Dipartimento di Statistica
"Giuseppe Parenti"

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

WORKING PAPER 2002/15

Centre Sampling for Estimating Elusive Population Size

Monica Pratesi, Emilia Rocco

Università degli Studi
di Firenze

Applied Statistics

Centre Sampling for Estimating Elusive Population Size

Monica Pratesi
Dipartimento di Matematica, Statistica, Informatica e Applicazioni
Via Dei Caniana 2- 24127 BERGAMO (ITALY)
pratesi@unibg.it

Emilia Rocco
Dipartimento di Statistica "Giuseppe Parenti"
Viale G.B. Morgagni 59 – 50134 FIRENZE (ITALY)
Rocco@ds.unifi.it

Abstract:
The estimation of the size of an elusive population is a problem frequently addressed in many fields of applications. In the paper a sampling strategy for the estimation of the size, named centre sampling, is proposed, and its properties are evaluated in a design-based approach. The estimator is admissible, consistent and the design is measurable. The expressions of the variance of the population size estimator and of its unbiased sample estimator are also proposed. The strategy is applied both to a simulated population and to the data collected with a CATI survey on the users of some libraries of the University of Florence.

Keywords: Centre sampling, elusive population size, design based inference

1. Introduction

Many elusive populations are detectable when the observation points are the centres of attraction and aggregation of the individuals. The examples are several: immigrants not listed in any administrative file gather in churches, streets or squares; also populations of different nature, as the 1users of a service, from libraries to supermarkets, can be individuated through their centres of aggregation, the service delivery points. The same individual is observable in more than one centre and the detection is limited to the individuals who frequent at least one centre.

The idea of using the centres of aggregation as a frame to estimate the parameters of an elusive population, advanced by Blangiardo (1996), is developed here, to estimate the population size, through the definition of a sampling strategy (p,T) composed by sampling design p and size estimator T, named centre sampling.

We propose to select a sample of centres, to observe all the individuals in the selected centres and to ask to each observed individual to which other centres he/she belongs (section 2). The collected

data are used to estimate the size of the elusive population. Actually, only the size of population of the individuals who frequent at least one centre is estimated. The expressions of the variance of the population size estimator and of its unbiased sample estimator are also proposed (section 3).
The approach can be compared with the Capture Mark Recapture (CMR) methodology due to Lincoln-Petersen and Chapman (Chapman 1951, Giommi and Pratesi 1994). In the CMR procedure the size is estimated using all the possible captures (centres), in our case only a sample of the possible captures is selected. The application of the centre sampling and the comparison are done on a simulated population (section 4). Also a "pseudo real" application of the centre sampling on the data collected with a CATI survey on the users of the faculty's libraries of the University of Florence is carried out (section 5).
This contribution is limited to the estimation of the population size, but the proposed estimator can be easily extended to estimate other parameters. Other contributions based on the idea of using the centres of aggregation as a frame to estimate the parameters of an elusive population give estimation of totals and percentages (Migliorati and Mecatti, 2001). Some contributions focus on parametric models for individual probability of detection (Haines and Pollock, 1998; Migliorati and Terzera, 2001), our procedures is completely design based and its properties are derived by the sampling design (section 6).

2. Centre sampling: the inclusion probabilities

The objective of the strategy (p,T) is the estimation of the number N_c of the individuals who frequent at least one centre. Suppose to have a list of M centres of aggregation. It is proposed a one stage sampling design: a simple random sample of m centres is selected and all the individuals in the selected centres are observed. So for each sample s, the sampling design for the centres is:

$$p(s)=\binom{M}{m}^{-1}$$

In this context, any individual observed in at least one of the selected centres is characterized by the exact number g $(1 \le g \le M)$ of centres that he declares to frequent.

Let I_{ig} be the sample membership indicator of an individual who frequents exactly g centres:

$$I_{ig}=\begin{cases}1 \ \textit{if} \ \ i \in s \\ 0 \ \textit{otherwise}\end{cases}$$

under the previous sampling design, the inclusion probability of the first order is:

$$\pi_{ig} = \Pr(I_{ig} = 1) = E_p(I_{ig}) = \sum_{s \ni i} p(s) = \sum_{h} \frac{\binom{m}{h}\binom{M-m}{g-h}}{\binom{M}{g}} = \pi_g$$

with $h = [\max(1, m+g-M), ..., \min(g, m)]$ and $\pi_g = 1$, $\forall g > M - m$

The first order inclusion probabilities depend on the total number of centres in the population M, the number of selected centres m and the number of centres g that an individual frequents; they are constant for the individuals who frequent the same number of centres.

