STATISTICA, anno LXIII, n. 3, 2003

CENTER SAMPLING: A STRATEGY FOR ELUSIVE POPULATION SURVEYS

F. Mecatti, S. Migliorati

1. INTRODUCTION

Center sampling (CS) has been developed in Italy (Blangiardo, 1996) in connection with surveys on immigrant population formed by regular and irregular people. The term regular refers to people with a residence permit according to Italian law while the term irregular indicates people illegally residing in the country.

With this kind of population, traditional finite population sampling is not allowed essentially due to the following reasons. The population size N is unknown and exhaustive lists of the target population are not readily available so that labeling is not possible; in addition units usually require to remain anonymous and in general there is a detectability problem.

CS has been successfully employed in a survey conducted by the Statistical Office of European Communities and it has been applied at a local level since the first half of the 1990s by the Regional Research Institute of Lombardia, Italy. However, few systematic theoretical results exist in the literature to support empirical findings (Blangiardo, 2000). One of the aims of the present paper is to formalize a theory for CS.

The essence of CS relies upon the Italian immigrant habit to congregate in particular places for social contacts, health care, religion, leisure or simply for everyday needs. The basic assumption is that every immigrant can be found, with some regularity, at one or more of these places. Therefore, the target population can be assimilated to the users of these sites which are then the natural and perhaps the only way to observe units. We shall refer to these sorts of places as "Centers". For instance a mosque could be a center for Muslim people, the train station is a usual meeting point for homeless and irregular immigrants, but also a partial list from any official source (e.g. a hospital, a police office or a register of birth or marriages and so on) could be considered as a center. Notice that the framework described is also suitable to deal with problems different from the immigrant one, for instance homelessness as it is accounted for in Dennis and Iachan (1992). Although it does not seem possible to give a formal definition of centers, nevertheless they have the following characterization. Though they are usually spread all

over the geographical area of interest, they have been identified purposively or by means of previous surveys, they are a finite number, say L, they usually overlap and they cover the entire target population.

Some appreciable analogies with multiple frame pattern, as originally proposed by Hartley (1974), are evident: centers might be treated essentially as frames, singularly incomplete and together covering the entire target population. Nevertheless, multiple frame survey can be viewed as a particular case of CS: due to generality of center's definition – as showed by the examples proposed above – both center sizes and overlapping among centers have to be assumed unknown. Thus multiple frame survey and CS are equivalent only if a particular definition of center is stated, i.e. when incomplete lists with known sizes are used, while CS is more general than multiple frame survey in all the other cases where centers are not necessarily incomplete lists.

Moreover, the number L of centers is usually taken greater than 2 in order to reach an adequate coverage. Consequently, dual frame estimators such as those proposed by Hartley (1974), Lund (1968), Fuller and Burmeister (1972), Skinner and Rao (1996), Lohr and Rao (2000) are not trivially adaptable in the context of CS.

In this article an estimation methodology capable of handling the CS framework is developed. We focus on the estimation of the mean μ of a quantitative or dichotomous characteristic; this seems the natural choice instead of the total, traditionally considered in the finite population literature, since both the population size N and center sizes are assumed unknown.

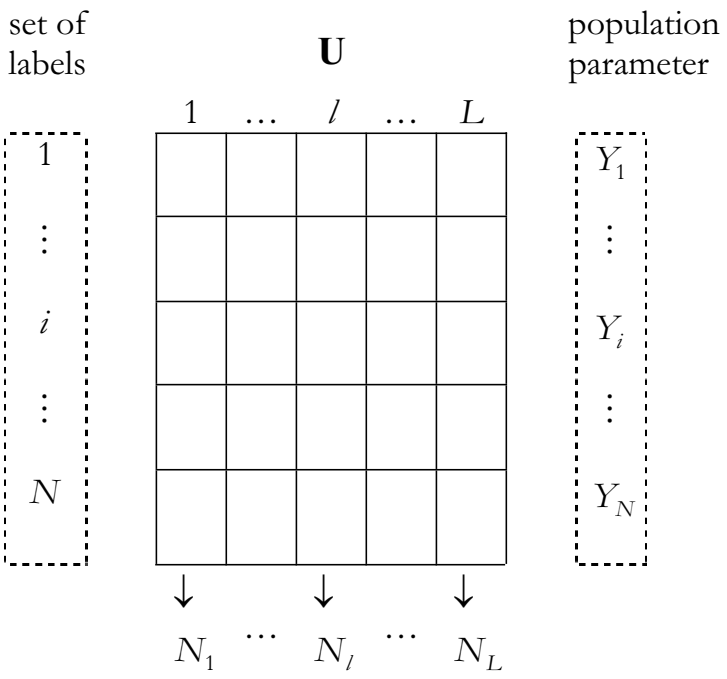
In Section 2 the formal context and a general estimation theory for CS is presented along with an unbiased estimator for μ and its exact variance. General results of Section 2 are particularized in Section 3 to the case of simple random sampling and a suitable variance estimator in a closed and simple form is provided. Some practical remarks and the optimum allocation of the sample size among centers subject to linear cost constraints are discussed in Section 4. In Section 5 single and double CS designs are considered. Finally, two illustrative applications to both artificial and real data are implemented in Section 6. Algebraic computations are gathered in the Appendix.

2. A GENERAL THEORY FOR CENTER SAMPLING

Let y be the characteristic of interest and let $\{Y_1,\cdots,Y_i,\cdots,Y_N\}$ define the population values, i.e. the *parameter* of the population if the units were identifiable (Cassel, Särndal and Wretman, 1977, p. 6).

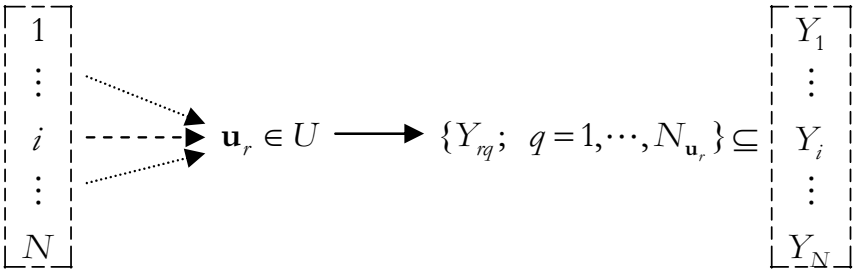
In the CS framework units are not identifiable by a label and, in general, no algebraic mapping between units and centers exists.

Consequently it is proposed to join the traditional pair "set of labels" and "population parameter" to the $N \times L$ matrix $\mathbf{U}$ where each il^{th} cell equals one if unit i attends the l^{th} center and zero otherwise as described in the following scheme ($i = 1,\cdots,N;\ l = 1,\cdots,L$)



The center size N_l $(N_l \geq 1)$ is given by the sum of each column and it is assumed unknown in the general case. Moreover $\sum_{l=1}^{L} N_l \geq N$ due to overlapping. The rows of $\mathbf{U}$ are named *profiles*. Profiles inform about each unit's attendance at each of the L centers and they are a fundamental tool in CS. Each unit has a unique profile which is one of the $2^L - 1$ possible ordered collections of L-tuples of the digits zero and one, whereas the null L-tuple is not a profile. The set U of such L-tuples always exists as the number of centers L is known; vice versa, since units are unidentifiable, the set of labels is unattainable. Besides, profiles are very important practical tools since they can be assumed observable by allowing units to remain anonymous. Particularly, profiles can be observed by simply asking the sampled unit the question «Which other centers – among the L considered and besides the one where the subject has been sampled – are you used to attend?». As a consequence a many-to-one mapping from the set of labels to the set U exists.

Let $N_{\mathbf{u}_r}, (r = 1, \cdots, 2^L - 1)$, denote the (unknown) number of units in the population with profile $\mathbf{u}_r = [u_{r1}, \cdots, u_{rl}, \cdots, u_{rL}]$ where $\mathbf{u}_r \in U$ and let $\{Y_{rq};\ q = 1, \cdots, N_{\mathbf{u}_r}\}$ be the subset of the parameter of the population with respect to units with profile $\mathbf{u}_r$ according to the highlighted mapping. Hence, as r varies from 1 to $2^L - 1$, the sets $\{Y_{rq};\ q = 1, \cdots, N_{\mathbf{u}_r}\}$ form a partition of the parameter of the population as outlined below



Notice that when centers are incomplete lists, i.e. in the multiple frame context, the sets of units with the same profile $\mathbf{u}_r$ are domains (Hartley, 1974).

