

Center Sampling: from Applicative Issues to Methodological Aspects

Campionamento per Centri: dalle Problematiche Applicative agli Sviluppi Metodologici

Gian Carlo Blangiardo

Dipartimento di Statistica, Università Milano-Bicocca, giancarlo.blangiardo@unimib.it

Sonia Migliorati

Dipartimento di Statistica, Università di Milano-Bicocca, sonia.migliorati@unimib.it

Laura Terzera

Dipartimento di Statistica, Università di Milano-Bicocca, laura.terzera@unimib.it

Riassunto: Il campionamento per centri è una tecnica proposta recentemente per trattare popolazioni elusive per le quali non esiste una lista completa né un insieme esaustivo di liste incomplete. Il presente lavoro offre una rilettura critica dei numerosi contributi recentemente apparsi in letteratura sull'argomento, evidenziando gli sviluppi metodologici nonché alcuni risultati applicativi e proponendo altresì una visione unitaria sia sotto profilo formale sia con riguardo alla collocazione di ogni contributo nel quadro generale del problema investigato.

Keywords: elusiveness, profile, multiplicity, detectability

1. Introduction

Over the last twenty years, the migrant phenomenon has been gathering momentum in Italy, becoming increasingly complex and varied. At the same time, it is extremely difficult to obtain reliable information about it, on account of the elusive, mobile nature of at least part of the population involved, together with the shiftiness of the illegal component. Facing an evasive population and without a complete list or even a set of incomplete lists - considering the need to preserve anonymity - means that obtaining reliable information cannot be done through traditional sampling techniques.

In order to deal with the above problems, Center Sampling (CS) (Blangiardo, 1996) was proposed. This technique is based on the migrants' need to attend at least one aggregation point (i.e. a "center") for social contacts, health care, religion, leisure or simply for everyday needs. Notice that although many different proposals for estimating rare and elusive populations are present in the literature, they are not suitable for tackling immigrant populations. For example, in the multiple frame theory (Hartley, 1974; Skinner, 1991; Lohr and Rao, 2000; Byczkowski *et al.*, 1998) frame sizes as well as overlapping sizes are supposed to be known, which is not a realistic hypothesis for the phenomenon under investigation. Then, the use of area frames (Haines and Pollock, 1998) assumes that the area of interest is divided into mutually exclusive and exhaustive sampling units called segments, and that only the sampled segments are exhaustively

observed. Conversely, in CS, no center can be left out due to its heterogeneity, which calls for samples to be drawn from each center. Similar remarks apply to methods based on joining list frames and area frames or to network sampling (Thompson, 2002). It is remarkable that even snowballing (Goodman, 1961; Berg, 1983) which has been used for surveys on illegal Mexican immigrants in the United States (Cornelius, 1982) does not work, because of the great number of migrant nationalities present in Italy, together with the fact that social links are limited to each ethnic group. Consequently, the identification of initial units would be very complex and expensive, some small ethnic communities could be easily missed, and inclusion probabilities are unknown, so that this technique seems more appropriate for explorative surveys. Finally, it is noteworthy that CS has some similar aspects with location sampling (Sudman and Kalton, 1986) for which, however, no methodological in-depth study is present in the literature.

The aim of the present paper is to give an overall view of the specific application problems that rise in the outlined context, together with the methodological solutions proposed in the literature. More precisely, in section 2 the CS technique is described; section 3 is devoted to an analysis of the main problems and early proposals, while section 4 focuses on the more recent ones. Finally, an overview of open problems and possible future developments will be given in section 5.

2. Center Sampling Technique

Some specific aspects of migrant reality prompted CS scheme, i.e. the need of setting up various kinds of relationships between the foreigner present on Italian soil (irrespective of her/his administrative position) and a combination of institutions, meeting places and various kinds of centers on one side, and the chance to make a kind of census of such centers on the other.

With this premise the CS technique can be described as follows. The basic assumptions are that the centers are known (or have been identified in previous surveys), they are a finite number L , they can overlap and they cover the entire target population. Nevertheless the size both of centers and of overlapping (i.e. the number of people attending more than one center) is generally unknown. In other words centers can be seen as overlapping strata with unknown sizes. Besides in CS sampling units are drawn from each center since these are heterogeneous and thus sources of specific information as mentioned above.