The expression of the second order inclusion probabilities is more complex. For each couple of individuals (i and j) the inclusion probability depends on the number of centres to which each one belongs and also on the number of centres to which both individuals belong (common centres).

Let g and g' denote the number of centres frequented by each of the two individuals and let c denote the number of common centres; the expression of the second order inclusion probability is:

$$\pi_{ig,jg'} = \pi_{jg',ig} = \sum_{s \ni i \& j} p(s) = \pi^{c}_{gg'} = \pi^{c}_{g'g}$$

and considering, from now on, only $g \le g'$, for the symmetry of the matrix of the probabilities of inclusion,

$$\pi^{c}_{gg'} = \begin{cases} \pi_g \quad if \;\; c = g \;\; i.e.\; if \; the \; number \; of \; common \; centres \; is \max \\ \dfrac{\binom{c}{l}\binom{g-c}{h}\binom{g'-c}{k}\binom{M-g-g'+c}{m-l-h-k}}{\binom{M}{m}} \quad otherwise \end{cases}$$

with the following possible values for l, h and k:

$l = (0, ..., \min(c, m))$;

if $c = 0$ or $l = 0$, $h = (1, ..., \min(g-c, m-1))$ and $k = (1, ..., \min(g-c, m-h))$;

if $l \neq 0$ $h = (0, ..., \min(g-c, m-l))$ and $k = (0, ..., \min(g-c, m-l-h))$.

We have $\pi^{c}_{gg'} = \pi_g \ \ \forall g' > M - m$ and also $\pi^{c}_{gg'} = 1 \ \ \forall g > M - m$.

3. The estimator of the elusive population size and its variance

The N_c individuals, which frequent at least one of the M centres, can be partitioned into G clusters $(G = M)$: each cluster contains all the individuals who frequent the same number of centres g. In other words $N_c = \sum_{g=1}^{G} N_g$, where N_g $(g = 1,...,G)$ denotes the number of individuals who frequent exactly g centres.

Obviously the N_g $(g = 1,...,G)$ are not known. Asking to each observed individual which other centres he frequents we know the corresponding partition of the sampled individuals: the number n_g $(n_g \le N_g)$ of individuals in the sample, which frequent exactly g centres. A possible estimator of the size of the population N_c is then:

$$T = \sum_{g=1}^{G} \frac{n_g}{\pi_g} = \sum_{g=1}^{G} \sum_{i=1}^{N_g} \frac{I_{ig}}{\pi_g}.$$

The T estimator (Horvitz-Thompson type) is of course unbiased for N_c.

The variance of T is:

$$Var[T] = \sum_{g=1}^{G} \frac{\mathrm{var}(n_g)}{\pi_g^2} + 2 \sum_{\substack{g=1 \\ g<g'}}^{G} \sum_{g'=1}^{G} \frac{\mathrm{cov}(n_g, n_{g'})}{\pi_g \pi_{g'}} = \sum_{g=1}^{G} \sum_{g'=1}^{G} \frac{1}{\pi_g \pi_{g'}} \sum_{i=1}^{N_g} \sum_{j=1}^{N_{g'}} \mathrm{cov}(I_{ig}, I_{jg'})$$

where, for every couple of individuals i and j, which frequent respectively g and g' centres and have c common centres, $\mathrm{cov}(I_{ig}, I_{jg'}) = \pi^{c}_{gg'} - \pi_g \pi_{g'}$.

An unbiased estimator of the variance is:

$$\hat{\mathrm{var}}\,[T] = \sum_{g=1}^{G} \sum_{g'=1}^{G} \frac{1}{\pi_g \pi_{g'}} \sum_{s} \sum_{s} \frac{\mathrm{cov}(I_{ig}, I_{jg'})}{\pi^{c}_{gg'}}.$$

The covariance terms are the same for all couples of individuals which have the same values of the triplet (g, g', c); let $N^{c}_{gg'}$ denote the number of these individuals in the population, the variance expression becomes:

$$Var[T]=\sum_{\substack{g=1\\g=g'}}^{G}\sum_{g'=1}^{G}\sum_{c}\frac{N_{gg'}^{c}\left(\pi_{gg'}^{c}-\pi_{g}\pi_{g'}\right)}{\pi_{g}\pi_{g'}}+2\sum_{\substack{g=1\\g'<g}}^{G}\sum_{g'=1}^{G}\sum_{c}\frac{N_{gg'}^{c}\left(\pi_{gg'}^{c}-\pi_{g}\pi_{g'}\right)}{\pi_{g}\pi_{g'}}$$

The number $N_{gg'}^{c}$ for the triplets (g , g' , c) is derived in the Appendix 2.