Notice also that when the l^{th} element of profile $\mathbf{u}_r$ equals 1 (i.e. $u_{rl}=1$ meaning that all of units with associated profile $\mathbf{u}_r$ frequent the l^{th} center) then the entire set $\{Y_{rq};\ q=1,\cdots,N_{\mathbf{u}_r}\}$ belongs to the l^{th} center.

The object of estimation is the population mean

$$\mu = \frac{1}{N}\sum_{r=1}^{2^L-1}\sum_{q=1}^{N_{\mathbf{u}_r}} Y_{rq} .$$

Let $m_r = \sum_{l=1}^{L} u_{rl}$ denote the number of centers attended by units with profile $\mathbf{u}_r$ so that it indicates the *multiplicity* of the r^{th} profile (Birnbaum and Sirken, 1965; Thompson, 2002 , p. 173).

By using profiles and multiplicity, μ can be expressed as

$$\mu = \frac{1}{N}\sum_{l=1}^{L}\sum_{r=1}^{2^L-1}\frac{1}{m_r}\sum_{q=1}^{N_{\mathbf{u}_r}} Y_{rq}\, u_{rl} .$$

We define

$$\tilde{\mu}_l = \frac{1}{N_l}\sum_{r=1}^{2^L-1}\frac{1}{m_r}\sum_{q=1}^{N_{\mathbf{u}_r}} Y_{rq}\, u_{rl} \qquad (1)$$

as the mean of *y* in the l^{th} center *adjusted for multiplicity* and by defining $\alpha_l = N_l/N$ as the weight of the l^{th} center with respect to the population size we have:

$$\mu = \sum_{l=1}^{L}\alpha_l\,\tilde{\mu}_l \qquad (2)$$

Notice that the weighted sum (2) has no counterpart with respect to the center's means $\mu_l = \sum_{i\in l} Y_i/N_l$ yet due to the overlap among centers. Whereas expression (2) is the convenient form for μ we will consider in discussing estimation.

Indeed, dealing with (2) where the overlapping is definitely embodied into the terms $\tilde{\mu}_l$ by using profiles and multiplicities, we can rely upon general results from stratified sampling theory. Hence, if an estimator $\overline{y}_l$ unbiased for $\tilde{\mu}_l$ under a general sampling design is given, then $\overline{y} = \sum_{l=1}^{L} \alpha_l \overline{y}_l$ is an unbiased estimator for μ and, by assuming independence of samples among centers, it has variance $\text{var}(\overline{y}) = \sum_{l=1}^{L} \alpha_l^2 \text{ var}(\overline{y}_l)$.

Furthermore, in real applications as cited in the introduction, it is usually possible to get information about the weight of centers, i.e. although the absolute sizes N and N_l are unknown, the ratio α_l is known, for instance from past data or as reported by experts and witnesses on the field.

However, in constructing $\overline{y}_l$ the analogy with stratified sampling does not help any more: as a matter of fact in the CS framework sampling is from overlapping centers, not from profiles so that some *ad hoc* adaptations are still called for.

Under a general sampling design, let us now refer to the sample from the l^{th} center.

As noticed in the introduction, the population units are not identifiable by a label but only through their profile hence let $\pi_{rq,l}$ be the first order inclusion probability of the q^{th} unit with profile $\mathbf{u}_r$ attending the l^{th} center.

Data are formed through the simultaneous observation of both y values and units' profiles.

Let $f_{\mathbf{u}_r,l}$ be the sample frequency of profile $\mathbf{u}_r$ and let $\{y_{rs}; s = 1,\cdots, f_{\mathbf{u}_r,l}\}$ be the set of sample values observed at units with profile $\mathbf{u}_r$.

As outlined above with respect to the population and by using a similar argument, as r varies from 1 to $2^L - 1$, the sets $\{y_{rs}; s = 1,\cdots, f_{\mathbf{u}_r,l}\}$ form a partition of the sample data from the l^{th} center.

We propose the following unbiased estimator for $\tilde{\mu}_l$

$$\overline{y}_l = \frac{1}{N_l} \sum_{r=1}^{2^L-1} \frac{1}{m_r} \sum_{s=1}^{f_{\mathbf{u}_r,l}} \frac{y_{rs}}{\pi_{rs,l}} \tag{3}$$

Estimator (3) is a linear combination (across profiles) of standard Horvitz-Thompson estimators (within profiles). Thus unbiasedness is readily proved.

Nevertheless, we can not carry this analogy further and rely upon known results about Horvitz-Thompson estimator basically because, as sampling is from centers, data are not still independent with respect to profiles.

By introducing the random variable $I_{rq,l}$ as the sample membership indicator of value Y_{rq} in l^{th} sample (Särndal *et al.*, 1992, p. 36), with $\mathrm{E}(I_{rq,l}) = \pi_{rq,l}$ and $\mathrm{E}(I_{rq,l} \cdot I_{tv,l}) = \pi_{rqtv,l}$ denoting first and second order inclusion probabilities, estimator (3) has exact variance

$$\begin{aligned}\mathrm{var}(\bar{y}_l) &= \mathrm{var}\left(\frac{1}{N_l} \sum_{r=1}^{2^L-1} \frac{1}{m_r} \sum_{q=1}^{N_{\mathbf{u}_r}} \frac{Y_{rq}\, u_{rq}\, I_{rq,l}}{\pi_{rq,l}} \right) \\ &= \frac{1}{N_l^2} \left\{ \sum_{r=1}^{2^L-1} \frac{1}{m_r^2} \left[\sum_{q=1}^{N_{\mathbf{u}_r}} Y_{rq}^2\, u_{rl} \left(\frac{1}{\pi_{rq,l}} - 1 \right) + \sum_{q=1}^{N_{\mathbf{u}_r}} \sum_{\substack{v=1 \\ v \neq q}}^{N_{\mathbf{u}_r}} Y_{rq}\, Y_{rv}\, u_{rl} \left(\frac{\pi_{rqrv,l}}{\pi_{rq,l}\, \pi_{rv,l}} - 1 \right) \right] + \right. \\ &\left. + \sum_{r=1}^{2^L-1} \sum_{\substack{t=1 \\ t \neq r}}^{2^L-1} \frac{1}{m_r\, m_t} \sum_{q=1}^{N_{\mathbf{u}_r}} \sum_{v=1}^{N_{\mathbf{u}_t}} Y_{rq}\, Y_{tv}\, u_{rl}\, u_{tl} \left(\frac{\pi_{rqtv,l}}{\pi_{rq,l}\, \pi_{tv,l}} - 1 \right) \right\}\end{aligned} \tag{4}$$

In expression (4) it is noticeable a first addend, in square brackets, related with the variability within profiles and dealing with the usual variance of the Horvitz-Thompson estimator but also an additional term clearly related with the variability across profiles $r \neq t$.

3. SIMPLE RANDOM SAMPLING WITHIN CENTERS

3.1 *Estimation*

In real applications, like the immigrant surveys mentioned in the introduction, CS has been performed by selecting a simple random sample of fixed size n_l from the l^{th} center where N_l units are present almost surely $(l = 1,...,L)$. This can be accomplished by sampling independently in each center at the moment of "maximum crowding" as suggested by the "nature" of the center. As an example, the shelters where homeless people can find a bed for the night will be visited at bedtime, the worship locations will be visited while rites are being officiated and so on. Obviously this is assured when centers are incomplete lists. Furthermore, we will assume α_l is known for all centers.

Some issues regarding the above assumptions and how to proceed in absence of frame, will be addressed in Section 4.

