) \tag{1}$$

Thus, with the two-stage restricted adaptive cluster sampling we can establish prior to sampling the upper limit of the final sample size and consequently choose the parameters of the survey, like the size of the primary units, the number of them initially selected and the limit on the number of observable units, according to the available budget.

Usually the researchers design the primary units of the same size, however two-stage restricted adaptive cluster sampling design may be applied also to the case in which the N_i are not constant. In this case the more general expression for the upper limit of υ_F is:

$$UP(\upsilon_F) = \max\left(\sum_{U_m^*} N_i, \upsilon + \max(N_i) - 1\right)$$

where U_m^* is the set of the *m* larger size primary units.

3. An Unbiased Total Estimator

Let (i, j) denote unit j in primary unit i with an associated measurement y_{ij}. Let $\tau_i = \sum_{j=1}^{N_i} y_{ij}$ be the sum of the y values in the primary unit i and $\tau = \sum_{i=1}^{M} \tau_i$ the population total. In order to get an unbiased estimator of τ we adapt to the present situation the Murthy's estimator (Murthy, 1957), originally proposed for one stage samples of fixed size and then extended by Salehi & Seber (2001) to any sequential one stage sampling scheme. For our two stage sequential sample design Murthy's estimator can be written as:

$$\hat{\tau} = \sum_{i=1}^{m_f} \hat{\tau}_i \frac{P(s \mid i)}{P(s)} \tag{2}$$

where s is the observed unordered set of distinct primary units, $P(s)$ is the probability of getting sample s, $P(s \mid i)$ is the conditional probability of getting the sample s, given primary unit i is select first, $\hat{\tau}_i$ is an unbiased estimator of the total of the y variable in primary unit i and m_f is the final number of sampled primary units.

To evaluate (2) we must determine: (a) the number of ordered samples giving rise to s, (b) the number of ordered sample giving rise to $\{I_i = 1, s\}$ where I_i is an indicator variable which takes the value 1 when the primary unit i is selected as the first one and 0 otherwise, (c) an explicit expression for $\hat{\tau}_i$.

(a) Let υ_F be the total number of distinct secondary units, including edge units, observed up to from a sample of m_f primary units. The last selected primary unit must satisfy the condition $\upsilon_F - \upsilon_{F-1} \geq \upsilon - \upsilon_{F-1} > 0$. Suppose that the number of primary units in the sample satisfying this condition is l. These primary units are indexed as $i=1,...,l$ and their set is referred as L, while the remaining primary units in the sample are indexed as $i=l+1,...,m_f$. Whereas the last selected primary unit must belong to L, the remaining m_f-1 can be permuted in $(m_f-1)!$ ways, and thus the number ε of ordered sample giving rise to s is:

$$\varepsilon = l(m_f-1)!$$

(b) If $i \in L$, the last selected primary unit must be chosen among the remaining $l-1$ in L; if $i \notin L$, then it can be any of the l satisfying the condition; in both cases the remaining m_f-2 primary units can be permuted in $(m_f-2)!$ ways, therefore the number of ordered sample giving rise to $\{I_i=1,s\}$ is:

$$\varepsilon_i = \begin{cases} (l-1)(m_f-2)! & \text{if } i \in L \\ l(m_f-2)! & \text{if } i \notin L \end{cases}$$

Since the probability to select the primary unit i in the first draw is $p_i = 1/M$,

$$\frac{P(s\mid i)}{P(s)} = \frac{\varepsilon_i}{\varepsilon p_i} = \begin{cases} \dfrac{M(l-1)}{l(m_f-1)} & \text{if } i \in L \\ \dfrac{M}{(m_f-1)} & \text{if } i \notin L \end{cases}$$

(c) For $\hat{\tau}_i$ we propose the following expression: $\hat{\tau}_i = \sum_{k=1}^{K_i} y_{ik}^{*} I_{ik} / \alpha_{ik}$. Here K_i, y_{ik}^{*}, and I_{ik} are, respectively, the number of networks in the primary unit i, the sum of the y values associated with network k in primary unit i, and an indicator variable which takes the value 1 when network k is selected in primary unit i and 0 otherwise. α_{ik} is the probability that the set of n units selected before starting the adaptive procedure in primary unit i intersects network k, and its expression is:

$$\alpha_{ik} = 1 - \binom{N-x_{ik}}{n} \Big/ \binom{N}{n}$$

where x_{ik} is the number of units in the network k in primary unit i. Actually $\hat{\tau}_i$ is nothing else than the modified Horvitz-Thompson total estimator proposed by Thompson (1990) and evaluated in primary unit i.

