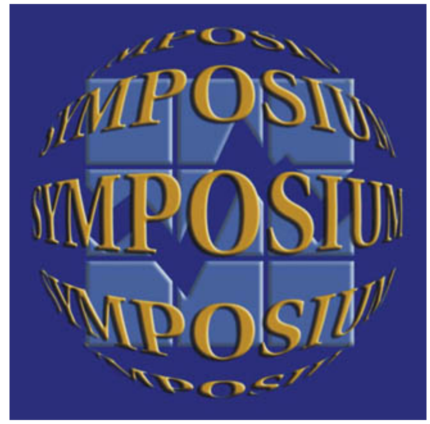
Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2004: Innovative Methods for Surveying Difficult-to-reach Populations

2004



Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2004
Innovative Methods for Surveying Difficult-to-reach Populations

CENTER SAMPLING: A STRATEGY FOR SURVEYING DIFFICULT-TO-SAMPLE POPULATIONS

Fulvia Mecatti[1]

ABSTRACT

During the last 20 years illegal immigration has become a key problem in Italy and the interest of public agencies in surveying immigrant population has consequently grown. The immigrant population is *difficult-to-sample* since exhaustive lists or frames are generally unavailable and units usually demand to remain anonymous so that the customary units' identification by a label is not allowed. Center sampling is a strategy recently developed in Italy for surveying immigrant population by addressing both the concerns "how to reach units for interviewing?" and "how to estimate parameters of interest?" Estimation of the mean of a quantitative variable is considered.

KEYWORDS: Multiplicity; Overlapping Frames; Profile; Sampling on Location.

1. INTRODUCTION

Center Sampling (CS) has been developed in Italy in connection with surveys on immigrant population.

Extra-European immigration has become a key problem in Italy in the last 20 years. Main reasons have perhaps to be seek in the fact that Italy has very rapidly turned from an *emigration country*, until late 60's, into an *immigration country*, since early 90's. Furthermore, its geographical position offers wide shore access so that immigrants are coming primarily from Northern Africa, from Eastern Europe and Middle East but also form Asia and Latin America. The phenomenon has quickly and constantly grown in the last decades often despite the Italian law for immigration so that a considerable part of Extra-European presence in Italy is presently illegal and as a consequence it is out of control and at social risk. In addition, two recent amnesties (1998, 2002) although having the objective of controlling the proportion of illegal residents and besides being limited to special categories, could have encouraged a new wave of illegal immigration. Consequently the need of quantifying and exploring such a phenomenon by public agencies in charge of territorial, migration and social interventions has increased. Periodical surveys on immigrant population started to be conducted since the first half of 90's both at European and local level.

In surveying both legal and illegal immigrants clearly a difficult-to-sample population has to be faced: an exhaustive and accurate list to be used as a frame does not exist, the population size N although finite is unknown so that units are not identifiable by a label and in general traditional sampling theory does not apply. With respect to the legal population component a set of partial lists, from official sources and possibly overlapping, might be available to be considered as multiple frame. On the other hand the illegal component is hidden and elusive. A considerable part may be reasonably supposed to be homeless and jobless, living in charity shelters or open areas like parks, cars or condemned buildings as well as on illegal activities. There is then a detectability problem and a severe requirement of anonymity.

CS is a strategy recently proposed in order to address both the issues "how to sample units" and "how to estimate parameters of interest". In Section 2 the first issue is concerned and the basic formal tools of CS are introduced. Section 3 refers to estimation of parameters of interest. In Section 4 the estimation of the mean under simple random sampling from all centers is considered. Single and Double stage designs are examined in Section 5. Some final remarks are given in Section 6.

[1] Department of Statistics, University of Milan-Bicocca, Via Bicocca degli Arcimboldi, 8, Ed. U7, 20126 Milano, Italy. Fulvia.Mecatti@unimib.it

2. CENTER, PROFILE AND MULTIPLICITY

In the early 90's, at the very beginning of surveys on immigrant population, link-tracing designs such as Network and Snowball sampling, where selected units are requested to report about any other potential unit they are linked with somehow, have been used (see Sudman and Kalton, 1986, for an extensive review). These previous surveys pointed out the characteristic habit of Italian immigrant population to congregate in particular places or territorial sites in order to satisfy several needs: religious, health, social contacts, recreation et cetera. These places are known and well located over the territorial area of interest so that sampling on location can be considered (Sudman and Kalton, 1986).