According to the assumed sample design, the collection $\{I_{rq,l},\ r = 1,\cdots,2^L - 1;\ q = 1,\cdots,N_{\mathbf{u}_r}\}$ is a Multihypergeometric random variable with

$$\mathrm{E}(I_{rq,l}) = \mathrm{E}(I^2_{rq,l}) = \frac{n_l\, u_{rl}}{N_l}$$

$$\mathrm{E}(I_{rq,l} \cdot I_{tv,l}) = \frac{n_l(n_l-1)\, u_{rl}\, u_{tl}}{N_l(N_l-1)} \tag{5}$$

$$\mathrm{var}(I_{rq,l}) = \frac{n_l\, u_{rl}}{N_l}\left(1 - \frac{n_l\, u_{rl}}{N_l}\right)$$

Yet, by remembering (2) and substituting in expression (3), an unbiased estimator for μ is

$$\bar{y} = \sum_{l=1}^{L} \alpha_l\, \bar{y}_l = \sum_{l=1}^{L} \frac{\alpha_l}{n_l} \sum_{r=1}^{2^L-1} \frac{y_{r,l}}{m_r} \tag{6}$$

where $y_{r,l} = \sum_{s=1}^{f_{\mathbf{u}_r,l}} y_{rs}$ is the sample total within the r^{th} profile in the l^{th} center.

It follows from (4) that the variance of the estimator (6) is given by

$$\mathrm{var}(\bar{y}) = \sum_{l=1}^{L} \frac{\alpha_l^2\,(N_l - n_l)}{n_l\, N_l(N_l-1)} \left[\sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \frac{Y^2_{rq}\, u_{rl}}{m_r^2} - \frac{1}{N_l} \left(\sum_{r=1}^{2^L-1} \frac{Y_r\, u_{rl}}{m_r} \right)^2 \right] \tag{7}$$

where $Y_r = \sum_{q=1}^{N_{\mathbf{u}_r}} Y_{rq}$ is the population total within the r^{th} profile (see the Appendix).

In order to obtain an estimator of $\mathrm{var}(\bar{y})$ given by (7), we first note that

$$\sum_{l=1}^{L} \frac{\alpha_l^2}{n_l^2\,(n_l-1)} \left(1 - \frac{n_l}{N_l}\right) \left[n_l \sum_{r=1}^{2^L-1} \sum_{s=1}^{f_{\mathbf{u}_r,l}} \frac{y^2_{rs}}{m_r^2} - \left(\sum_{r=1}^{2^L-1} \frac{y_{r,l}}{m_r} \right)^2 \right] \tag{8}$$

is unbiased, as shown in the Appendix. A conservative variance estimator, not depending on N_l (assumed unknown), is obtained from (8) by replacing $(1 - n_l/N_l)$ by 1:

$$\hat{v}(\bar{y}) = \sum_{l=1}^{L} \frac{\alpha_l^2}{n_l^2\,(n_l-1)} \left[n_l \sum_{r=1}^{2^L-1} \sum_{s=1}^{f_{\mathbf{u}_r,l}} \frac{y^2_{rs}}{m_r^2} - \left(\sum_{r=1}^{2^L-1} \frac{y_{r,l}}{m_r} \right)^2 \right]. \tag{9}$$

3.2 *Comparison with dual frame estimators*

If the center sizes, N_l, are known, we can obtain an estimator of the total, $N\mu$, from (6) as

$$\hat{y} = \sum_{l=1}^{L} N_l \, \overline{y}_l \tag{10}$$

Estimator (10) agrees with the estimator suggested by Hartley (1974) for a dual frame survey, assuming that partial lists with known sizes are used as centers. Setting L=2 we have $2^L - 1 = 3$ profiles: $\mathbf{u}_1 = (1, 0)$ referring to units belonging to center 1 only, $\mathbf{u}_2 = (0, 1)$ referring to units belonging to center 2 only and $\mathbf{u}_3 = (1, 1)$, with multiplicity 2, referring to units attending both center 1 and center 2, namely the overlap domain in the dual frame literature. Hence the estimator (10) reduces to

$$\hat{y} = \frac{N_1}{n_1}\left(y_{1,1} + \frac{1}{2} y_{3,1} \right) + \frac{N_2}{n_2}\left(y_{2,2} + \frac{1}{2} y_{3,2} \right). \tag{11}$$

This estimator coincides with Hartley's dual frame estimator with a simple choice of weights, p=q=1/2.

Estimator (11) has been compared with some of its major competitors (Mecatti, 2002). Theoretical considerations and simulation results indicate that estimator (11) is a feasible alternative with respect to inferential properties. Moreover, from a practical point of view, estimator (11) may be preferable over the usual dual frame estimators due to its simplicity, the availability of an unbiased variance estimator in a simple and closed form, and its immediate generalisation to any number of frames as given by (10).

4. REMARKS

4.1 *Optimum choice of sample sizes* n_l

We first assume the simple cost function $n = \sum_{l=1}^{L} n_l$ where n is the overall sample size specified on the basis of available resources.

In order to pursue the optimum allocation we first need the variance of y, say σ^2, to enter $\mathrm{var}(\overline{y})$ as given by (7). The matter of concern is obviously the variability "within centers" since we sample from every center. Hence we suggest (see the Appendix) the following decomposition

$$\begin{aligned} \sigma^2 = & \left[\frac{1}{N} \sum_{l=1}^{L} \left(\sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \frac{Y_{rq}^2 \, u_{rl}}{m_r^2} - N_l \tilde{\mu}_l^2 \right) \right] + \\ & + \left[\frac{1}{N} \sum_{l=1}^{L} N_l \tilde{\mu}_l^2 - \mu^2 + \frac{1}{N} \sum_{l=1}^{L} \sum_{\substack{k=1 \\ k \neq l}}^{L} \left(\sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \frac{Y_{rq}^2 \, u_{rl} \, u_{rk}}{m_r^2} \right) \right] \end{aligned} \tag{12}$$

where the first addend in (12) clearly refers to the variability within center. Indeed, if there were no overlapping among centers, the first addend in (12) reduces to the usual variance within stratum.

On the other hand, the second addend in (12) has to do with the variability across centers: the first two terms in it are a measure of variability across centers *adjusted for multiplicity* while the third one is a sort of correction due to overlapping. In fact, if there were no overlapping among centers it is easy to prove that, in the second addend in (12), the third term vanishes while the remaining two terms reduce to the usual variance among strata.

Notice that the first addend in (12) is always non negative but can be greater than σ^2 so that the second addend in (12) can be negative. This is because σ^2 is a definitely theoretical quantity since the parent population is not gathered in a unique set but it is spread over a number of overlapping centers. For instance, the first term in (12) results greater than σ^2 if $\alpha_l \to 0 \ \forall l$ i.e. when the N units are quite scattered over the L centers. This yields a variability within center greater than the theoretical σ^2. Viceversa, if $\alpha_l \to 1 \ \forall l$ each center is *similar* to the population of N units and, as a consequence, the variability within center results closer to σ^2.

Let us focus on the variability within center. By the same argument used for the mean in Section 2, we set

$$\tilde{\sigma}_l^2 = \frac{1}{N_l} \sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \frac{Y_{rq}^2 \, u_{rl}}{m_r^2} - \tilde{\mu}_l^2 \tag{13}$$

denoting the variance of y in the l^{th} center *adjusted for multiplicity*. Yet, the first addend in (12) reduces to the simpler form

$$\frac{1}{N} \sum_{l=1}^{L} \left[\sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \frac{Y_{rq}^2 \, u_{rl}}{m_r^2} - N_l \tilde{\mu}_l^2 \right] = \sum_{l=1}^{L} \alpha_l \, \tilde{\sigma}_l^2 \tag{14}$$

which agrees with the structure of equation (2).