The expression of our estimator is then :

$$\hat{\tau} = \left(\sum_{i=1}^{l} \frac{(l-1)M}{l(m_f-1)} \hat{\tau}_i + \sum_{i=l+1}^{m_f} \frac{M}{(m_f-1)} \hat{\tau}_i \right) \tag{3}$$

In the appendix we show that an unbiased variance estimator of $\hat{\tau}$ is given by

$$
\begin{aligned}
v\mathrm{ar}(\hat{\tau}) = {} & M\,\frac{(M-1)l(l-2)(m_f-1) - M(l-1)^2(m_f-2)}{l^2(m_f-1)^2(m_f-2)}\sum_{i=1}^{l-1}\sum_{j=i+1}^{l}(\hat{\tau}_i - \hat{\tau}_j)^2 + \\
& + M(l-1)\frac{(M-1)(m_f-1) - M(m_f-2)}{l(m_f-1)^2(m_f-2)}\sum_{i=1}^{l}\sum_{j=1}^{m_f-l}(\hat{\tau}_i - \hat{\tau}_j)^2 + \\
& + M\,\frac{(M-1)(m_f-1) - M(m_f-2)}{(m_f-1)^2(m_f-2)}\sum_{i=1}^{m_f-l-1}\sum_{j=i+1}^{m_f-l}(\hat{\tau}_i - \hat{\tau}_j)^2 + \\
& + \sum_{i=1}^{l}\frac{(l-1)M}{l(m_f-1)}v_{2,i} + \sum_{i=l+1}^{m_f}\frac{M}{(m_f-1)_i}v_{2,i}
\end{aligned}
\tag{4}
$$

with $v_{2,i} = \sum_{k=1}^{K_i}\sum_{k'=1}^{K_i} y^*_{ik} I_{ik} y^*_{ik'} I_{ik'}(\alpha_{ikk'} - \alpha_{ik}\alpha_{ik'})/(\alpha_{ikk'}\alpha_{ik}\alpha_{ik'})$.

It is important to note that for evaluating this variance estimator we need to select at least two primary units and at least two secondary units in each of them.

Observing that, when $m_f = m$, the two-stage restricted adaptive cluster sampling corresponds to the two-stage adaptive cluster sampling proposed by Salehi & Seber (1997), $\hat{\tau}$ can be also written as

$$\hat{\tau} = I_{(m_f=m)}\hat{\tau}_1 + I_{(m_f>m)}\hat{\tau}_2$$

where $\hat{\tau}_1 = N_T\hat{\mu}_1 = M\sum_{i=1}^{m}\hat{\tau}_i / m$ is N_T times the mean estimator proposed by Salehi & Seber (1997, pp.963) and the expression of $\hat{\tau}_2$ coincides with that of $\hat{\tau}$ in (3).

Likewise its variance estimator can be rewritten as

$$v\mathrm{ar}(\hat{\tau}) = I_{(m_f=m)}v\mathrm{ar}(\hat{\tau}_1) + I_{(m_f>m)}v\mathrm{ar}(\hat{\tau}_2)$$

where $v\mathrm{ar}(\hat{\tau}_1)$ is $N_T{}^2$ times the variance estimator proposed by Salehi & Seber (1997, pp.963) and $v\mathrm{ar}(\hat{\tau}_2)$ is evaluated through (4).

4. A SIMULATION STUDY AND SOME FINAL CONSIDERATIONS

A simulation has been performed in order to evaluate the properties of two-stage restricted adaptive cluster sampling and, above all, to evaluate its behaviour in terms of sample size variability. To highlight the potential of the two-stage restricted adaptive cluster sampling in reducing the variance of the sample size without producing negative effects on the efficiency of the total estimator, we compare the two-stage restricted adaptive cluster sampling strategy proposed in this paper with the restricted adaptive cluster sampling design first proposed by Brown (1994) and Brown & Manly (1998), and the total unbiased estimator proposed for this sample design by Salehi &

Seber (2002). Brown (1994) and Brown & Manly (1998) used the standard but biased estimators that, according to the results of Salehi & Seber (2002) have also a worse performance than their unbiased estimator in terms of mean square error.