Thus let N be the unknown population size and N_l the unknown size of center C_l ($l=1, \dots, L$). Notice that $\sum_l N_l \geq N$ since each unit can attend more than one center. Besides let Y be the characteristic of interest and let $\{Y_1, \dots, Y_i, \dots, Y_N\}$ define the population values, i.e. the parameter of the population if the units were identifiable. Notice that in the CS framework units cannot be identified with a label. A fundamental tool in CS theory is the “profile”, i.e. a vector of length L , whose l^{th} element can be 0 or 1: 0 if the unit does not attend the center l and 1 otherwise. Hence profiles inform about the units’ center attendance. It is remarkable that the overall number p of profiles is $p = 2^L - 1$ as a profile with null elements is not considered. Let $\mathbf{u}_r = [u_{r1}, \dots, u_{rl}, \dots, u_{rL}]$ be the generic profile ($r = 1, \dots, p$), $N_{\mathbf{u}_r}$ be the unknown number of units with profile $\mathbf{u}_r$.

in the population and $\{Y_{rq}; q=1, \dots, N_{\mathbf{u}_r}\}$ be the subset of the parameter of the population with respect to units with profile $\mathbf{u}_r$.

CS requires that an overall sample of size n is selected drawing a simple random sample of fixed size n_l independently from each center, where $\sum_l n_l = n$. Every sampled unit is asked to fill in a questionnaire which includes the unit's profile; notice that giving one's profile does not compromise anonymity. Thus such a sample provides the quantities $f_{\mathbf{u}_r, l}$ which denote the sample frequencies of the r^{th} profile from center C_l ($r=1, \dots, p; l=1, \dots, L$), $f_{\mathbf{u}_r} = \sum_l f_{\mathbf{u}_r, l}$, i.e. the frequencies of units with profile $\mathbf{u}_r$ ($r=1, \dots, p$) and the set of sample values $\{y_{rs}; s=1, \dots, f_{\mathbf{u}_r}\}$ observed for units with profile $\mathbf{u}_r$.

3. Early proposals

3.1. Calculating the weights

Notice that the sample obtained by means of the CS technique is not representative; in fact the unknown probability of inclusion of each unit is higher the more "ls" appear in corresponding profile $\mathbf{u}_r$ and, moreover the lower the size N_l of centers frequented. This problem can be solved (Blangiardo, 1996, 2000) by providing a set of weighting coefficients $c_{\mathbf{u}_r}$, functions of the profile, so as to guarantee that the weighted sample possesses, in relation to the distribution of the profiles of the n sample units, the same representational requisites as a hypothetically simple random sample. The weighted coefficients should take the form:

$$c_{\mathbf{u}_r} = \frac{N_{\mathbf{u}_r} / N}{f_{\mathbf{u}_r} / n} \quad (1)$$

which is not operative, the numerator being unknown. Notice that the estimator $F_{\mathbf{u}_r} / n$ described by the relative frequency of profile $f_{\mathbf{u}_r} / n$ is biased for $N_{\mathbf{u}_r} / N$. Its expected value is in fact

$$E\left[\frac{F_{\mathbf{u}_r}}{n}\right] = \frac{N_{\mathbf{u}_r}}{n} \sum_{l=1}^L \frac{n_l}{N_l} u_{rl} \quad (2)$$

From the latter it can be seen that when the quantities

$$\alpha_l = N_l / N \quad (3)$$

are known, from (2) it is possible to get an unbiased estimate for $N_{\mathbf{u}_r} / N$ which leads to the following estimate of the weighting coefficients:

$$\hat{c}_{u_r} = \left(\sum_{l=1}^L \frac{n_l}{n \alpha_l} u_{rl} \right)^{-1} \quad (4)$$