The term *Center* appeared in the first proposal of CS (Blangiardo, 1996) referring to either a partial list or a territorial place where units congregate. For instance, a mosque is a center where to reach Muslim immigrants, charity shelters are centers where to reach homeless immigrants and also any official register referring to foreign presence in the Country can be considered as a center.

A collection of L centers is then identified in order to ensure an adequate coverage of the population under the assumption that every unit belongs to, or regularly visits, at least one center. Table 1 lists a set of L=13 centers purposively identified to cover the immigrant population in Milan, Italy, in a survey conducted in 2002 by ISMU (Institution for Integration and Multi-Ethnicity). Three types of center are considered: type 1 are essentially partial lists from administrative sources; type 2 are centers where, although a list is not available, nevertheless some sort of units enumeration is allowed for instance centers where people is given a ticket or centers supplying a fixed number of services such as meals or beds. Finally, type 3 are centers lacking of any kind of frame and of information about their absolute size.

Table 1: Centers set covering immigrant population in Milan, 2002

l	Center	type
1	Reception centers	2
2	Welfare service centers	2
3	Language courses	1
4	Religious centers	3
5	Medical treatment centers	1
6	Legal and work aid centers	1
7	Cultural associations	2
8	Service and information centers	2
9	Public offices	2
10	Fun centers	2-3
11	Malls and ethnic shops	3
12	Open air locations	3
13	Private houses	3

In order to deal with the general case, besides the population size N also the center sizes $N_l \ (l = 1 \cdots L)$ are assumed unknown. As units might belong to or visit more than one center, centers usually overlap so that $\sum_l N_l > N$. In addition, due to the anonymity requirement, also the overlapping structure and the overlapping sizes have to be assumed unknown and, in general, no algebraic mapping between units and centers is allowed.

With the purpose of overcoming the lacking of unit identification and of dealing with the overlapping, CS combines elements from Multiple frame survey and from Network sampling by introducing *Profile* and *Multiplicity*.

The collection of labeled units $\{1 \cdots i \cdots N\}$, traditionally denoting the population P, is substituted by a $N \times L$ matrix with units on the rows and centers on the columns, whose entries equal either the digit 1, if unit i belongs to center l, or the digit 0 otherwise as sketched in Figure 1. Note that the sum of each column reproduces the center size N_l.

Every row of the (theoretical) matrix in Figure 1 defines a unit's profile. Thus a profile is a vector of length L informing about in which centers unit i can or cannot be found. Let $\mathbf{u}_r$ be the generic profile included in the (known) class $\mathbf{U}$ of all possible profiles $\left(r = 1 \cdots 2^L - 1\right)$. Since each unit has a unique profile, this implies a many to one mapping from units $\{1 \cdots i \cdots N\}$, or equally from the population set of y-values $\{y_1 \cdots y_i \cdots y_N\}$, onto the class $\mathbf{U}$ of all profiles so that a theoretical partition of the population P is allowed as pictured in Figure 1. Hence, by means of profiles a sort of identification is recovered and the overlapping is overcome.

The collection D_r of units having the same profile $\mathbf{u}_r$ defines a *Domain* as in the Multiple frame context as well as a *Network* in the Network sampling language.

Figure 1: Partitioning the population by profiles

$$
\begin{array}{c}
 \quad 1 \quad \textbf{center} \quad \cdots \quad L \\
\textbf{units}\begin{bmatrix} \\ \vdots \\ N \end{bmatrix}
\begin{bmatrix} 0 & \cdots & 1 & \cdots & 1 \\ & & \vdots & & \\ 1 & \cdots & 0 & \cdots & 1 \\ & & \vdots & & \\ 0 & \cdots & 1 & \cdots & 0 \end{bmatrix} \\
\downarrow \qquad \downarrow \qquad \downarrow \\
N_1 \;\cdots\; N_l \;\cdots\; N_L
\end{array}
\longrightarrow
\begin{array}{c} \mathbf{u}_r \in \mathbf{U} \\ \\ r = 1 \cdots 2^L - 1 \end{array}
\longrightarrow
\begin{array}{c} D_r = \{ i \in P \text{ with profile } \mathbf{u}_r \} \\ \\ \text{with } \bigcup_r D_r = P \text{ and } D_r \bigcap D_t = \varnothing \quad \forall r \neq t \end{array}
$$

The multiplicity of profile $\mathbf{u}_r$, whose elements u_{rl} equal 0 or 1, is defined as $m_r = \sum_l u_{rl}$. Thus the multiplicity indicates the number of centers every unit $i \in D_r$ belongs to. The CS estimation theory presented in the next Sections relies upon multiplicity for dealing with the different chance of each unit to be sampled which clearly depends on the number of centers it belongs to. Parameters and estimators will be weighted via multiplicity in order to apply as widely as possible standard results from sampling theory.