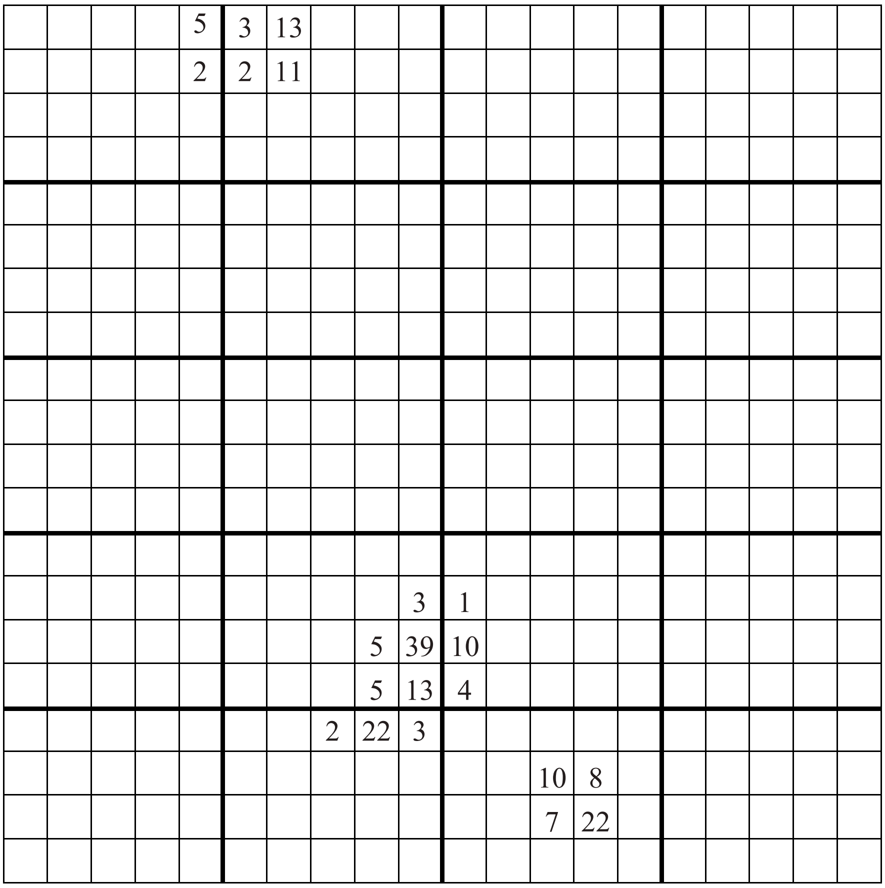

Figure 1. Population 1 coinciding with point-objects population of Thompson (1990). The number in each cell (unit) denotes the number of objects. Bold lines show the partition into primary units.

To compare the two strategies we use the same population adopted by Salehi & Seber (2002): the point-objects population of Thompson (1990) shown in Figure 1 and hereafter denoted as Population 1. It contains $N_T = 400$ units and the total number of objects in the population is 190. A unit satisfies the condition C_0 if the number of objects inside it is greater than 0. For each stopping values υ=20, 30, 40 we simulated 20.000 two-stage restricted adaptive cluster samples and 20.000 restricted adaptive cluster samples. The two-stage restricted adaptive cluster samples are selected after the partition of the population in 20 primary units each of 20 units (see Figure 1) and adopting $m = 2$ and $n = 2$. For each two-stage restricted adaptive cluster sample we calculate $\hat{\tau}$ and its variance estimator and record the final sample size, υ_F. In the same way for each restricted adaptive cluster sample we calculate the Salehi & Seber's estimator denoted as $\hat{\tau}_{RB}$ and its variance estimator and record the final sample size, υ_F. All sampling procedures use the same sequence of random values. Some results are shown in Table 1: the first five rows report some empirical statistics about the sample size in the two designs, and the last row the empirical variance of the two total estimators. Both total estimators and their corresponding variance estimators are unbiased so we do not report their empirical expected values.