Resorting to (14), equation (7) can be expressed as

$$\mathrm{var}(\bar{y}) = \sum_{l=1}^{L} \left(\frac{N_l}{N_l - 1} \right) \left(\frac{N_l - n_l}{N^2 \, n_l} \right) N_l \tilde{\sigma}_l^2 \tag{15}$$

and ignoring the negligible factors $N_l/(N_l - 1)$, (15) finally reduces to

$$\mathrm{var}(\bar{y}) \cong \frac{1}{N^2} \sum_{l=1}^{L} \left(\frac{N_l}{n_l} - 1 \right) N_l \, \tilde{\sigma}_l^2 \ . \tag{16}$$

Equation (16) is somewhat adequate since center's overlapping is definitely embodied into the variances $\tilde{\sigma}_l^2$ defined by (13). The structure of (16) is now meeting the traditional form of the variance of the estimator of the population mean in the stratified random sampling (Thompson, 2002, p. 120). Accordingly, the optimum allocation can be identified by minimizing the following function

$$f(n_l,\cdots,n_L)=\sum_{l=1}^{L}\left(\frac{N_l}{n_l}-1\right)N_l\,\tilde{\sigma}_l^2$$

with respect to $n_1,\cdots,n_L$ subject to the constraint $n=\sum_{l=1}^{L}n_l$. This yields to the well known solution

$$n_l = n\cdot\frac{\alpha_l\,\tilde{\sigma}_l}{\sum_{l=1}^{L}\alpha_l\,\tilde{\sigma}_l} \qquad (l=1,\cdots,L)\,. \tag{17}$$

Furthermore we can rely on a more realistic cost function assuming that the cost of sampling differs from center to center and the total cost is described by the linear relationship $c=c_0+\sum_{l=1}^{L}c_l n_l$ where c_0 is a fixed cost and c_l is the cost per unit observed in the l^{th} center. It is well known that the optimum is achieved with

$$n_l=(c-c_0)\cdot\frac{\alpha_l\,\tilde{\sigma}_l}{\sqrt{c_l}\sum_{l=1}^{L}\alpha_l\,\tilde{\sigma}_l\sqrt{c_l}} \qquad (l=1,\cdots,L)\,. \tag{18}$$

Accordingly to the stratified sampling case, when no auxiliary information about $\tilde{\sigma}_l^2$ is available, a proportional allocation should be recommended, i.e. $n_l=n\cdot\alpha_l\Big/\sum_{l=1}^{L}\alpha_l$, $(l=1,\cdots,L)$.

4.2 *Circular systematic sampling*

In real applications of CS it is not unusual to deal with centers such as a city square, a rail station or other territorial places for which no frame is available or can be realistically built.

In these practical cases CS is performed by taking a *bunch* of people for interviewing, say the first n_l the interviewer is able to meet in the l^{th} center.

Such a procedure can be formally treated as *circular systematic sampling* since the following issues yield:

– the N_l units present (almost surely) in the l^{th} center at the moment of interview can be considered in random order;

– the first unit met/interviewed can be considered as randomly selected out of the N_l;

– the remaining $n_l - 1$ units taken for interviewing besides the first selected in the random order, can be considered as systematically selected with unitary skip.

The resulting sample of size n_l is circular systematic according to the Singh and Singh's Multiple Distance approach (Hedayat and Sinha, 1991, p. 238) by putting $u \in \{1, \cdots, n_l - 1\}$ and $\delta = 1$. Hence it is equivalent to simple random sample as the inclusion probabilities are the same under both methods (Särndal *et al.*, 1992, p. 77). As a consequence, the results in Section 3 remain unaltered and data can be treated as a simple random sample. In addition, a sufficient condition for second order inclusion probabilities to be positive is $n_l \geq N_l/2 + 1$ (Hedayat and Sinha, 1991, p. 238) i.e. center sample size has to be *sufficiently large*.

4.3 *Hypotheses about* N_l *and* α_l

Up to now we have dealt with the two assumptions

i) the weights $\alpha_l = N_l/N$ are known for each center;

ii) at the moment of selection all of the N_l units attending the l^{th} center are present almost surely.

As these hypotheses could be difficult to respect in the applications, some practical indications are now provided.

Assumption *i)* can be relaxed by resorting upon suitable estimation procedures already proposed in the literature (Migliorati, 2003). To start with notice that, in order to estimate weights α_l, simply an estimator for the relative profile size $N_{\mathbf{u}_r}/N$ is required since $\alpha_l = \sum_{r=1}^{2^L-1} N_{\mathbf{u}_r} u_{rl}/N$. In fact, although estimating the profile size $N_{\mathbf{u}_r}$ is quite a difficult task, estimating the corresponding relative sizes might be easier. A solution consists to rely upon the expected value of the frequencies $f_{\mathbf{u}_r} = \sum_l f_{\mathbf{u}_r,l}$, $(r = 1, \cdots, 2^L - 1)$, for they involve many quantities and the profile sizes are among them. So as r varies from 1 to $2^L - 1$, such expected values form a set of equations which can be simultaneously solved with respect to the unknown $N_{\mathbf{u}_r}$. Since the equation related to the unitary profile is not independent from the others, only a solution for the relative profile sizes $N_{\mathbf{u}_r}/N$ can be found. Although dealing with a non linear system, so that an unbiased estimator for $N_{\mathbf{u}_r}/N$ does not exist, we can achieve an estimator by solving it numerically and then substituting the unknown expected values with the corresponding sample values.

With respect to assumption *ii)*, if not all of the N_l units attending the l^{th} center could be supposed present almost surely while drawing the sample, a possible

way to attack the problem relies on the following observation. If $N_l^{'} \le N_l$ units are actually present then it is possible to define a new set of unknown parameters $\vartheta_{\mathbf{u}_r} = \frac{N'_{\mathbf{u}_r}}{N_{\mathbf{u}_r}}$, $r = 1,\ldots,2^L - 1$ and $\vartheta_{N_l} = \frac{N'_l}{N_l}$, $l = 2,\ldots,L$ expressing the *detectabilities* of units (Thompson, 2002, p. 185). Notably $\vartheta_{\mathbf{u}_r}$ is the probability to detect (i.e. that a unit is present at the moment of drawing) a unit with profile $\mathbf{u}_r$ and ϑ_{N_l} is the probability to detect a unit in the l^{th} center. With those new parameters the fact that the exact number of units present at moment of sampling is unknown is concerned. On the other hand, they are nuisance quantities which have to be eliminated from the analysis. A possible solution is to consider them as random variables and by incorporating any information about them in a suitable prior distribution finally integrate them out. This leads to an integrated likelihood function, i.e. the fundamental inferential tool of this type of approach.

The analysis of consequences on the estimation process is currently under investigation.

5. OTHER SAMPLING DESIGNS

5.1 *Single-stage CS design*

In real applications, in order to achieve a complete population coverage it could be necessary to consider a large number of centers each with low sizes. In this case, a feasible sampling design seems to be a single stage sampling where n centers out of L are selected under a specified design and a census of units attending each selected center is then executed. Yet data are formed through the simultaneous observation of both y values and units profiles.

With the purpose of estimating the population mean μ, the census of the j^{th} selected center $(j = 1,\cdots,n)$ provides the real value of the mean adjusted by multiplicity $\tilde{\mu}_j$ as defined by (1). Hence

$$\overline{y} = \sum_{j=1}^{n} \frac{\alpha_j \, \tilde{\mu}_j}{\pi_j} \tag{19}$$

where π_l stands for the first order inclusion probability of center l, is an unbiased estimator for μ according to known results from single stage cluster sampling (Hedayat and Sinha, 1991, p. 204).

Notice that we can resort to standard results from single stage cluster sampling since overlapping among centers is overcome by means of multiplicity and profiles embedded into $\tilde{\mu}_l$.

Notice also that a similar set up has been considered by Maiti, Pal and Sinha (1993) with the purpose of estimating the population size *N*. It is easy to see that estimator (19), by simply substituting the sample total instead of the sample mean, generalizes Maiti, Pal and Sinha's estimator of *N* to the case of the total of any quantitative characteristic *y* and it coincides when *y*'s values are substituted by 1's. Furthermore, profiles are observable by insuring units to remain anonymous while although this seems an important issue in real applications, it is not guaranteed by the Maiti, Pal and Sinha procedure.

If a simple random sampling design is adopted in selecting centers, estimator (19) reduces to

$$\overline{y} = \frac{L}{n} \sum_{j=1}^{n} \alpha_j \, \tilde{\mu}_j \, . \tag{20}$$

According to standard results of single stage cluster sampling, estimator (20) has variance

$$\mathrm{var}(\overline{y}) = \frac{L(L-n)}{n(L-1)} \left(\sum_{l=1}^{L} \alpha_l^2 \, \tilde{\mu}_l^2 - \frac{\mu^2}{L} \right) \tag{21}$$

which can be unbiasedly estimated by

$$\hat{v}(\overline{y}) = \frac{L(L-n)}{n(L-1)} \left(\sum_{j=1}^{n} \alpha_j^2 \, \tilde{\mu}_j^2 - \frac{n \, \overline{y}^2}{L^2} \right). \tag{22}$$

5.2 *Double-stage CS design*

Due to budgetary constraints or in case of a number of centers substantially large in size or when centers are expected to be essentially homogenous with respect to the characteristic *y* observed, it might not be practical to completely enumerate a selected center. In such situations we can take advantage of subsampling into the selected centers.