In the same way, an unbiased estimator of the variance is:

$$\hat{\text{var}}[T]=\sum_{\substack{g=1\\g=g'}}^{G}\sum_{g'=1}^{G}\sum_{c}\frac{n_{gg'}^{c}\left(\pi_{gg'}^{c}-\pi_{g}\pi_{g'}\right)}{\pi_{g}\pi_{g'}\pi_{gg'}^{c}}+2\sum_{\substack{g=1\\g'<g}}^{G}\sum_{g'=1}^{G}\sum_{c}\frac{n_{gg'}^{c}\left(\pi_{gg'}^{c}-\pi_{g}\pi_{g'}\right)}{\pi_{g}\pi_{g'}\pi_{gg'}^{c}}$$

where $n_{gg'}^{c}$ denote the number of individuals in the sample characterised by the same values of the triplet (g , g' , c).

We note that $\text{cov}(I_{ig},I_{jg'})=\pi_{gg'}^{c}-\pi_{g}\pi_{g'}=0$, $\forall g'>M-m$, and so the corresponding addenda in the expressions of the variance and of the variance's estimator vanish.

4. A numerical example

Consider a population of 3100 individuals who frequent at least one of 4 centres A , B , C and D and are distributed in the centres as showed in Table 1.

Just to understand the notation, we mean that there are 200 people who frequent only the centre A (first row of the table) and 50 people who frequent all the 4 centres (last row of the table). The $N_c=3100$ individuals who belong to at least one of the 4 centres can be partitioned in 4 clusters, made of all the individuals which frequent exactly from one to four centres. The size of these clusters are: $N_1=1200$, $N_2=1400$, $N_3=450$ and $N_4=50$.

If we select a sample of two centres, we obtain the following values for the first and second order inclusion probabilities:

$\pi_1=\frac{1}{2}$, $\pi_2=\frac{5}{6}$, $\pi_3=1$, $\pi_4=1$,

$$\pi_{11}=\begin{cases}\pi_1=\frac{1}{2}\ if\ \ c=1\\ \frac{1}{6}\ if\ \ c=0\end{cases},\quad \pi_{12}=\pi_{21}=\begin{cases}\pi_1=\frac{1}{2}\ if\ \ c=1\\ \frac{1}{3}\ if\ \ c=0\end{cases},\quad \pi_{22}=\begin{cases}\pi_2=\frac{5}{6}\ if\ \ c=2\\ \frac{2}{3}\ if\ \ c=1\\ \frac{2}{3}\ if\ \ c=0\end{cases},$$

Table 1: Population distributed by centres of aggregation

CENTRES				FREQUENCY
A	*B*	*C*	*D*	
1	0	0	0	200
0	1	0	0	300
0	0	1	0	400
0	0	0	1	300
0	1	1	0	200
0	0	1	1	300
1	0	1	0	300
0	1	0	1	400
1	1	0	0	100
1	0	0	1	100
1	1	1	0	200
1	0	1	1	50
1	1	0	1	100
0	1	1	1	100
1	1	1	1	50

The other second order inclusion probabilities are not relevant for the calculation of the variance and of the variance estimator because the corresponding addenda in the variance and variance estimator expressions vanish.