3. ESTIMATION

In the natural context where CS has developed as described in the Introduction, interest first focused on the estimation of the population size N and several proposals have already appeared in literature (see Blangiardo, Migliorati and Terzera, 2004, for a review). As the attention on the topic has increased, population characteristics either than size have been concerned. Let y be a quantitative or dichotomous survey variable. Since N is unknown, we focus on the estimation of the population mean $\overline{Y}$. By turning the attention from the unlabeled units to the profiles $\mathbf{u}_r \in \mathbf{U}$, the parameter $\overline{Y}$ can be expressed as: $\overline{Y} = \sum_r \sum_{i \in D_r} y_i / N$ and by introducing multiplicity

$$\overline{Y} = \frac{1}{N} \sum_l \sum_r \frac{1}{m_r} \sum_{i \in r} y_i \, u_{rl} = \sum_l \alpha_l \, \tilde{\overline{Y}}_l \tag{1}$$

where $\alpha_l = N_l / N$ denotes the weight of center l and

$$\tilde{\overline{Y}}_l = \frac{1}{N_l} \sum_r \frac{1}{m_r} \sum_{i \in D_r} y_i \, u_{rl} \tag{2}$$

indicates the center mean *adjusted* with multiplicity. Note that by introducing profile and multiplicity, the overlapping among center is definitely embedded in the adjusted mean $\tilde{\tilde{Y}}_l$ so that customary associative relationship holds as showed by (1) while it does not, owing to the overlapping, with respect to the (not adjusted) center means $\bar{Y}_l = \sum_{i \in l} y_i / N_l$. Furthermore in typical CS surveys on Italian immigrants, auxiliary information about center weights are usually available, i.e. although the absolute sizes N and N_l are unknown, the center weight α_l can be assumed known. Thus, general results from stratified sampling theory apply: if an unbiased estimator $\bar{y}_l$ for the adjusted center mean $\tilde{\tilde{Y}}_l$ is given, then $\bar{y} = \sum_l \alpha_l \bar{y}_l$ is an unbiased estimator for the population mean $\bar{Y}$ and, by assuming independence among centers, it has variance $V(\bar{y}) = \alpha_l^2 V(\bar{y}_l)$. On the other hand, since sampling can be performed into the overlapping centers only and not among the profiles, hence the analogy with stratified sampling does not help any more and some *ad hoc* theoretical adaptations are required.

In real applications of CS, where immigrant population is concerned, centers are usually heterogeneous with respect to the survey variable so that sampling from all centers, independently, is suggested. Moreover it has been experienced that profiles are observable by allowing units to remain anonymous. Thus, each selected unit is asked to report, along with the survey value y_i, also his/her profile, i.e. what centers, besides the one in which it has been sampled, he/she belongs to.

Let us focus on sample from center l where n_l units are selected according to a given design. Sample data from each center l can therefore be partitioned with respect to the class $\mathbf{U}$ of all profiles. Let $d_r^{(l)}$ be the set of units sampled in center l and sharing the same profile $\mathbf{u}_r$. Since units are not identified but through their profiles, the inclusion probabilities refer to profiles. Let $\delta_{ri}^{(l)}$ be the sample membership indicator of unit i with profile $\mathbf{u}_r$ in the sample from center l i.e. $\delta_{ri}^{(l)}$ is the indicator random variable taking value 1 if unit $i \in D_r$ is included in the sample from center l and 0 otherwise. Then $E\left(\delta_{ri}^{(l)}\right) = \pi_{ri}^{(l)}$ denotes the first order inclusion probability under the given design. The following is an unbiased estimator under a general design for the adjusted mean $\tilde{\tilde{Y}}_l$ as defined by (2) (Mecatti and Migliorati, 2003)

$$\bar{y}_l = \frac{1}{N_l} \sum_r \frac{1}{m_r} \sum_{i \in d_r^{(l)}} \frac{y_i}{\pi_{ri}^{(l)}} \qquad (3)$$