	Sampling strategy					
Statistics	TSRACS design and $\hat{\tau}$ estimator			RACS design and $\hat{\tau}_{RB}$ estimator		
υ	20	30	40	20	30	40
$E(\upsilon_F)$	21.52	31.60	41.63	25.29	34.24	43.79
$\max(\upsilon_F)$	32	42	53	43	53	63
$\mathrm{var}(\upsilon_F)$	6.32	6.39	6.25	43.50	39.24	34.97
% sample with $\upsilon_F > UP(\upsilon_F)$	-	-	-	3.58	3.98	3.01
Variance of total estimator	72005.71	37835.36	24272.15	439953.20	88735.89	50704.10

Table 1. Empirical statistics concerning the final size of two-stage restricted adaptive cluster samples and restricted adaptive cluster samples and simulation variances of point-objects total estimators $\hat{\tau}$ and $\hat{\tau}_{RB}$ evaluated for υ equal to 20, 30 and 40. The study population is Population 1. The other parameters of the two-stage restricted adaptive cluster samples are $m = 2$, $n = 2$ and $M = 20$ primary units of 4×5 units.

From Table 1 we see that, for any stopping value, the final sample size in two-stage restricted adaptive cluster sampling is unequivocally better controlled than in restricted adaptive cluster sampling, with its expected value closer to υ and its variance by a long way lower. In restricted adaptive cluster sampling we observe also a percentage between 3.01 and 3.98 of samples with a final size over the upper limit achievable with the two-stage restricted adaptive cluster sampling design that is 40, 49 and 59 for υ respectively equal to 20, 30 and 40. These samples might identify situations in which the experimenter may not be able to complete the survey because factors such as budget, time, etc have run out.

The reduction in the variability of the final sample size also has a positive effect on the variability of the total estimator. From last row of Table 1 we can observe a relative gain of about 84%, 57% and 52% respectively for υ equal to 20, 30 and 40. Finally we note that the gain in terms of both control of the sample size and variability of the total estimator, even if relevant for any υ, decreases as υ increases.

Some more considerations on the role of the population partition and the patchiness of the population are important. If the population is partitioned in a few big primary units and the clusters does not intersect them but are wholly in one of them, it is obvious that the two-stage restricted adaptive cluster sampling control of the sample size with respect to the one of restricted adaptive cluster sampling decreases. At the worst with only one primary unit the two sample design coincide and the upper limit of the sample size is equal to the population size. On the other hand an excessive reduction of the size of the primary units may not be opportune as the other extreme case is that in which the primary units coincide with the secondary units. In this case the final size of the samples selected through the two-stage restricted adaptive cluster sampling design is always equal to υ, but the two-stage restricted adaptive cluster sampling design corresponds to the inverse sampling design which cannot take into account the aggregation of the units satisfying condition C_0 and therefore works worse for the estimation of a total or mean than the adaptive cluster sampling designs for clustered populations (Rocco, 2003).