Applying the estimator T, defined above, to each of the six possible sample of size 2, we obtain the following results:

Table 2: Sampling distribution of the size estimator T

Sample		n_1	n_2	n_3	n_4	$T = \frac{n_1}{\pi_1} + \frac{n_2}{\pi_2} + \frac{n_3}{\pi_3} + \frac{n_4}{\pi_4}$
A	*B*	500	1100	450	50	2820
A	*C*	600	1000	450	50	2900
A	*D*	500	1200	450	50	2940
B	*C*	700	1300	450	50	3460
B	*D*	600	1100	450	50	3020
C	*D*	700	1300	450	50	3460

$$E_p[T]=\binom{4}{2}^{-1}(2820+2900+2940+3460+3020+3460)=3100=N_c\,.$$

The true variance of T is:

$$Var_p[T]=\sum_{g=1}^{G}\frac{\mathrm{var}(n_g)}{\pi_g^2}+\sum_{\substack{g=1\\g\neq g'}}^{G}\sum_{g'=1}^{G}\frac{\mathrm{cov}(n_g,n_{g'})}{\pi_g\pi_{g'}}=\frac{\mathrm{var}(n_1)}{\pi_1^2}+\frac{\mathrm{var}(n_2)}{\pi_2^2}+2\frac{\mathrm{cov}(n_1,n_2)}{\pi_1\pi_2}=$$

$$=\frac{6666{,}667}{\left(\frac{1}{2}\right)^2}+\frac{12222{,}22}{\left(\frac{5}{6}\right)^2}+2\frac{5000}{\frac{1}{2}\times\frac{5}{6}}=68266{,}67$$

We obtain the same result also applying:

$$Var_p[T]=\sum_{\substack{g=1\\g=g'}}^{G}\sum_{g'=1}^{G}\sum_{c}\frac{N_{gg'}^{c}\left(\pi_{gg'}^{c}-\pi_g\pi_{g'}\right)}{\pi_g\pi_{g'}}+2\sum_{\substack{g=1\\g'<g}}^{G}\sum_{g'=1}^{G}\sum_{c}\frac{N_{gg'}^{c}\left(\pi_{gg'}^{c}-\pi_g\pi_{g'}\right)}{\pi_g\pi_{g'}}=68266{,}67\ \text{(see Appendix 2).}$$

The variance is unbiasedly estimated from each of the six possible samples (see Table 3):

Table 3: Estimated variance of the size estimator T

samples		$\hat{\mathrm{var}}[T]=\sum_{g}\sum_{g'}\sum_{i\in S}\sum_{j\in S}\frac{\pi_{gg'}-\pi_g\pi_{g'}}{\pi_{gg'}\pi_g\pi_{g'}}$
A	*B*	400
A	*C*	212000
A	*D*	1600
B	*C*	3600
B	*D*	128400
C	*D*	63600

$$E_p\left[\hat{\mathrm{var}}_p[T]\right]=\binom{4}{2}^{-1}(400+212000+1600+3600+128400+63600)=68266{,}67=Var_p[T].$$

5. A "pseudo-real" application

The application of centre sampling in order to estimate the population size requires that the elements of the population gather in centres of aggregation. That is the case of the population of users of a service, for whom the centres of aggregation are the service delivery points. In order to test the proposed strategy on survey data, we apply the centre sampling to estimate the number of students who are users of at least one of the 13 libraries active today in the University of Florence and for which data are available.

Actually, it is a "pseudo-real" application. A CATI survey on customer satisfaction of the libraries' users collected data on a stratified sample selected from the population of students registered as users of the 13 libraries mentioned above. The sample size was 1063 students. The interviewers asked each observed user to indicate how many libraries and which libraries he/she usually visits.

The CATI survey data has the same structure of the population described in Table 1: the centres are the 13 libraries and the users are clustered in frequency profiles.

The collected data can be used to test the centre sampling strategy with a Monte Carlo experiment, if we allow the following restrictions:

1. the available sample is the target population, so the elusive population size is 1063;
2. the population is composed of users of at least one of the 13 libraries of the Florence University mentioned above. These libraries are all the possible centres of aggregation.

The Monte Carlo experiment has been carried out with this procedure:

- a simple random sample of 4 centres (libraries) is selected;
- the frequency profile of the users of the sampled libraries is obtained from CATI survey data. In other words, for each user we know how many and which other libraries he/she frequents;
- for each sample the estimator T and is calculated.

The whole procedure is repeated 10000 times: the estimates of the expected value of T, of its median $ME(T)$, of its Mean Squared Error $\sqrt{MSE(T)}$ and of its variance $\sqrt{VAR(T)}$ are obtained on the replications. The results are shown in Table 4.