Estimator (3) is a linear combination of profile's estimators of the Horvitz-Thompson type referring to profiles and weighted by multiplicity so that unbiasedness is readily proved. Furthermore, it has (exact) variance:

$$V(\bar{y}_l) = \frac{1}{N_l^2} \left\{ \sum_r \frac{1}{m_r^2} \left[\sum_{i \in D_r} y_i^2 \, u_{rl} \, \frac{1 - \pi_{ri}^{(l)}}{\pi_{ri}^{(l)}} + \sum_{i \neq j} \sum_{\in D_r} y_i y_j u_{rl} \, \frac{\pi_{rij}^{(l)} - \pi_{ri}^{(l)} \pi_{rj}^{(l)}}{\pi_{ri}^{(l)} \pi_{rj}^{(l)}} \right] + \sum_{r \neq t} \sum \frac{1}{m_r m_t} \sum_{i \in D_r} \sum_{j \in D_t} y_i y_j u_{rl} u_{tl} \, \frac{\pi_{ritj}^{(l)} - \pi_{ri}^{(l)} \pi_{tj}^{(l)}}{\pi_{ri}^{(l)} \pi_{tj}^{(l)}} \right\} \qquad (4)$$

where $\pi_{rij}^{(l)} = E\left(\delta_{ri}^{(l)} \cdot \delta_{rj}^{(l)}\right)$ denotes the joint inclusion probability referring to a pair of different units with the same profile $i \neq j \in D_r$ and $\pi_{ritj}^{(l)} = E\left(\delta_{ri}^{(l)} \cdot \delta_{tj}^{(l)}\right)$ indicates the joint inclusion probability related to a pair of units $i \in D_r$ and $j \in D_t$ with different profiles $\mathbf{u}_r \neq \mathbf{u}_t$. Note that the two terms in square brackets in (4) clearly refer to the variability within profiles according to the customary variance of the Horvitz-Thompson estimator while the third addend in (4) is due to the additional variability between different profiles.

4. ESTIMATION UNDER SIMPLE RANDOM SAMPLING

CS has been effectively employed for surveying the Italian immigrant population under simple random sampling (SRS) from every center (see for example, Eurostat, 2000). SRS can be customarily performed when a list of center's units is available or in case of centers in which some sort of units' enumeration is allowed, for instance centers of type 1 and 2 in Table 1. For centers lacking of any kind of list or of units' enumeration, for instance centers of type 3 in Table 1, sampling on location may be performed by previously detecting a time period for visiting the center according to the hypothesis: "the N_l units belonging to the center l are all present *almost surely*". Known habits about the Italian immigrant population and/or special characteristics of the center allow such hypothesis to be addressed. More precisely, a unique day and/or a time indicated as "moment of maximum occupancy of the center" is planned to visit the center and to take a *bunch* of people for interviewing, generally the first n_l the interviewer is able to meet at the entrance or inside the center. As a consequence, sampling on location factors such as time segments, multiple visits of the same center and detectability of units can be ignored (Sudman and Kalton, 1986). Moreover, when the following issues hold:

i) the N_l units, present almost surely in center l at the moment of the interviewers visit, can be assumed in random order,

ii) the first unit met/interviewed can be considered as randomly selected,

iii) the remaining $(n_l - 1)$ units taken for interviewing are supposed to be systematically selected with unitary skip,

then the collection of n_l selected units results in a circular systematic sample according to the Singh & Singh Multiple Distance approach (see, for example, Hedayat and Sinha, 1991, p. 238). Since circular systematic sample is equivalent to SRS, i.e. the inclusion probabilities are the same under both designs (Särndal, Swensson and Wretman, 1992, p. 77), the collection of n_l units selected in center l as described above can be treated as SRS. In addition, the sufficient condition $n_l \geq N_l/2 + 1$ for joint inclusion probabilities to be all positive follows straightforward, i.e. center sample size has to be *sufficiently large*.