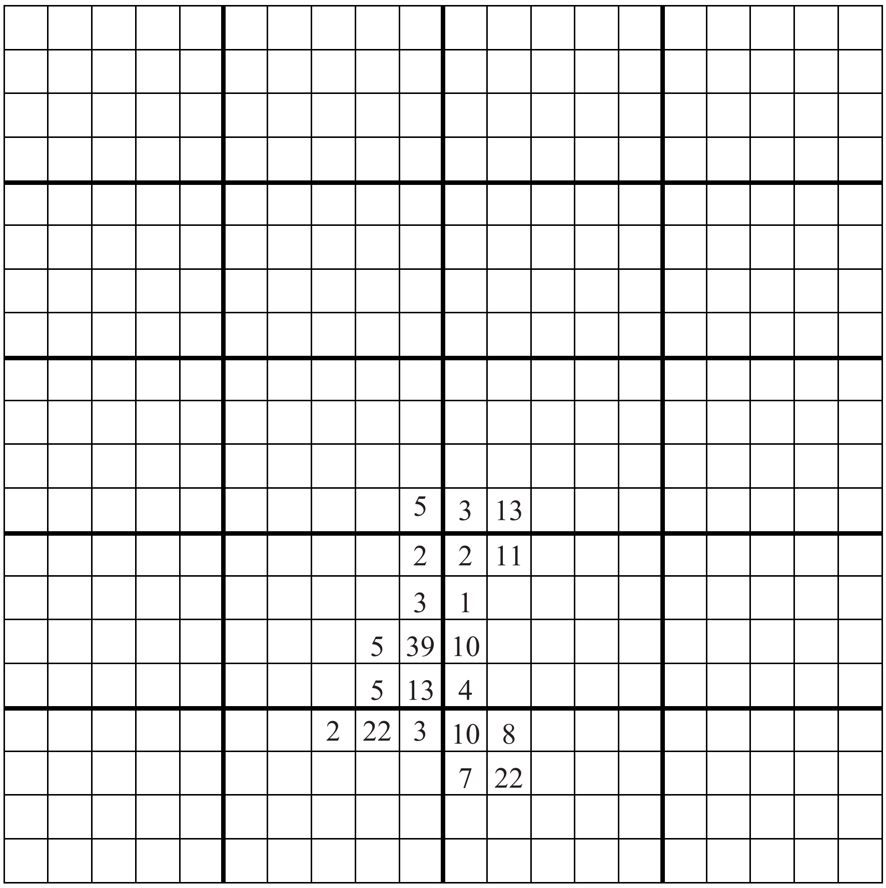

Figure 2. Population 2: Point-objects population obtained from joining all the networks in Population 1. Bold lines show the partition into primary units

It is known that the final sample size for adaptive cluster sampling will depend on the patchiness of the population and for populations with many large clusters we believe that the control achievable with our strategy could be more sensible. At this end, the same Monte Carlo experiment described above is repeated on one more population denoted as Population 2 and obtained joining all the networks in Population

1. Population 2 is represented in Figure 2 and the simulation results concerning it are shown in Table 2. From this table, as expected, we can conclude that with this second population the advantages of the two-stage restricted adaptive cluster sampling with respect to the restricted adaptive cluster sampling are striking. Note that the constant value for $v\mathrm{ar}(\hat{\tau}_{RB})$ is due to the presence in population 2 of only a network for which the sum of the y values is non zero and to the size of the corresponding cluster that is 42. For these reasons for any υ lesser or equal to 42 $\hat{\tau}_{RB}$ is zero if the first selected unit does not belong to the only relevant network and assume the same non zero value in all the other cases. Therefore, as all the sampling procedure use the same sequence of random number, the empirical distribution of $\hat{\tau}_{RB}$ is the same for υ equal to 20, 30 and 40.

	Sampling strategy					
Statistics	TSRACS design and $\hat{\tau}$ estimator			RACS design and $\hat{\tau}_{RB}$ estimator		
υ	20	30	40	20	30	40
$E(\upsilon_F)$	21.44	31.52	41.59	39.69	47.98	52.64
$\max(\upsilon_F)$	30	40	50	61	71	81
$\mathrm{var}(\upsilon_F)$	5.53	5.48	5.52	208.93	121.57	103.25
% sample with $\upsilon_F > UP(\upsilon_F)$	-	-	-	67.13	44.54	24.70
Variance of total estimator	65904.03	35895.42	22656.11	665015.58	665015.58	665015.58

Table 2. Empirical statistics concerning the final size of two-stage restricted adaptive cluster samples (TSRACS) and restricted adaptive cluster samples (RACS) and simulation variances of point-objects total estimators $\hat{\tau}$ and $\hat{\tau}_{RB}$ evaluated for υ equal to 20, 30 and 40. The study population is Population 2. The other parameters of the two-stage restricted adaptive cluster samples are $m = 2$, $n = 2$ and $M = 20$ primary units of 4×5 units.

In all the simulations described until now we initially select $m = 2$ primary units and $n = 2$ secondary units, the number of primary units is then increased if the size of the final sample does not reach the predefined stopping value, υ; while the number of secondary units to be selected not adaptively in each primary unit is fixed a priori and does not change until the selection process stops. To evaluate the effects of different values of n, for each value n = 2, 3, 5, 6, 10 we generate 20.000 two-stage restricted adaptive cluster samples using the study Population 1 and assuming as stopping value υ=30. The results of these simulations are shown in Table 3, from which we see that the variance of the final sample size increases with n; otherwise the effect of n on the variance of the total estimator is not obvious depending on the distribution of the study variable which is obviously unknown and therefore cannot be used in the choice of the best n.