In particular, if attention is fixed on a proportional allocation of sample size n , namely $n_l = n q_l / q^*$ where q_l represents a previously known initial weight and $q^* = \sum_l q_l$, the estimator (4) takes the form:

$$\hat{c}_{u_r} = q^* / \sum_{l=1}^L \frac{q_l}{\alpha_l} u_{rl} \quad (5)$$

Notice that the hypothesis that ratios (3) are known can be considered realistic only where preliminary information on the relative importance of the centers is available. In all other cases a possible solution to avoid such a problem is to estimate α_l not in absolute but in relative terms, that is to focus on ratios $\alpha_l / \alpha_{l'}$ ($l, l' = 1, \dots, L; l \neq l'$) which can be estimated using the following expression

$$(\alpha_l / \alpha_{l'}) = \frac{f_{l|l'} / n_{l'}}{f_{l'|l} / n_l} \quad (6)$$

where $f_{l'|l} = \sum_r f_{u_r, l} u_{rl'}$ is the number of those who, having been sampled in center l , also frequent center l' and viceversa for $f_{l|l'}$.

Besides, having chosen a “base” center of reference (from now on marked as “b”), (6) also makes it possible to estimate the relative version c'_{u_r} of the coefficients c_{u_r} :

$$c'_{u_r} = \frac{q^*}{\sum_{l=1}^L \frac{q_l / q_b}{\alpha_l / \alpha_b} u_{rl}} \quad \hat{c}'_{u_r} = \frac{q^*}{\sum_{l=1}^L \frac{(q_l / q_b)(f_{l'|l} / n_{l'})}{(f_{l|l'} / n_{l'})} u_{rl}}$$

Finally, bearing in mind the constraint of normalization $\sum_r \hat{c}_{u_r} f_{u_r} = n$, the definitive form of the coefficients estimate is:

$$\hat{c}_{u_r} = \frac{n}{\sum_{l=1}^L \frac{(q_l / q_b)(f_{l'|l} / n_{l'})}{(f_{l|l'} / n_{l'})} u_{rl} \sum_{r=1}^p \frac{f_{u_r}}{\sum_{l=1}^L \frac{(q_l / q_b)(f_{l'|l} / n_{l'})}{(f_{l|l'} / n_{l'})} u_{rl}}} \quad (7)$$

It has been shown (Migliorati, 1997) that the estimates proposed, though not endowed with any particular properties for finite size samples, are asymptotically unbiased and consistent as the size of the sample drawn from the base center diverges. From this, the hypothesis being of proportional allocation, the operative indication emerges to favour as base center the one with the highest a priori weight q_b .

3.2 Estimators

Having re-proportioned the sample by means of the estimates of the weighting coefficients illustrated in the previous section, it is possible to determine the estimates of the quantities of interest, particularly of the mean μ of a generic quantitative character Y and of the overall size of the population N .

As regards the first we proceed by simply calculating the weighted sample mean, where the weights of each sample unit are represented by weighting coefficients $\hat{c}_{u_r}$, coming from (4) where the ratios α_i are known, or from (7) in the opposite case (Blangiardo, 2000). In this way we obtain the estimate:

$$\hat{y} = \frac{1}{n} \sum_{r=1}^p \sum_{s=1}^{f_{u_r}} y_{rs} \hat{c}_{u_r} . \quad (8)$$

Formula (8) is proposed in Mecatti (2002a), where it is shown that if the α_i ratios are known, the corresponding estimator is unbiased for μ and the variance of the estimator is provided, as well as an estimator for the latter. With reference to the estimate of N , it is possible to proceed singularly for each nationality (a characteristic always observed in every survey concerning foreigners) and to then aggregate the estimates as regards the great areas of origin (Blangiardo, 2000). Notice that, in the absence of any information about the sizes of centers N_i and about the overlappings, there are not enough elements to obtain an estimate of N . However, it is always possible to fall back on information from official sources. In particular, the method proposed resorts to registry records where the size N_A is known also at an individual nationality level (from now on indicated by N_{Ai} , where i represents the generic i -th nationality).

Having thus focused on the latter, let $_{nat(i)}Y$ be the “presence of the i -th nationality” characteristic and $_AY$ the “registry listed” one. Both prove to be dichotomous variables useful for determining the following sample quantities:

$$n'_i = \sum_{r=1}^p \sum_{s=1}^{f_{u_r}} _{nat(i)}y_{rs} \hat{c}_{u_r} , \quad n'_{Ai} = \sum_{r=1}^p \sum_{s=1}^{f_{u_r}} _AY_{rs} _{nat(i)}y_{rs} \hat{c}_{u_r} \quad (9)$$

of which the first represents the number of sample units belonging to the i -th nationality and the second the number of units of the said nationality also officially registered.