Under SRS the inclusion probabilities result

$$E\left(\delta_{ri}^{(l)}\right) = \pi_{ri}^{(l)} = \frac{n_l\, u_{rl}}{N_l} \qquad \text{for all } i \in D_r$$

$$E\left(\delta_{ri}^{(l)} \cdot \delta_{rj}^{(l)}\right) = \pi_{rij}^{(l)} = \frac{n_l(n_l - 1)u_{rl}}{N_l(N_l - 1)} \qquad \text{for all } i \neq j \in D_r \tag{5}$$

$$E\left(\delta_{ri}^{(l)} \cdot \delta_{tj}^{(l)}\right) = \pi_{ritj}^{(l)} = \frac{n_l(n_l - 1)u_{rl}\, u_{tl}}{N_l(N_l - 1)} \qquad \text{for all } i \in D_r,\ j \in D_t \text{ and } r \neq t = 1 \cdots 2^L - 1$$

An unbiased estimator for the mean $\bar{Y}$ under SRS follows by substituting in (3) the first equation in (5) and therefore by associating overall centers (Mecatti, Migliorati and Thompson, 2001)

$$\bar{y}_{SRS} = \sum_l \frac{\alpha_l}{n_l} \sum_r \frac{1}{m_r} \sum_{i \in d_r^{(l)}} y_i \ . \tag{6}$$

Substituting equations (5) in (4), after some simplifications, the (exact) variance of estimator (6) results

$$V\left(\bar{y}_{SRS}\right)=\sum_{l}\frac{\alpha_l^2\left(N_l-n_l\right)}{n_l N_l\left(N_l-1\right)}\left[\sum_{r}\sum_{i\in D_r}\frac{y_i^2\,u_{rl}}{m_r^2}-\frac{1}{N_l}\left(\sum_{r}\frac{Y_r\,u_{rl}}{m_r}\right)^2\right] \tag{7}$$

where $Y_r=\sum_{i\in D_r} y_i$.

Finally, a variance estimator unbiased under SRS is given by

$$\hat{v}\left(\bar{y}_{SRS}\right)=\sum_{l}\frac{\alpha_l^2}{n_l^2\left(n_l-1\right)}\left(1-\frac{n_l}{N_l}\right)\left[n_l\sum_{r}\sum_{i\in d_r^{(l)}}\frac{y_i^2}{m_r^2}-\left(\sum_{r}\sum_{i\in d_r^{(l)}}\frac{y_i}{m_r}\right)^2\right]. \tag{8}$$

From (8), a conservative variance estimate is readily derived by neglecting the finite population corrections $\left(1-n_l/N_l\right)$.

As an example, let us consider real data referring to n=1100 immigrants sampled in Milan, Italy, in 2002 under SRS from every of L=13 centers listed in Table 1. A collection of 8191 possible profiles is concerned. We focus on the survey variable *y:*"age". Table 2 reports the center weights α_l as deduced from 2001 data, the sample sizes proportionally allocated so that $n_l=n\alpha_l\big/\sum_l\alpha_l$ and the estimated center means adjusted with multiplicity according to equation (3). Notice that the values of the estimated center means adjusted with multiplicity do not inform about the correspondent (not adjusted) center means owing to the overlapping; nevertheless they might offer interesting information for demographic and social analyses. For instance younger people turns up needing reception services (center 1) as well as looking for amusement (center 10) while adults look integrated enough to be involved in cultural activities (center 7).

Table 2: Example on real data (Milan, 2002): estimates of center means adjusted for multiplicity

Center l	1	2	3	4	5	6	7	8	9	10	11	12	13
α_l	0.0992	0.0769	0.3772	0.3449	0.2084	0.1489	0.0893	0.3549	0.1737	0.2754	0.1663	0.4045	0.0099
n_l	40	31	152	139	84	60	36	143	70	111	67	163	4
$\bar{y}_l$	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

The estimate for the mean of the population age according to (6) results $\bar{y}_{SRS}=32.74$. The conservative variance estimate, obtained from (8) by neglecting the finite population corrections for centers of type 3, results $\hat{v}\left(\bar{y}_{SRS}\right)=0.3226$ while by neglecting all the finite population corrections is $\hat{v}\left(\bar{y}_{SRS}\right)=0.5131$.

5. OTHER SAMPLING DESIGNS

Besides SRS in every center, two other sampling designs are now concerned.