By combining results as proposed in Sections 2 and 5.1, a double stage CS can be performed.

Particularly, under a specified design, *n* out of *L* centers are selected at the first stage with inclusion probabilities π_l, $(l = 1,...,L)$. At the second stage, n_j units are drawn from the j^{th} selected center, under a specified sampling design not necessarily equal to the first stage one.

By using profiles, $\pi_{rq,j}$ represents the second stage inclusion probability of the q^{th} unit with profile $\mathbf{u}_r$ attending the j^{th} center selected at first stage.

Referring to the j^{th} center selected at first stage the following estimator

$$\bar{y}_j = \frac{1}{N_j} \sum_{r=1}^{2^L-1} \frac{1}{m_r} \sum_{s=1}^{f_{u_r,j}} \frac{y_{rs}}{\pi_{rs,j}} \tag{23}$$

is unbiased for $\tilde{\mu}_j$ under the second stage sampling design. Hence

$$\bar{y} = \sum_{j=1}^{n} \alpha_j \frac{\bar{y}_j}{\pi_j} \tag{24}$$

is an unbiased estimator for μ according to standard theory for two-stage cluster sampling (Hedayat and Sinha, 1991, p. 209).

Notice that estimator (24), by simply substituting sample totals instead of sample means, agrees with Maiti, Pal and Sinha's estimator for the population size N under two-dimensional sampling (Maiti, Pal and Sinha, 1993). As noted in the previous section, estimator (24) generalises Maiti, Pal and Sinha's estimator of N to the case of the total of any quantitative characteristic y and it coincides when y's values are substituted by 1's.

Resorting to standard results from two-stage cluster sampling, estimator (24) has variance

$$\mathrm{var}(\bar{y}) = \sum_{l=1}^{L} \alpha_l^2 \tilde{\mu}_l^2 \left(\frac{1}{\pi_l} - 1 \right) + \sum_{l=1}^{L} \sum_{\substack{l'=1 \\ l' \neq l}}^{L} \alpha_l \alpha_{l'} \tilde{\mu}_l \tilde{\mu}_{l'} \left(\frac{\pi_{ll'}}{\pi_l \pi_{l'}} - 1 \right) + \\ + \sum_{l=1}^{L} \frac{\alpha_l^2}{\pi_l} \mathrm{var}(\bar{y}_l) \tag{25}$$

where $\pi_{ll'}$ stands for joint inclusion probability of the pair of centers l and l' at the first stage selection and $\mathrm{var}(\bar{y}_l)$ agrees with the general form (4).

Finally, by assuming $\pi_{ll'} > 0$, $(l \neq l' = 1, \cdots, L)$, an unbiased estimator of the variance (25) is given by

$$\hat{v}(\bar{y}) = \sum_{j=1}^{n} \frac{\alpha_j^2 \bar{y}_j^2}{\pi_j} \left(\frac{1}{\pi_j} - 1 \right) + \sum_{j=1}^{n} \sum_{\substack{j'=1 \\ j' \neq j}}^{n} \frac{\alpha_j \alpha_{j'} \bar{y}_j \bar{y}_{j'}}{\pi_{jj'}} \left(\frac{\pi_{jj'}}{\pi_j \pi_{j'}} - 1 \right) + \\ + \sum_{j=1}^{n} \frac{\alpha_j^2}{\pi_j} \hat{v}(\bar{y}_j). \tag{26}$$

6. ILLUSTRATIVE APPLICATIONS

In order to illustrate CS under operational aspects an example in the simple case of three centers (Section 6.1) based upon artificial data and a real application

(Section 6.2) based on data from a real survey on immigrant population conducted in Milan, Italy, in 2002 are now proposed.

In both cases the sample design is simple random sampling with proportional allocation.

6.1 *An example*

Let us consider a simple example involving $L = 3$ centers.

Then we have $2^L - 1 = 7$ possible profiles with multiplicities m_r as shown in Table 1.

TABLE 1

Profiles and multiplicities for $L = 3$

l \ r	1	2	3	4	5	6	7
1	0	0	0	1	1	1	1
2	0	1	1	0	0	1	1
3	1	0	1	0	1	0	1
m_r	1	1	2	1	2	2	3

An artificial population of size $N = 15$ has been randomly generated as well as a quantitative characteristic y taking values over the integers between 16 and 60. The matrix $\mathbf{U}$ of the profiles in the population, the population values Y_i, the center sizes N_l and the center weights $\alpha_l = N_l / N$ are summarised in Table 2.

TABLE 2

Artificial population with N=*15*

l \ i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	N_l	α_l
1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	0	12	0.800
2	1	0	1	0	1	0	1	1	1	0	1	0	0	1	0	8	0.533
3	0	0	0	0	0	0	0	1	1	0	1	0	0	0	1	4	0.267
Y_i	40	39	27	52	56	24	50	44	37	49	36	47	38	44	43		

The population values have been partitioned according to profiles as shown in Table 3.

TABLE 3

Partitioning the population by profiles

r	Y_{rq} $(q = 1, \cdots, N_{\mathbf{u}_r})$	$N_{\mathbf{u}_r}$	Y_r
1	43	1	43
2	40 27	2	67
3		0	0
4	39 52 24 49 47 38	6	249
5		0	0
6	56 50 44	3	150
7	44 37 36	3	117
		$N = 15$	$\sum_r Y_r = 626$

The above framework allows us to compute the means of each center adjusted for multiplicities $\tilde{\mu}_l$ and the population mean μ as given by (1) and (2) respectively:

$$\tilde{\mu}_1 = 30.25, \quad \tilde{\mu}_2 = 22.625, \quad \tilde{\mu}_3 = 20.5, \quad \mu = 41.733$$

Let us now consider a sample of size $n = 10$ proportionally allocated, i.e. $n_1 = 5$, $n_2 = 3$ and $n_3 = 2$. Sample data are reported in Table 4.

TABLE 4

Sample data

l							n_l
1	sampled labels	8	11	7	1	4	
	sampled values	49	38	37	39	24	5
	profiles	1 0 0	1 0 0	1 1 1	1 0 0	1 0 0	
	r	4	4	7	4	4	
2	sampled labels	8	7	2			
	sampled values	44	36	27			3
	profiles	1 1 0	1 1 1	0 1 0			
	r	6	7	2			
3	sampled labels	3	1				
	sampled values	36	44				2
	profiles	1 1 1	1 1 1				
	r	7	7				
							$n = 10$

In Table 5 sample data are partitioned according to profiles.

TABLE 5

Partitioned sample data

l \ r		1	2	3	4	5	6	7	$\bar{y}_l$	n_l
1	y_{rs}	–	–	–	(24, 38, 39, 49)	–	–	(37)		
	$y_{r,l}$	0	0	0	150	0	0	37	32.467	
	$f_{\mathbf{u}_r,l}$	0	0	0	4	0	0	1		5
2	y_{rs}	–	(27)	–	–	–	(44)	(36)		
	$y_{r,l}$	0	27	0	0	0	44	36	20.333	
	$f_{\mathbf{u}_r,l}$	0	1	0	0	0	1	1		3
3	y_{rs}	–	–	–	–	–	–	(36, 44)		
	$y_{r,l}$	0	0	0	0	0	0	80	13.333	
	$f_{\mathbf{u}_r,l}$	0	0	0	0	0	0	2		2
										$n = 10$

With the above results we can compute the sample value of estimator $\bar{y}$ as given by (6) $\bar{y} = 40.373$

while the variance $\text{var}(\bar{y})$ as given by (7) results $\text{var}(\bar{y}) = 24.893$

Finally, the sample value of the unbiased variance estimator as given by (8) and the corresponding conservative estimate (in parenthesis) as given by (9) are

$$\hat{v}(\overline{y}) = 18.895, \quad (32.015)$$

6.2 *An application to real data*

During the last 15 years, illegal immigration has become a key problem in Italy. Essentially because of its geographical position, immigrants come from Easthern Europe, from Northern Africa and from Middle East but also from Asia and from Latin America. The norther Italian regions, which are the heart of the economic and productive activities, attracted the greater part of immigrants and the necessity of quantifying such a phenomenon has consequently increased. Particularly in Lombardia, surveys based on CS have been carried out during the 1990s and in 2000 a permanent Observatory for Integration and Multi-Ethnicity has been founded. Hence, since 2001 an annual survey is conducted as a main tool to plan territorial and migration interventions in the region. Data from a survey carried out in 2002 is now concerned as an application.