The n'_{Ai}/n'_i ratio is an estimate of the corresponding relative frequency. However, since the i -th nationality can be seen as an element of a macro-area (for example, if the nationality in question is Albanian the matching macro-area will be East Europe) it seems right to adjust the estimate, bringing it closer to that of the corresponding macro-area. Given the empiric proof of similarity of behaviour depending on macro-area of origin, this can be accounted for the higher sample size of the macro-area and thus the greater reliability of the estimates pertaining to it. Therefore, indicating with $_jn'$ the sample size of the j -th macro-area and with $_jn'_A$ the frequency of individuals belonging

to macro-area j and officially registered, the adjustment applied to the sample proportion relating to the i -th nationality leads us to identify the quantity ${}_j\tilde{z}_{Ai}$, which is:

$${}_j\tilde{z}_{Ai} = \frac{n'_{Ai}}{n'_i} - \frac{\sqrt{{}_jn'}}{(\sqrt{{}_jn'} + \sqrt{n'_i})} \left(\frac{n'_{Ai}}{n'_i} - \frac{{}_jn'_A}{{}_jn'} \right) \quad (10)$$

In (10) the original proportion is adjusted by means of a quantity proportional to the distance between the same and the corresponding proportion relating to the macro-area. Notice also that the proportionality constant is a function of the estimates of the variances of the said proportions, estimates obtained by considering a condition of maximum variability. This expression can be formally justified by means of the Chebyshev inequality (Marasini e Migliorati, 2003). Finally, for the i -th nationality an estimate of its size can be obtained in the following way:

$${}_j\hat{N}_i = N_A / {}_j\tilde{z}_{Ai} \quad (11)$$

and, aggregating, we obtain an estimate of the overall size:

$$\hat{N} = \sum_j \sum_i {}_j\hat{N}_i \quad (12)$$

where j numbers the macro-areas and i the nationalities in each of them. Since we are interested in knowing whether the foreign component is legal or not, with similar ratios to (11) and (12) we obtain an estimate of the size of legal positions and a break-down of the overall estimate into component parts: residents, legal non-residents and illegal ones.

3.3 Application to real data

The CS technique has found various applications in the subject of migration since 1990. The most important contributions, both for variety and territorial detail of the information provided, as well as the ample cover of the population, can be traced to the surveys conducted by I.S.MU. (Foundation for Integration and Multi-ethnicity), as part of the “provincial observer” activities launched in 1996 in some Lombard areas, and joining up with the Regional Observatory experience in 2001. It has thus been possible to get to know and assess the characteristics of the foreign presence in all its component parts – both legal and illegal – over such a vast territory, particularly interesting for the study of the phenomenon. Below can be found some simple examples concerning the estimates of the number of foreigners obtained from the researches.

Through the methodology illustrated in section 3.2, the regional surveys have afforded a particularly interesting picture (table 1) for assessing the reliability of these estimates.

The results of these surveys, making it possible to estimate the illegality rate, afford the opportunity of understanding both the effects of containment of the illegal factor produced by the amnesties of 1998 and 2002, and its significant (and generalized) increase following close on the amnesties themselves. Notice, for example, that the

percentage of illegality rises from 21% for the whole of Lombardy on 1 January 2001 to 31% a year later. This situation can feasibly be ascribed as the effect of the 2002 amnesty itself. It can be held that in the course of 2001 expectation of a tightening up by the government in the matter of immigration flows, together with the chance of another amnesty, even if limited to very specific categories, encouraged a new wave of illegal immigration a year after the effects of the amnesty of 1998 seemed to have run their course.