5.1 Single-stage Center Sampling

In order to reach a sufficient coverage of the population it might be necessary to consider a large number L of centers with low sizes N_l . In this situation it seems convenient to select $n<L$ centers under a given design and then to perform a complete observation of the units in the selected centers. We shall call this procedure *single-stage CS*. Let $l*$ denotes a selected center. Following the CS approach based on profile and multiplicity, both the size N_{l*} and the actual mean adjusted for multiplicity $\tilde{\tilde{Y}}_{l*}$ as given by (2), are known since to center l* belong all units

$i \in D_r$ having the same profile $\mathbf{u}_r$ with $u_{rl*} = 1$. Thus, an unbiased estimator for the population mean, under a general design, is given by (Mecatti and Migliorati, 2003)

$$\bar{y}_{SS} = \sum_{l*} \alpha_{l*} \frac{\tilde{\tilde{Y}}_{l*}}{\pi_{l*}} \tag{9}$$

where π_{l*} denotes (first order) in1clusion probability of the selected center. Moreover, standard theory for single stage cluster sampling applies (see, for example, Hedayat and Sinha, 1991, pp. 204-205). For instance, if centers are selected under SRS, i.e. $\pi_l = n/L$ for all center $l = 1 \cdots L$, estimator (9) has variance

$$V(\bar{y}_{SS}) = \frac{L(L-n)}{n(L-1)} \left(\sum_l \alpha_l^2 \tilde{\tilde{Y}}_l - \bar{Y}^2 / L \right) \tag{10}$$

so that a variance estimator unbiased under SRS is

$$\hat{v}(\bar{y}_{SS}) = \frac{L(L-n)}{n(L-1)} \left(\sum_l \alpha_l^2 \tilde{\tilde{Y}}_l - n\, \bar{y}_{SS}^2 / L^2 \right). \tag{11}$$

5.2 Double-stage Center Sampling

If L is large and several centers are also large in size N_l, it might be impractical to observe units exhaustively in the selected centers. Subsampling into the selected center would also be suggested for budgetary constraints or when units belonging to the same center are expected to be essentially homogenous with respect to the survey variable.

CS approach based upon profile and multiplicity applies by combining results of Sections 3 and 5.1. We shall call this procedure *double-stage CS*. At first stage, $n<L$ centers are selected under a given design with (first order) inclusion probability π_l. Let $l*$ denotes a selected center. At second stage n_{l*} units are drawn into the selected center $l*$ under a possibly different design. An estimator for the center mean adjusted with multiplicity $\tilde{\tilde{Y}}_{l*}$, unbiased under the second stage design is given by (Mecatti and Migliorati, 2003)

$$\bar{y}_{l*} = \frac{1}{n_{l*}} \sum_r \frac{1}{m_r} \sum_{i \in d_r^{(l*)}} \frac{y_i}{\pi_{ri}^{(l*)}} \tag{12}$$

where $\pi_{ri}^{(l*)}$ denotes (first order) inclusion probability under the second stage design referring to unit $i \in D_r$ belonging to center $l*$ selected at first stage. Hence, standard theory for two-stage cluster sampling applies (see, for example, Hedayat and Sinha,1991, p. 209). An unbiased estimator for the population mean is

$$\bar{y}_{DS} = \sum_{l*} \alpha_{l*} \frac{\bar{y}_{l*}}{\pi_{l*}} \tag{13}$$

with variance

$$V(\bar{y}_{DS}) = \sum_l \alpha_l^2\, \tilde{\tilde{Y}}_l^2\, \frac{1-\pi_l}{\pi_l} + \sum_{l \neq h} \sum \alpha_l \alpha_h\, \tilde{\tilde{Y}}_l \tilde{\tilde{Y}}_h\, \frac{\pi_{lh} - \pi_l \pi_h}{\pi_l \pi_h} + \sum_l \frac{\alpha_l^2}{\pi_l} V(\bar{y}_l) \tag{14}$$

where π_{lh} denotes joint inclusion probability of centers $l \neq h$ and $V(\bar{y}_l)$ indicates the variance of estimator (12) according to the general form given in (4). Finally, by assuming $\pi_{lh} > 0$ for all center $l \neq h = 1 \cdots L$, an unbiased variance estimator is given by

$$\hat{v}(\bar{y}_{DS}) = \sum_{l^*} \alpha_{l^*}^2 \bar{y}_{l^*}^2 \frac{1-\pi_{l^*}}{\pi_{l^*}^2} + \sum_{l^* \neq h^*}\sum \alpha_{l^*}\alpha_{h^*}\bar{y}_{l^*}\bar{y}_{h^*} \frac{\pi_{l^*h^*} - \pi_{l^*}\pi_{h^*}}{\pi_{l^*h^*}\pi_{l^*}\pi_{h^*}} + \sum_{l^*} \frac{\alpha_{l^*}^2}{\pi_{l^*}} \hat{v}(\bar{y}_{l^*}). \tag{15}$$

6. REMARKS AND CONCLUSIONS

We conclude with some remarks which could lead to future research on the subject of CS.