Budgetary constraints led to $n = 8000$. The sharing out among the different regional districts took into account activity and magnitude criterions under the constraint that at least 400 immigrants were sampled in each district.

We shall focus on $n = 1100$ immigrants sampled in Milan under a simple random CS design with proportional allocation. Weights $\alpha_l = N_l / N$ were deduced from 2001 data so that $n_l = n \cdot \alpha_l \Big/ \sum_l \alpha_l$.

Furthermore $L = 13$ centers were purposively identified to cover the population of interest as listed in Table 6.

TABLE 6

Centers

l	Center	type	Sample size n_l	$\alpha_l = N_l / N$
1	Reception centers	2	40	0.0992
2	Welfare service centers	2	31	0.0769
3	Language courses	1	152	0.3772
4	Religious centers	3	139	0.3449
5	Medical treatment centers	1	84	0.2084
6	Legal and work aid centers	1	60	0.1489
7	Cultural associations	2	36	0.0893
8	Service and information centers	2	143	0.3549
9	Public offices	2	70	0.1737
10	Fun centers	2-3	111	0.2754
11	Malls and ethnic shops	3	67	0.1663
12	On the street	3	163	0.4045
13	Private houses	3	4	0.0099
	Total		1100	2.73

Center identification took into account administrative, social and private needs of immigrant people. Table 6 shows different typologies of aggregation points such as partial lists (type 1), centers where a list is not available nevertheless some

form of enumeration is possible (type 2) and also centers lacking of any frame and information about sizes (type 3). Notably, type 1 concerns centers where people is requested to give his/her personal data and usually refers to legal immigrant while type 2 refers to centers where attending people are given a ticket or where a fixed number of services is supplied (e.g. meals or beds).

The questionnaire contemplates 32 items in addition to the profile request. We shall focus on two quantitative characteristics: the age $y(A)$ and the monthly private income $y(I)$ expressed in Euros. Notice that the number of possible profiles is $2^L - 1 = 8191$ so that profiles' arrangement can not be listed and some computational support is also needed. Implementation has been fulfilled by *Mathematica* 4.1.

TABLE 7

Estimates of center's means adjusted for multiplicity (Age and Income)

l	1	2	3	4	5	6	7	8	9	10	11	12	13
$\bar{y}(A)_l$	8.64	15.90	16.85	13.01	10.62	15.54	24.79	10.15	14.11	6.29	10.95	8.40	6.92
$\bar{y}(I)_l$	160.3	172.5	254.2	244.7	174.7	126.9	213.9	198.5	189.5	163.1	263.3	189.3	68.5

For both characteristics in each center, the unbiased estimator (3) for the center mean adjusted for multiplicity has been computed and estimates are reported in Table 7. Notice that, although center's means adjusted for multiplicity have little to do with center's mean due to overlapping, comparisons are allowed so that they might be a useful source of information for demographic analyses as it is possible to discriminate among the different centers and to better understand their features. For instance, it clearly emerges that younger people need reception services (center 1) or look for amusement (center 10) while older people feel integrated enough to be involved in cultural activities (center 7).

By applying (6) and data in Table 7, the population's mean estimates result

$$\bar{y}(A) = 32.7435 \qquad \bar{y}(I) = 553.304.$$

Finally, conservative variance estimates obtained from the unbiased estimator (8) by neglecting the finite population corrections $(1 - n_l / N_l)$ for centers of type 3 only are

$$\hat{v}(\bar{y}(A)) = 0.3226\ (0.5131) \qquad \hat{v}(\bar{y}(I)) = 282.199\ (418.82)$$

where numbers in parenthesis are obtained by neglecting all the finite population corrections as given by (9).

APPENDIX: PROOFS

A.1 *Proof of equation (7)*

By substituting in (4) simple random sample inclusion probabilities yields

$$\operatorname{var}(\bar{y}_l) = \frac{1}{N_l^2}\left\{ \sum_{r=1}^{2^L-1} \frac{1}{m_r^2}\left[\sum_{q=1}^{N_{u_r}} \frac{Y_{rq}^2\, u_{rl}\, N_l}{n_l} + \sum_{q=1}^{N_{u_r}} Y_{rq}^2\, u_{rl} \right] \right.$$

$$+ \sum_{r=1}^{2^L-1} \frac{1}{m_r^2} \sum_{q=1}^{N_{u_r}} \sum_{\substack{v=1 \\ v\neq q}}^{N_{u_r}} Y_{rq}\, Y_{rv}\, u_{rl} \left(\frac{1}{N_l - 1} - \frac{N_l}{n_l (N_l - 1)} \right)$$

$$\left. + \sum_{r=1}^{2^L-1} \sum_{\substack{t=1 \\ t\neq r}}^{2^L-1} \frac{1}{m_r\, m_t} \sum_{q=1}^{N_{\mathbf{u}_r}} \sum_{v=1}^{N_{\mathbf{u}_t}} Y_{rq}\, Y_{tv}\, u_{rl}\, u_{tl} \left(\frac{1}{N_l - 1} - \frac{N_l}{n_l (N_l - 1)} \right) \right\}$$

Hence

$$\operatorname{var}(\bar{y}) = \sum_{l=1}^{L} \alpha_l^2 \operatorname{var}(\bar{y}_l)$$

$$= \sum_{l=1}^{L} \frac{\alpha_l^2}{n_l} \left[\sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \frac{Y_{rq}^2\, u_{rl}}{m_r^2\, N_l} - \sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \frac{Y_{rq}^2\, n_l\, u_{rl}}{m_r^2\, N_l^2} - \sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \sum_{\substack{v=1 \\ v\neq q}}^{N_{\mathbf{u}_r}} \frac{Y_{rq}\, Y_{rv}\, u_{rl}}{m_r^2\, N_l (N_l - 1)} \right.$$

$$+ \sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \sum_{\substack{v=1 \\ v\neq q}}^{N_{\mathbf{u}_r}} \frac{Y_{rq}\, Y_{rv}\, n_l\, u_{rl}}{m_r^2\, N_l^2 (N_l - 1)} - \sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \sum_{\substack{t=1 \\ t\neq r}}^{2^L-1} \sum_{v=1}^{N_{\mathbf{u}_t}} \frac{Y_{rq}\, Y_{tv}\, u_{rl}\, u_{tl}}{m_r\, m_t\, N_l (N_l - 1)}$$

$$\left. + \sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \sum_{\substack{t=1 \\ t\neq r}}^{2^L-1} \sum_{v=1}^{N_{\mathbf{u}_t}} \frac{Y_{rq}\, Y_{tv}\, n_l\, u_{rl}\, u_{tl}}{m_r\, m_t\, N_l^2 (N_l - 1)} \right]$$

$$= \sum_{l=1}^{L} \frac{\alpha_l^2}{n_l} \left[\sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \frac{Y_{rq}^2\, u_{rl}}{m_r^2 (N_l - 1)} - \sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \frac{Y_{rq}^2\, n_l\, u_{rl}}{m_r^2\, N_l (N_l - 1)} + \sum_{r=1}^{2^L-1} \frac{Y_r^2\, n_l\, u_{rl}}{m_r^2\, N_l^2 (N_l - 1)} \right.$$

$$-\sum_{r=1}^{2^L-1}\frac{Y_r^2\,u_{rl}}{m_r^2\,N_l(N_l-1)}-\sum_{r=1}^{2^L-1}\sum_{\substack{t=1\\t\neq r}}^{2^L-1}\frac{Y_r\,Y_t\,u_{rl}\,u_{tl}}{m_r\,m_t\,N_l(N_l-1)}+\sum_{r=1}^{2^L-1}\sum_{\substack{t=1\\t\neq r}}^{2^L-1}\frac{Y_r\,Y_t n_l\,u_{rl}\,u_{tl}}{m_r\,m_t\,N_l^2(N_l-1)}\Bigg]$$

where $Y_r=\sum_{q=1}^{N_{\mathbf{u}_r}}Y_{rq}$

$$=\sum_{l=1}^{L}\frac{\alpha_l^2}{n_l}\Bigg[\left(1-\frac{n_l}{N_l}\right)\sum_{r=1}^{2^L-1}\sum_{q=1}^{N_{\mathbf{u}_r}}\frac{Y_{rq}^2\,u_{rl}}{m_r^2(N_l-1)}+\frac{n_l}{N_l^2(N_l-1)}\left(\sum_{r=1}^{2^L-1}\frac{Y_r\,u_{rl}}{m_r}\right)^2$$

$$-\frac{1}{N_l}\left(\sum_{r=1}^{2^L-1}\frac{Y_r\,u_{rl}}{m_r}\right)^2\Bigg]$$

$$=\sum_{l=1}^{L}\frac{\alpha_l^2(N_l-n_l)}{n_l\,N_l(N_l-1)}\left[\sum_{r=1}^{2^L-1}\sum_{q=1}^{N_{\mathbf{u}_r}}\frac{Y_{rq}^2\,u_{rl}}{m_r^2}-\frac{1}{N_l}\left(\sum_{r=1}^{2^L-1}\frac{Y_r\,u_{rl}}{m_r}\right)^2\right]$$