The exact expression for the variance of T has not been applied, because of the complexities of the calculation when the centres in the population are many and the sampling fraction of the centres is not high (see section 3). Not having the exact variance of T, an indication of its variability is given by the empirical coefficient of variation. The estimates do not concentrate by the true parameter: the coefficient of variation is about 25% and this suggests large estimation intervals for N. Anyway the variability of the proposed methodology should be compared with that of analogous procedure for elusive populations, as it is done in the next section for the data of the numerical examples (see

section 4). The results of that comparison indicate that the centre sample strategy is competing with the CMR methodology.

Table 4: Some results of the Monte Carlo experiment

$E(T)/N$	1.0017
$ME(T)/N$	0.9984
$\sqrt{MSE(T)}/N$	0.2592
$\sqrt{VAR(T)}/N$	0.2592
average sample size of users detected by the centre sampling (percentage on the total)	40.98%

The aim of the pseudo real application is simply to test the factual applicability of the proposed strategy and the obtained results (Table 4) seems to confirm this possibility: the simple random sample of 4 centres allowed to detect, in average, the 41% of the students users of the libraries and this was enough to obtain an unbiased estimate for N (both $E(T)/N$ and $ME(T)/N$ are close to 1).

6. Some properties of the sampling strategy

The centre sampling strategy is characterised by a measurable design and an admissible unbiased estimator of the population size. The estimator is also consistent for N_c. These properties are described in the following settings:

- The proposed centre sampling design, under the restricted condition that at least two centres are selected, is measurable (Särndal et al.,1992). In other words, $\pi_g > 0 \ \forall g$ and $\pi^c_{gg'} > 0 \ \forall \ (g, g', c)$. So the design allows the calculation of valid variance estimates and valid intervals estimates based on the observed survey data.
- As showed in the section 2, under the centre sampling design, the proposed estimator is unbiased for N_c and, as Horvitz-Thompson estimator, is admissible in the class of all the

unbiased estimator of N_c. Under the proposed design there is no better estimator (with smaller MSE) in the class of the unbiased estimators of N_c (Cassel et al.,1977).

- For elusive populations of increasing size and centre samples of increasing size in terms of individuals, the variance of the estimator tends to vanish. This is a sufficient condition for the consistency of our estimator. The condition holds: in fact, given an elusive population, the increase in the population size and the increase of the size of the sample of individuals can be achieved only increasing the number of selected centres m. The number m can increase till M, that, for elusive populations of increasing size, can remain constant or increase as well. If m become greater, the most of the terms $\pi^c_{gg'} - \pi_g \pi_{g'}$ in the expression of the variance vanish. It happens because the condition $g' > M - m$ become more and more frequent when m tends to M. Some authors call consistent the estimator that equals the parameter of interest when the whole population is observed. ($T = N_c$) (Cochran, 1977). The T estimator is consistent also according to this definition: when all the centres are observed ($m = M$) the partition n_g ($g = 1,...,G$) is equal to the corresponding partition of the individuals in the population and all the $\pi_g = 1$. As a result the value of the T estimator is equal to N_c.

The comparison of the efficiency of the proposed strategy (p,T) with that of other traditional strategies for the estimation of the elusive population size is not a simple task.
Particularly, referring to cluster sampling we remark that cluster sampling strategy is based on a totally different approach to the problem. This makes it difficult to compare directly the properties of the two strategies. However, the following considerations are in favour of centre sampling:

- In the elusive population case, the clusters are the areas where the individuals are likely to be present, our centres are instead points of sure attraction for the elements of the population.
- It is easier to build a list of centres than a list of areas: list of areas require costly mapping of the population.
- When nothing is known on the spatial distribution of the elusive population, the cluster sampling can be inefficient because of the effect of the positive intra-cluster correlation on the variance of the estimates.

Among the CMR solutions to the estimation of the size of an elusive population, the model that is closest to the assumptions of the centre sampling strategy is the Lincoln-Petersen model (Giommi and Pratesi, 1994). We remind that, under this model, the CMR estimators are based on the data collected by two independent captures, that the population is assumed to be closed, and that the probability of capture is homogeneous among the individuals. Under these assumptions that are common also to the centre sampling strategy, the estimator usually used is the Chapman one :

$$\hat{N}_C = \frac{(M+1)\cdot(n+1)}{m+1} - 1$$

where M is the size of the first capture (first selected centre), n is the size of the second capture (second selected centre), and m are the individual captured in both the occasions (individuals present in both the selected centres).

The Chapman estimator is a modified version of the Lincoln Petersen estimator $\left(\hat{N}_{LP} = (M \cdot n)/m\right)$; it has less bias in small sample and it is defined also for captures with $m = 0$ (Seber, 1982).