First note that estimation theory presented in the previous sections relies upon the hypothesis that center weights α_l are known. At one hand, if on the basis of specific characteristics of the Italian immigrants for which CS has been developed, it would not be difficult to handle with this assumption, on the other hand that might be impractical in dealing with other difficult-to-sample populations. Some proposals for estimating α_l by using the same sample data with both N and N_l unknown, have been appeared in literature (Migliorati, 2002). Thus, results in previous sections can be applied by inserting estimated weights into the proposed estimators. To this extent more research is needed, under both aspects theoretical and empirical, for instance in order to study the effect of estimated weights upon unbiasedness of the estimator of the mean and on the variance estimator.

Secondly, several analogies between CS and Dual/Multiple frame survey are noticeable. In case of all centers having lists available and known sizes N_l, centers are essentially frames and the collections D_r of units with the same profile are domains as traditionally defined (Hartley, 1974). Hence, an unbiased estimator of the total $Y = N\bar{Y}$ follows straightforward from equation (3)

$$\hat{y} = \sum_l N_l \bar{y}_l \tag{16}$$

Note that estimator (16) provides an unbiased estimator for the unknown population size N by simply substituting y-values by 1's. Moreover, for the Dual frame case, i.e. when L=2 , estimator (16) coincides with the Hartley's dual frame estimator with the simplest choice for weighting data from the overlapping domain, i.e. $1/2$. Estimator (16) has been already compared with its major competitors in the Dual frame context (Mecatti, 2002). Simulation results indicate estimator (16) as a feasible alternative with respect to both inferential properties and practical aspects. However, since formulae (3) and (16) hold for any number of frames L>2, further research is needed considering the potentialities of CS approach in the Multiple frame context.

ACKNOWLEDGEMENTS

Thanks are due to Jon N.K. Rao for useful discussions and valuable comments on the subject of CS, and also to Wesley Yung for refereeing an earlier version of this paper and for giving a number of helpful suggestions.

REFERENCES

Blangiardo, G. C. (1996), "Center Sampling or Aggregation Points Sampling for Survey on Foreign Presence" (in Italian), *Studi in onore di Giampiero Landenna*, Milano: Giuffré Ed.

Blangiardo, G. C., Migliorati, S. and Terzera, L. (2004), "Center Sampling: from Applicative Issues to Methodological Aspects", *Proceedings of the XLII Meeting of the Italian Statistical Society*, pp. 377-388.

Eurostat (2000), "Push and Pull Factors of International Migration", *Country Report-Italy, 3/2000/E/n.5* Brussells: Office for Official Publications of the European Communities.

Hartley, H. O. (1974), "Multiple Frame Methodology and Selected Applications", *Sankhyā*, Series C, 36, pp. 99-118.

Hedayat, A. S. and Sinha, B. K. (1991), *Design and Inference in Finite Population Sampling*, New York: J.Wiley & Sons.

Mecatti, F. (2002) "Center Sampling and Dual Frame surveys: comparisons among estimators for the total", paper presented at the XLI Meeting of the Italian Statistical Society, Milan, Italy.

Mecatti, F. and Migliorati, S. (2003), "Center Sampling: a strategy for elusive population surveys", unpublished paper.

Mecatti, F., Migliorati, S and Thompson, S. K. (2001), "Center Sampling: Theory and Estimation", Technical Report #01-06, Department of Statistics, Pennsylvania State University.

Migliorati, S. (2002), "New Developments in Center Sampling", paper presented at the XLI Meeting of the Italian Statistical Society, Milan, Italy.

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

Sudman, S. and Kalton G. (1986), "New Developments in the Sampling of Special Populations", *Annual Review of Sociology*, 12, pp. 401-429.