A.2 *Proof of unbiasedness of (8)*

Expression (8) can be rewritten as

$$\sum_{l=1}^{L}\frac{\alpha_l^2}{n_l^2(n_l-1)}\left(1-\frac{n_l}{N_l}\right)\left[n_l\sum_{r=1}^{2^L-1}\sum_{s=1}^{f_{\mathbf{u}_r,l}}\frac{y_{rs}^2}{m_r^2}-\left(\sum_{r=1}^{2^L-1}\frac{y_{r,l}}{m_r}\right)^2\right]$$

$$=\sum_{l=1}^{L}\frac{\alpha_l^2}{n_l(n_l-1)}\left(\frac{1}{n_l}-\frac{1}{N_l}\right)\Bigg[n_l\sum_{r=1}^{2^L-1}\sum_{s=1}^{f_{\mathbf{u}_r,l}}\frac{y_{rs}^2}{m_r^2}-\sum_{r=1}^{2^L-1}\sum_{s=1}^{f_{\mathbf{u}_r,l}}\frac{y_{rs}^2}{m_r^2}-\sum_{r=1}^{2^L-1}\sum_{s=1}^{f_{\mathbf{u}_r,l}}\sum_{\substack{j=1\\j\neq s}}^{f_{\mathbf{u}_r,l}}\frac{y_{rs}\,y_{rj}}{m_r^2}$$

$$-\sum_{r=1}^{2^L-1}\sum_{s=1}^{f_{\mathbf{u}_r,l}}\sum_{\substack{t=1\\t\neq r}}^{2^L-1}\sum_{j=1}^{f_{\mathbf{u}_t,l}}\frac{y_{rs}\,y_{tj}}{m_r\,m_t}\Bigg]$$

$$=\sum_{l=1}^{L}\frac{\alpha_l^2}{n_l}\Bigg[\left(\frac{1}{n_l}-\frac{1}{N_l}\right)\sum_{r=1}^{2^L-1}\sum_{s=1}^{f_{\mathbf{u}_r,l}}\frac{y_{rs}^2}{m_r^2}-\frac{1}{n_l-1}\left(\frac{1}{n_l}-\frac{1}{N_l}\right)\sum_{r=1}^{2^L-1}\sum_{s=1}^{f_{\mathbf{u}_r,l}}\sum_{\substack{j=1\\j\neq s}}^{f_{\mathbf{u}_r,l}}\frac{y_{rs}\,y_{rj}}{m_r^2}$$

$$-\frac{1}{n_l-1}\left(\frac{1}{n_l}-\frac{1}{N_l}\right)\sum_{r=1}^{2^L-1}\sum_{s=1}^{f_{\mathbf{u}_r,l}}\sum_{\substack{t=1\\t\neq r}}^{2^L-1}\sum_{j=1}^{f_{\mathbf{u}_t,l}}\frac{y_{rs}\,y_{tj}}{m_r\,m_t}\Bigg]$$

by using sample membership indicators, results

$$= \sum_{l=1}^{L} \frac{\alpha_l^2}{n_l} \left[\left(\frac{1}{n_l} - \frac{1}{N_l} \right) \sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \frac{Y_{rq}^2}{m_r^2} I_{rq,l} - \frac{1}{n_l - 1} \left(\frac{1}{n_l} - \frac{1}{N_l} \right) \sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \sum_{\substack{v=1 \\ v \neq q}}^{N_{\mathbf{u}_r}} \frac{Y_{rq} Y_{rv}}{m_r^2} I_{rq,l} I_{rv,l} \right.$$

$$\left. - \frac{1}{n_l - 1} \left(\frac{1}{n_l} - \frac{1}{N_l} \right) \sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \sum_{\substack{t=1 \\ t \neq r}}^{2^L-1} \sum_{v=1}^{N_{\mathbf{u}_t}} \frac{Y_{rq} Y_{tv}}{m_r m_t} I_{rq,l} I_{tv,l} \right]$$

Recurring to (5), its expected value has the form

$$\sum_{l=1}^{L} \frac{\alpha_l^2}{n_l} \left[\left(\frac{N_l - n_l}{N_l} \right) \sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \frac{Y_{rq}^2 u_{rl}}{m_r^2 N_l} - \left(\frac{N_l - n_l}{N_l} \right) \sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \sum_{\substack{v=1 \\ v \neq q}}^{N_{\mathbf{u}_r}} \frac{Y_{rq} Y_{rv} u_{rl}}{m_r^2 N_l (N_l - 1)} \right.$$

$$\left. - \left(\frac{N_l - n_l}{N_l} \right) \sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \sum_{\substack{t=1 \\ t \neq r}}^{2^L-1} \sum_{v=1}^{N_{\mathbf{u}_t}} \frac{Y_{rq} Y_{tv} u_{rl} u_{tl}}{m_r m_t N_l (N_l - 1)} \right]$$

$$= \sum_{l=1}^{L} \frac{\alpha_l^2}{n_l} \left[\sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \frac{Y_{rq}^2 u_{rl}}{m_r^2 N_l} - \sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \frac{Y_{rq}^2 n_l u_{rl}}{m_r^2 N_l^2} - \sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \sum_{\substack{v=1 \\ v \neq q}}^{N_{\mathbf{u}_r}} \frac{Y_{rq} Y_{rv} u_{rl}}{m_r^2 N_l (N_l - 1)} \right.$$

$$+ \sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \sum_{\substack{v=1 \\ v \neq q}}^{N_{\mathbf{u}_r}} \frac{Y_{rq} Y_{rv} n_l u_{rl}}{m_r^2 N_l^2 (N_l - 1)} - \sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \sum_{\substack{t=1 \\ t \neq r}}^{2^L-1} \sum_{v=1}^{N_{\mathbf{u}_t}} \frac{Y_{rq} Y_{tv} u_{rl} u_{tl}}{m_r m_t N_l (N_l - 1)}$$

$$\left. + \sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \sum_{\substack{t=1 \\ t \neq r}}^{2^L-1} \sum_{v=1}^{N_{\mathbf{u}_t}} \frac{Y_{rq} Y_{tv} n_l u_{rl} u_{tl}}{m_r m_t N_l^2 (N_l - 1)} \right]$$

which coincides with $\mathrm{var}(\bar{y})$ as shown in A.1. This yields to unbiasedness of (8).