	Number of units selected not adaptively in each primary unit				
Statistics	2	3	5	6	10
$E(\upsilon_F)$	31.60	31.91	32.36	32.83	33.07
$\max(\upsilon_F)$	42	43	44	46	46
$\mathrm{var}(\upsilon_F)$	6.39	6.62	7.67	8.87	9.42
$\mathrm{var}(\hat{\tau})$	37835.36	36322.24	36944.10	38247.23	41803.73

Table 3. Simulation statistics about the sample size and simulation variances of point-objects total estimators $\hat{\tau}$ for the two-stage restricted adaptive cluster sampling applied to Population 1 with different values for the number n of units selected not adaptively in each primary units.

We are aware that while all the simulation results are favourable to the two-stage restricted cluster adaptive sampling strategy, they are not exhaustive. However the main feature of our design is the possibility to know before sampling the exact upper limit of the final sample size; this property holds on for each population and for each choice of the design parameters.

APPENDIX: UNBIASED VARIANCE ESTIMATOR

In order to find an unbiased variance estimator of $\hat{\tau}$ we note that it is nothing else that a particular case of a typical estimator of the total in multi-stage sampling. When the first-stage sample is taken without replacement, a typical estimator of τ is of the form:

$$e = \sum_{i \in s_A} d_i(s_A)\hat{\tau}_i$$

where s_A is the sample of primary unit labels selected in the first stage, $\hat{\tau}_i$ is an unbiased estimator of the total within A_i and $d_i(s_A)$ $(i = 1,...,M)$ are coefficients chosen so that for any $\xi_1, \xi_2,...,\xi_M$, $\sum_{i \in s_A} d_i(s_A)\xi_i$ is an unbiased or nearly unbiased estimator of $\sum_{i=1}^{M} \xi_i$ with respect to the first stage of sampling. As shown in Thompson (1997, p. 35) an unbiased estimator of $\mathrm{var}(e)$ is:

$$var(e) = v_1 + \sum_{i \in s_A} d_i^*(s_A) v_{2,i} \tag{5}$$

where $v_{2,i}$ is an unbiased estimator of the variance of $\hat{\tau}_i$ respect to the second stage of sampling, $d_i^*(s_A)$ satisfies the same unbiasedness condition as $d_i(s_A)$ and v_1 is an unbiased estimator with respect to the first stage design of $\text{var}_1(e \mid s_i \ i = 1,...,M)$ and s_i is the sample of secondary units selected from A_i.

$\hat{\tau}$ is a particular case of e, with

$$d_i(s_A) = \frac{P(s \mid i)}{P(s)} = \begin{cases} \dfrac{M(l-1)}{l(m_f - 1)} & \text{if } i \in L \\ \dfrac{M}{(m_f - 1)} & \text{if } i \notin L \end{cases}$$

Its variance estimator is derived from (5) observing that:

(a) $d_i^*(s_A) = d_i(s_A)$ and so:

$$var(\hat{\tau}) = v_1 + \sum_{i=1}^{l} \frac{(l-1)M}{l(m_f - 1)} v_{2,i} + \sum_{i=l+1}^{m_f} \frac{M}{(m_f - 1)_i} v_{2,i} \tag{6}$$

(b) s_i $(i = 1,...,M)$ is nothing else than a classical adaptive cluster sampling in A_i and $\hat{\tau}_i$ the corresponding total modified Horvitz-Thompson estimator proposed by Thompson (1990), so:

$$v_{2,i} = \sum_{k=1}^{K_i} \sum_{k'=1}^{K_i} y_{ik}^* I_{ik} y_{ik'}^* I_{ik'} (\alpha_{ikk'} - \alpha_{ik}\alpha_{ik'}) / (\alpha_{ikk'}\alpha_{ik}\alpha_{ik'}) \tag{7}$$

where all the symbols are defined above apart from $\alpha_{ikk'}$ which is the probability that the set of n units initially selected, before starting the adaptive procedure, in primary unit i intersects both the k and k' networks, its expression being:

$$\alpha_{ikk'} = \alpha_{ik} + \alpha_{ik'} - \left(1 - \binom{N - x_{ik} - x_{ik'}}{n} \Big/ \binom{N}{n} \right)$$