The variance of $\hat{N}_C$ is estimated by :

$$v(\hat{N}_C) = \frac{(M+1)\cdot(n+1)\cdot(M-m)\cdot(n-m)}{(m+1)^2\cdot(m+2)} - 1.$$

The comparison with the CMR model is done on the data simulated in section 4. Table 5 shows the values M, n, m for each couple of captures of Table 2. The centres A, B, C, D are considered as captures from the population (see Table 1).

Table 5: The captures sizes

samples	M	n	m
AB	1100	1250	450
AC	1100	1600	600
AD	1100	1400	300
BC	1250	1600	550
BD	1250	1400	650
CD	1600	1400	500

Table 6 shows the sampling distribution of the CMR estimator. In table 6 and 7 the results are compared with those obtained with centre sampling (see Tables 2 and 3). E_T and V_T are respectively the expected values and the variances of the sampling distributions of the compared estimators, cv and B are respectively the coefficient of variation and the relative bias (bias divided by the square root of the Mean Squared Error) of each estimator. The true population size is 3100.

Table 6: Sampling distributions of T and $\hat{N}_C$

samples	$\hat{N}_C$	$v(\hat{N}_C)$	T	$v(T)$
AB	3053,993	2808,508	2820	400
AC	2932,947	1521,55	2900	212000
AD	5124,588	10658,51	2940	1600
BC	3634,938	3023,11	3460	3600
BD	2692,244	1326,298	3020	128400
CD	4477,048	6302,097	3460	63600
E_T	3652,626	4273,346	3100	68266,67
V_T	774214,8		68266,67	

Table 7: Comparison of the size estimator T with $\hat{N}_C$

samples	$cv(\hat{N}_C)$	$\hat{N}_C / 3100$	$B_{\hat{N}_C}$	$cv(T)$	$T/3100$	B_T
AB	0,017353	0,985159	-0,65556	0,007092	0,909677	-0,99746
AC	0,0133	0,946112	-0,97381	0,158771	0,935484	-0,39841
AD	0,020146	1,653093	0,998702	0,013605	0,948387	-0,97014
BC	0,015126	1,172561	0,994759	0,017341	1,116129	0,986394
BD	0,013527	0,868466	-0,99604	0,118652	0,974194	-0,21789
CD	0,017732	1,444209	0,998342	0,072887	1,116129	0,819028

The CMR estimator is biased for the population size as well as the variance estimator $v(\hat{N}_C)$ (see Table 6). The sampling variability of T is less in average than the variability of the CMR estimator (see V_T in Table 6) and is unbiasedly estimated by $v(T)$. In average, the absolute value of the

empirical relative bias is bigger in the CMR model . This is an expected result given the unbiasedness of the T estimator (see Table 6).

7. Concluding remarks

The estimation of the size of an elusive population is a problem frequently addressed in many fields of applications. Sampling methods such as snowball sampling, network or multiplicity sampling, adaptive cluster sampling and graph sampling are examples of methods usually used to estimate parameters of elusive populations (Thompson 1997). Many contributions deal only with model based approaches (Thompson and Frank, 2000), others apply both design-based and model-based approach (Frank and Snijders, 1994).

In the case of centre sampling proposed in this paper, the approach is completely design based, the proposed estimator is non parametric and consistent; the centre design is measurable and the estimator has good empirical properties when we consider efficiency in term of variance. The comparison with the CMR methodology has highlighted elements in favour of centre sampling. Also the remarks on cluster sampling are in favour of the our strategy when the auxiliary information does not allow an improvement of the efficiency of the cluster design.

The centre sampling strategy can be easily extended do estimate other parameters of the elusive population: the T estimator can be extended to estimate the totals of variables collected on the individuals observed in the centres. Analogously, the expressions of the variance and of the variance's estimator can be extended too.

Future research will be devoted both to centre sampling strategies for the estimation of totals and percentages and to sampling designs that select a subset of the individuals detected in the centres and stratify the centres of aggregation on the basis of auxiliary information. The auxiliary information can be also included in the definition of parametric models for individual probabilities of detection, as already prospected in previous contributions (Alho, 1990; Huggins 1989; Pratesi and Rocco, 1999).

The approximate evaluation of the variance of the centre sampling strategy and development of a software for the application of the strategy are other two issues on which our attention will focus.