Table 1 – Estimate of the number of illegal immigrants for every 100 foreigners present in Lombardy

Area	31.12.1998	30.6.1999	30.6.2000	1.1.2001	1.1.2002	1.7.2003
Milano city	20	17	16	22	36	14
Prov. of Milano	30	24	20	22	35	13
Varese			22	17	26	7
Como				20	29	8
Sondrio				24	26	13
Bergamo				22	23	8
Brescia				18	27	9
Pavia				27	28	11
Cremona			14	19	29	8
Mantova			16	15	19	8
Lecco				16	26	9
Lodi		29		23	25	7
Lombardy				21	31	11

Source: Lombardy - I.S.MU. Foundation and Regional Observatory for Integration and Multi-ethnicity.

4. More recent proposals

Many aspects of the earlier proposals have been investigated from the methodological point of view giving rise to interesting new solutions for the problems described above, particularly where the estimate of the mean, population size and relative profile sizes are concerned.

4.1 Estimate of the mean

In Mecatti, Migliorati (2003a, 2003b) an estimate of the mean μ alternative to (8) is proposed under the hypothesis that ratios α_l defined by (3) are known, i.e.:

$$\bar{y} = \sum_{l=1}^L \frac{\alpha_l}{n_l} \sum_{r=1}^p \frac{1}{m_r} \sum_{s=1}^{f_{u_r,l}} y_{rs} = \sum_{l=1}^L \sum_{r=1}^p \sum_{s=1}^{f_{u_r,l}} \frac{y_{rs}}{\beta_{rl}} \quad (13)$$

where $m_r = \sum_{l=1}^L u_{rl}$ denotes the number of centers attended by units with profile $\mathbf{u}_r$, i.e. the *multiplicity* of profile r , and $\beta_{rl} = (n_l m_r) / \alpha_l$. The exact variance and an estimate of it are also provided. Estimate (13) is then generalized to adapt to different sampling designs, i.e. a single stage sampling where a sample (not necessarily simple random)

from each center is drawn, a single stage sampling where a fixed number of centers are selected under a specified design and a census of units attending each selected center is performed and, moreover, a double stage design where both centers and units inside each center are drawn. Besides the optimum choice of sample sizes n_l under different cost constraints is carried out.

The two estimators $\hat{Y}$ and $\bar{Y}$, described by (8) and (13) under the hypothesis that ratios α_l are known, are compared in Mecatti, Migliorati (2003b) where it emerges that the main difference between the two lies in the way sampling information from each center is synthesized. In fact $\hat{Y}$ makes use of the weights $\hat{c}_{u_r}$ given by (4) which average over the L centers while $\bar{Y}$ resorts to center specific weights β_{rl} . A further investigation shows that if the sampling fraction n_l / N_l is constant, i.e. if the allocation of n among the centers is proportional to their size, the two estimators coincide; if the sampling fractions exhibit small variability $\hat{Y}$ is more efficient while the reverse is true where the sampling fraction variability is large.

Let us now consider an applicative example which focuses on 2002 data coming from 2003 I.S.MU. survey. Calculation of estimates (8) and (13) concerning immigrants sampled in Milan centers (1100 altogether) leads to the following values $\bar{y} = 32.74$ $\hat{y} = 32.77$ for the “Age” character and $\bar{y} = 553.30$ $\hat{y} = 598.24$ for the “Private monthly income (expressed in euros)” character. Notice that the first two are quite similar as the variability of the sampled values is quite small while the second two differ appreciably. Obviously the ratios α_l , which are supposedly known in the application of (13), have been estimated from previous surveys. In a more general context where this is not so feasible the identification of an estimate for such ratios becomes a fundamental issue.

4.2 Estimate of relative profile sizes

Estimating the relative profile sizes N_{u_r} / N is quite useful as they make it possible to estimate the weights c_{u_r} defined by (1). This gives rise to an alternative way of adjusting the sample as described in section 3.1 and, consequently, to alternative estimates of the mean μ and of the population size N given by (8) and by (13). Furthermore observing that

$$\alpha_l = \frac{N_l}{N} = \sum_{r=1}^p N_{u_r} u_{rl} / N$$

once the relative profile sizes are estimated an obvious estimate of the ratios α_l defined by (3) is also attainable. Thus the assumption the ratios α_l are known can be relaxed and estimate (13) becomes operative in any context.