(c) The expression of v_1 is immediately derived from the expression of the variance estimator of Murthy's estimator:

$$v_1 = \sum_{i=1}^{m_f} \sum_{j<i}^{m_f} \left(\frac{P(s \mid i, j)}{P(s)} - \frac{P(s \mid i) P(s \mid j)}{P(s)^2} \right) \left(\frac{\hat{\tau}_i}{p_i} - \frac{\hat{\tau}_j}{p_i} \right)^2 p_i p_j$$

To evaluate it we must determine the number of ordered samples in which units i and j are the first two selected primary units, in any order. Following the same reasoning used to define ε_i, we have:

$$\varepsilon_{ij} = \begin{cases} (l-2)(m_f-3)! & \text{if } i \in L \text{ and } j \in L \\ (l-1)(m_f-3)! & \text{if } i \in L \text{ or } j \in L \\ l(m_f-3)! & \text{if } i \notin L \text{ and } j \notin L \end{cases}$$

and:

$$\frac{P(s \mid i,j)}{P(s)} = \begin{cases} \dfrac{M(M-1)(l-2)}{l(m_f-1)(m_f-2)} & \text{if } i \in L \text{ and } j \in L \\ \dfrac{M(M-1)(l-1)}{l(m_f-1)(m_f-2)} & \text{if } i \in L \text{ or } j \in L \\ \dfrac{M(M-1)}{(m_f-1)(m_f-2)} & \text{if } i \notin L \text{ and } j \notin L \end{cases}$$

and

$$\begin{aligned} v_1 = M\, & \frac{(M-1)l(l-2)(m_f-1)-M(l-1)^2(m_f-2)}{l^2(m_f-1)^2(m_f-2)} \sum_{i=1}^{l-1}\sum_{j=i+1}^{l}\left(\hat{\tau}_i-\hat{\tau}_j\right)^2 + \\ & + M(l-1)\frac{(M-1)(m_f-1)-M(m_f-2)}{l(m_f-1)^2(m_f-2)}\sum_{i=1}^{l}\sum_{j=1}^{m_f-l}\left(\hat{\tau}_i-\hat{\tau}_j\right)^2 + \\ & + M\,\frac{(M-1)(m_f-1)-M(m_f-2)}{(m_f-1)^2(m_f-2)}\sum_{i=1}^{m_f-l-1}\sum_{j=i+1}^{m_f-l}\left(\hat{\tau}_i-\hat{\tau}_j\right)^2 \end{aligned} \tag{8}$$

Final variance estimator expression is obtained putting together (6), (7) and (8).

REFERENCES

Brown, J.A. (1994) The application of adaptive cluster sampling to ecological studies. In Statistics in Ecology and Environmental Monitoring, D.J. Fletcher & B.F.J. Manly (eds), University of Otago Press, Dunedin, New Zeland. 2, pp. 86-97.

Brown, J.A., Manly, B.J.F. (1998) Restricted adaptive cluster sampling. Environmental and Ecological Statistics, 5, 49-63.

Felix-Medina, M.H., Thompson, S.K. (2004) Adaptive cluster double sampling. Biometrika, 91, 877-891.

Murthy, M.N. (1957) Ordered and unordered estimators in sampling without replacement. Sankhyā,18, 379-390.

Rocco, E. (2003) Constrained inverse adaptive cluster sampling. Journal of Official Statistics, 19, 45-57.

Salehi, M.M., Seber, G.A.F. (1997) Two stage adaptive cluster sampling. Biometrics, 53, 959-970.

Salehi, M.M., Seber, G.A.F. (2001) A new proof of the Murthy's estimator which applies to sequential sampling. Australian and New Zealand Journal of Statistics, 43, 281-286.

Salehi, M.M., Seber, G.A.F. (2002) Unbiased estimators for restricted adaptive cluster sampling. Australian and New Zealand Journal of Statistics, 44, 63-74.

Thompson, M.E. (1997) Theory of Sample Surveys, Chapman & Hall, London.

Thompson, S.K. (1990) Adaptive cluster sampling. Journal of the American Statistical Association, 85, 1050-1059.

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