Appendix 1

This appendix shows how the numbers $N^c_{gg'}$ are calculated. It is important to remember that $N^c_{gg'} = N^c_{g'g}$ and so the following consideration will be made only considering $g' < g$ (under this assumption the max number of common centres is g).

Let k denote one of the possible group of g centres extractible from the M centres, N^k_g the number of individuals of this group which frequent g centres and $N^k_{g'}$ the number of individuals of this group which frequent g' centres:

if $c = g$ $N^c_{gg'} = \sum_{k=1}^{\binom{M}{g}} N^k_g N^k_{g'}$;

if $c = g-1$ $N^c_{gg'} = \sum_{k=1}^{\binom{M}{g}} N^k_g \left(\sum_{h=1}^{\binom{g}{g-1}} N^h_{g'} - \binom{g}{g-1} N^k_{g'} \right)$ where the meanings of h and $N^h_{g'}$ are similar to which of k and $N^k_{g'}$;

if $c = g-2$ $N^c_{gg'} = \sum_{k=1}^{\binom{M}{g}} N^k_g \left(\sum_{l=1}^{\binom{g}{g-2}} N^l_{g'} - \binom{g-1}{g-2} \sum_{h=1}^{\binom{g}{g-1}} N^h_{g'} + \binom{g}{g-1} N^k_{g'} \right)$;

the other $N^c_{gg'}$ for each possible value of c, until c gets to its minimal value can be draw at the same manner; only for $c = 0$ we have another possible expression:

if $c = 0$ $N^c_{gg'} = \sum_{k=1}^{\binom{M}{g}} N^k_g \left(N_g - \bigcup_{h=1}^{\binom{g}{1}} N^h_g \right)$.

The expression for the $n^c_{gg'}$ are the same, it is sufficient to substitute the $N^c_{gg'}$ with the corresponding $n^c_{gg'}$.

Appendix 2

This appendix shows, for the data simulated in section 4, the calculus of the variance of T.

$$Var[T]=\sum_{\substack{g=1\\g=g'}}^{G}\sum_{g'=1}^{G}\sum_{c}\frac{N_{gg'}^{c}\left(\pi_{gg'}^{c}-\pi_{g}\pi_{g'}\right)}{\pi_{g}\pi_{g'}}+2\sum_{\substack{g=1\\g'<g}}^{G}\sum_{g'=1}^{G}\sum_{c}\frac{N_{gg'}^{c}\left(\pi_{gg'}^{c}-\pi_{g}\pi_{g'}\right)}{\pi_{g}\pi_{g'}}=$$

$$=\left[\left(N_1^A\right)^2+\left(N_1^B\right)^2+\left(N_1^C\right)^2+\left(N_1^D\right)^2\right]\left(\pi_{11}^1-\pi_1\pi_1\right)+$$

$$+\left[N_1^A\left(N_1-N_1^A\right)+N_1^B\left(N_1-N_1^B\right)+N_1^C\left(N_1-N_1^C\right)+N_1^D\left(N_1-N_1^D\right)\right]\left(\pi_{11}^0-\pi_1\pi_1\right)+$$

$$+2\left[N_1^AN_2^A+N_1^BN_2^B+N_1^CN_2^C+N_1^DN_2^D\right]\left(\pi_{12}^1-\pi_1\pi_2\right)+$$

$$+2\left[N_1^A\left(N_2-N_2^A\right)+N_1^B\left(N_2-N_2^B\right)+N_1^C\left(N_2-N_2^C\right)+N_1^D\left(N_2-N_2^D\right)\right]\left(\pi_{12}^0-\pi_1\pi_2\right)$$

$$+\left[\left(N_2^{AB}\right)^2+\left(N_2^{AC}\right)^2+\left(N_2^{AD}\right)^2+\left(N_2^{BC}\right)^2+\left(N_2^{BD}\right)^2+\left(N_2^{CD}\right)^2\right]\left(\pi_{22}^2-\pi_2\pi_2\right)+$$

$$+[N_2^{AB}\left(N_2^A+N_2^B-2N_2^{AB}\right)+N_2^{AC}\left(N_2^A+N_2^C-2N_2^{AC}\right)+N_2^{AD}\left(N_2^A+N_2^D-2N_2^{AD}\right)+$$