A.3 *Proof of equation (12)*

$$\sigma^2 = \frac{1}{N} \sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} Y_{rq}^2 - \left(\frac{1}{N} \sum_{r=1}^{2^L-1} Y_r \right)^2$$

$$= \frac{1}{N}\sum_{l=1}^{L}\sum_{r=1}^{2^L-1}\sum_{q=1}^{N_{\mathbf{u}_r}}\frac{Y_{rq}^2\,u_{rl}}{m_r} - \frac{1}{N^2}\left(\sum_{l=1}^{L}\sum_{r=1}^{2^L-1}\frac{Y_r\,u_{rl}}{m_r}\right)^2$$

$$= \frac{1}{N}\sum_{l=1}^{L}\sum_{r=1}^{2^L-1}\sum_{q=1}^{N_{\mathbf{u}_r}}\frac{Y_{rq}^2\,u_{rl}}{m_r} - \frac{1}{N^2}\left[\sum_{l=1}^{L}\left(\sum_{r=1}^{2^L-1}\frac{Y_r\,u_{rl}}{m_r}\right)^2 + \sum_{l=1}^{L}\sum_{\substack{k=1\\k\neq l}}^{L}\left(\sum_{r=1}^{2^L-1}\frac{Y_r\,u_{rl}}{m_r}\right)\left(\sum_{r=1}^{2^L-1}\frac{Y_r\,u_{rk}}{m_r}\right)\right]$$

$$= \frac{1}{N}\sum_{l=1}^{L}\left(\sum_{r=1}^{2^L-1}\sum_{q=1}^{N_{\mathbf{u}_r}}\frac{Y_{rq}^2\,u_{rl}}{m_r^2}\right) + \frac{1}{N}\sum_{l=1}^{L}\sum_{\substack{k=1\\k\neq l}}^{L}\left(\sum_{r=1}^{2^L-1}\sum_{q=1}^{N_{\mathbf{u}_r}}\frac{Y_{rq}^2\,u_{rl}\,u_{rk}}{m_r^2}\right)$$

$$- \frac{1}{N^2}\left[\sum_{l=1}^{L}\left(\sum_{r=1}^{2^L-1}\frac{Y_r\,u_{rl}}{m_r}\right)^2 + \sum_{l=1}^{L}\sum_{\substack{k=1\\k\neq l}}^{L}\left(\sum_{r=1}^{2^L-1}\frac{Y_r\,u_{rl}}{m_r}\right)\left(\sum_{r=1}^{2^L-1}\frac{Y_r\,u_{rk}}{m_r}\right)\right]$$

$$= \frac{1}{N}\sum_{l=1}^{L}\left(\sum_{r=1}^{2^L-1}\sum_{q=1}^{N_{\mathbf{u}_r}}\frac{Y_{rq}^2\,u_{rl}}{m_r^2} - N_l\,\tilde{\mu}_l^2\right) + \frac{1}{N}\sum_{l=1}^{L}\left(N_l\,\tilde{\mu}_l^2 - \frac{N_l^2\,\tilde{\mu}_l^2}{N}\right)$$

$$+ \frac{1}{N}\sum_{l=1}^{L}\sum_{\substack{k=1\\k\neq l}}^{L}\left(\sum_{r=1}^{2^L-1}\sum_{q=1}^{N_{\mathbf{u}_r}}\frac{Y_{rq}^2\,u_{rl}\,u_{rk}}{m_r^2} - \frac{N_l\tilde{\mu}_l N_k\tilde{\mu}_k}{N}\right)$$

$$= \left[\frac{1}{N}\sum_{l=1}^{L}\left(\sum_{r=1}^{2^L-1}\sum_{q=1}^{N_{\mathbf{u}_r}}\frac{Y_{rq}^2\,u_{rl}}{m_r^2} - N_l\tilde{\mu}_l^2\right)\right]$$

$$+ \left[\frac{1}{N}\sum_{l=1}^{L}N_l\tilde{\mu}_l^2 - \mu^2 + \frac{1}{N}\sum_{l=1}^{L}\sum_{\substack{k=1\\k\neq l}}^{L}\left(\sum_{r=1}^{2^L-1}\sum_{q=1}^{N_{\mathbf{u}_r}}\frac{Y_{rq}^2\,u_{rl}\,u_{rk}}{m_r^2}\right)\right]$$

Dipartimento di Statistica
Università degli Studi di Milano-Bicocca

FULVIA MECATTI
SONIA MIGLIORATI

REFERENCES

Z.W. BIRNBAUM and M. G. SIRKEN (1965), *Design of sample surveys to estimate the prevalence of rare diseases: three unbiased estimates*, "Vital and Health Statistics". Ser. 2, 11, Government Printing Office, Washington.

G.C. BLANGIARDO (1996), *Il campionamento per centri o ambienti di aggregazione nelle indagini sulla presenza straniera*, in: "Studi in onore di Giampiero Landenna", Giuffrè Ed., Milano.

G.C. BLANGIARDO (2000), *Push and pull factors of international migration. Country report-Italy*, 3/2000/E/n. 5 Eurostat , European Communities Printing Office, Bruxelles.

C. M. CASSEL, C. E. SÄRNDAL and J. H. WRETMAN (1977), *Foundations of inference in survey sampling*, John Wiley & Sons, New York.

M. L. DENNIS and R. IACHAN (1992), *Sampling people who are homeless: implications of multiple definitions and sampling frames*, "Proceedings of the Social Statistics Section, American Statistical Association", pp. 87-96.

W. A. FULLER and L. F. BURMEISTER (1972), *Estimators for samples selected from two overlapping frames*, "Proceedings of the Social Statistics Section, American Statistical Association", pp. 245-249.

H. O. HARTLEY (1974), *Multiple frame metodology and selected applications*, Sankhyā, 36, Series C, pp. 99-118.

A. S. HEDAYAT and K. SINHA, BIKAS (1991), *Design and inference in finite population sampling*, John Wiley & Sons, New York.

S. L. LOHR and J. N. K. RAO (2000), *Inference from dual frame surveys*, "Journal of the American Statistical Association", 95, pp. 271-280.

R. E. LUND (1968), *Estimators in multiple frame surveys*, "Proceedings of the Social Statistics Section, American Statistical Association", pp. 282-288.

P. MAITI, M. PAL and K. SINHA, BIKAS (1993), *Estimating unknown dimensions of a binary matrix with application to estimation of the size of a mobile population*, In *Probability and Statistics*, S. K. Basu and K. Sinha, Bimal. Editors. Narosa Publishing House, New Delhi, pp. 220-233.

F. MECATTI (2002), *Center sampling and dual frame surveys: comparisons among estimators for the total*, "Technical Report", Department of Statistics, University of Milan-Bicocca, Italy. Submitted.

S. MIGLIORATI (2003), *New developments in center sampling*, "Technical Report", Department of Statistics, University of Milan-Bicocca, Italy. Submitted.

C. E. SÄRNDAL, B. SWENSSON and J. H. WRETMAN (1992), *Model assisted survey sampling*, Springer-Verlag, New York.

C. J. SKINNER and J. N. K. RAO (1996), *Estimation in dual frame surveys with complex Designs*, "Journal of the American Statistical Association", pp. 91, 349-356.

S. K. THOMPSON (2002), *Sampling*, 2nd Ed., Wiley, New York.

RIASSUNTO

Campionamento per centri: una strategia per indagini relative a popolazioni elusive

Il campionamento per centri è stato recentemente proposto per condurre indagini relative a popolazioni per le quali non esistono una o più liste esaustive e le unità sono in via naturale aggregate in ambienti o centri sovrapposti. E' quanto accade, ad esempio, per la popolazione straniera illegalmente presente in una nazione. Tale tecnica di campionamento è già stata utilizzata con successo sia in Italia (in particolare in Lombardia) sia a livello europeo. Tuttavia gli aspetti metodologici della medesima sono stati sviluppati solo par-

zialmente. Il presente lavoro propone una formalizzazione di tali aspetti e fornisce uno stimatore corretto per la media di un carattere quantitativo presente sulle unità della popolazione nonché la sua varianza esatta. Inoltre vengono proposti: una stima della varianza corretta nell'ipotesi di campionamento casuale semplice, l'allocazione ottima della numerosità campionaria nel caso di vincoli di costo lineari, l'estensione a disegni di campionamento a due stadi e, infine, una discussione critica degli aspetti operativi della tecnica considerata.

SUMMARY

Center sampling: a strategy for elusive population surveys

Center sampling is useful in finite population surveys when exhaustive lists of all units are not available and the target population is naturally clustered into a number of overlapping sites spread over an area of interest such as, for instance, the immigrant population illegally resident in a country. Center sampling has been successfully employed in official European surveys; nevertheless few systematic theoretical results have been given yet to support empirical findings. In this paper a general theory for Center sampling is formalized and an unbiased estimator for the mean of a quantitative or dichotomous characteristic is proposed together with its exact variance. A suitable estimator for the variance, unbiased under simple random sampling, is also derived and the optimum allocation of the sample size among centers subject to linear cost constraints is discussed. Other sampling designs, useful under operational aspects, are also considered.