$$+N_2^{BC}\left(N_2^B+N_2^C-2N_2^{BC}\right)+N_2^{BD}\left(N_2^B+N_2^D-2N_2^{BD}\right)+N_2^{CD}\left(N_2^C+N_2^D-2N_2^{CD}\right)]\times\left(\pi_{22}^1-\pi_2\pi_2\right)+$$

$$+[N_2^{AB}\left(N_2-N_2^A-N_2^B+N_2^{AB}\right)+N_2^{AC}\left(N_2-N_2^A-N_2^C+N_2^{AC}\right)+$$

$$+N_2^{AD}\left(N_2-N_2^A-N_2^D+N_2^{AD}\right)+N_2^{BC}\left(N_2-N_2^B-N_2^C+N_2^{BC}\right)+$$

$$+N_2^{BD}\left(N_2-N_2^B-N_2^D+N_2^{BD}\right)+N_2^{CD}\left(N_2-N_2^C-N_2^D-N_2^{CD}\right)]\left(\pi_{22}^0-\pi_2\pi_2\right)=68266{,}67$$

where: $N_1^A=200$, $N_1^B=300$, $N_1^C=400$, $N_1^D=300$, $N_2^A=500$, $N_2^B=700$, $N_2^C=800$, $N_2^D=800$, $N_2^{AB}=100$, $N_2^{AC}=300$, $N_2^{AD}=100$, $N_2^{BC}=200$, $N_2^{BD}=400$ and $N_2^{CD}=300$ are calculated using the expression in appendix 1.

References

Alho J. M. (1990), Logistic Regession in capture-Recapture Models, *Biometrics*, 46, pp. 623-635.

Blangiardo G. C. (1996), Il campionamento per centri o ambienti di aggregazionenelle indagini sulla presenza straniera, in *Studi in onore di Giampiero Landenna*, Giuffrè, Milano, pp.13-30.

Cassel C. M., Särndal C. E. and Wretman J. H. (1977), *Foundation of Inference in Survey Sampling*, John Wiley & Sons, New York.

Chapman R. M. (1951), Some properties of the Hypergeometric Distribution with Application to Zoological Sample Censuses, University of California Publication in Statistics, 1, pp. 131-159.

Cochran W. G. (1977), *Sampling techniques*, Wiley, New York.

Frank O. and Snijders T. (1994), Estimating the Size of Hidden Populations Using Snowball Sampling, *Journal of Official Statistics*, 10, 1, pp. 53-67.

Giommi A, Pratesi M. (1994), Liste incomplete: metodi cattura e ricattura per la stima della dimensione della popolazione di aziende, *Atti della XXXVII Riunione Scientifica della Società Italiana di Statistica*, San Remo 6-8 aprile 1994, pp.211-218.

Haines D.E., Pollock K. H. (1998), Combining Multiple Frames to Estimate Population Sizes and Totals, *Survey Methodology,* vol.24, n.1 pp. 79-88.

Huggins R. M. (1989), On the Statistical Analysis of capture Experiments, *Biometrika*, 76, pp.133-140.

Mecatti F., Migliorati S. (2001), Confronti fra stimatori per la media nel campionamnto per centri, relazione presentata al Convegno Metodi di Inferenza Statistica per problemi Complessi Bressanone 2-4 novembre 2001.

Migliorati S., Terzera L. (2001), Una proposta di stima della numerosità nel campionamento per centri, relazione presentata al Convegno Metodi di Inferenza Statistica per problemi Complessi Bressanone 2-4 novembre 2001.

Pratesi M., Rocco E. (1999), Un disegno di campionamento per centri per il controllo della copertura del censimento, in *Atti del Convegno della Società Italiana di Statistica "Verso i censimenti del 2000"*, pp. 514-530.

Särndal C. E., Swensson B., Wretman J. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Seber, G. A. F. (1982), *The Estimation of Animal Abundance and Related Parameters (Second Edition)*, E. Arnold, London.

Thompson S. K. (1997), Adaptive Sampling in Behavioural Surveys, In The Validity of Self-Reported Drug use: Improving the Accuracy of survey Estimates, Harrison L. e Hughes A. eds.,pp.269-319. NIDA Research Monograph 167, Rockville, MD:National Institute of Drug Abuse.

Thompson S. K. and Frank O. (2000), Model-Based Estimation With Link-Tracing sampling Designs, *Survey Methodology*, 26, 1, pp. 87-98.