The problem of estimating N_{u_r} / N has been treated in Migliorati (2002), where observing that the center sizes N_l are themselves a function of the profile sizes N_{u_r} , the expected value (2) is written in the form

$$E(F_{u_r}) = N_{u_r} \sum_l n_l u_{rl} \left(\sum_r N_{u_r} u_{rl} \right)^{-1} = N_{u_r} \left(\mathbf{n}' \mathbf{D} \mathbf{u}_r' \right) \quad (14)$$

($r = 1, \dots, p$) where $\mathbf{n}' = [n_1, \dots, n_L]$ is the vector of center sample sizes and $\mathbf{D}$ is the $L \times L$ diagonal matrix whose l^{th} entry on the diagonal is $(\mathbf{N}'_{\mathbf{u}} \mathbf{u}_{\cdot l})^{-1}$, $\mathbf{N}'_{\mathbf{u}} = [N_{\mathbf{u}_1}, \dots, N_{\mathbf{u}_{2L-1}}]$ is the vector of profile sizes and $\mathbf{u}_{\cdot l}$ is the l^{th} column of the matrix $\mathbf{U}$ whose rows are the profiles $\mathbf{u}_r$. Solving equations (14) simultaneously with respect to $N_{\mathbf{u}_r} / N$ leads to an estimate of them once the expected values of the random variable $F_{\mathbf{u}_r}$ have been substituted by the corresponding sampling values. The properties of the estimators proposed and some computational issues are also discussed in the paper cited.

4.3 Estimate of the population size

Some alternative proposals exist for estimating the size N . In Migliorati, Terzera (2002) the hypothesis that at least one center can be compared to an incomplete list, and thus has a supposedly known size, is assumed. Besides, two methods are proposed for estimating the profile sizes, both relying on the observation that the equality $N = \sum_r N_{\mathbf{u}_r}$ holds.

The first method is based on the maximizing of the likelihood function relating to the units drawn from the center with a known size, thus arriving at an estimate $\hat{N}_{\mathbf{u}_r}$ of the sizes of profiles that entail the center attendance. Having substituted these estimates in the likelihood function relating to the samples coming from the remaining centers, the maximization of the latter makes it possible to obtain an estimate for the size of the residual profiles through the solution of the linear system $\mathbf{A}\tilde{\mathbf{N}}_{\mathbf{u}} = \mathbf{c}$, where $\mathbf{A}$ is a $(p-1)/2 \times (p-1)/2$ matrix, whose generic element a_{rj} is

$$a_{rj} = \begin{cases} 1 & \text{se } r = j \\ -\frac{f_{\mathbf{u}_{r+(p-1)/2}}}{\sum_{l=2}^L n_l - f_{\mathbf{u}_{r+(p-1)/2}}} & \text{se } r \neq j \end{cases} \quad (15)$$

$\tilde{\mathbf{N}}'_{\mathbf{u}} = [N_{\mathbf{u}_{(p+1)/2}}, \dots, N_{\mathbf{u}_{p-1}}]$ is a sub-vector of the vector of the unknown sizes referring to profiles which do not entail the attendance of the center assimilated to an incomplete list, and the generic element c_r of vector $\mathbf{c}$ of the known terms has the following expression:

$$c_r = \left(\sum_{i=1}^{(p-1)/2} \hat{N}_{\mathbf{u}_i} \right) \frac{f_{\mathbf{u}_{r+(p-1)/2}}}{\sum_{l=2}^L n_l - f_{\mathbf{u}_{r+(p-1)/2}}}, \quad (r = 1, \dots, (p-1)/2).$$

The second method introduces a corrective element with the aim of removing the hypothesis that all the units are almost certainly present at the moment of sampling. This can be carried out by introducing the notion of detectability (Thompson, 2002),

that is the probability that a statistical unit can be observed at the moment of sampling and defined as

$$\mathcal{G}_{\mathbf{u}_r} = N'_{\mathbf{u}_r} / N_{\mathbf{u}_r}, \quad r = 1, \dots, p$$

where $N'_{\mathbf{u}_r}$ represents the actual number of units visible with profile $\mathbf{u}_r$.

Quantities $\mathcal{G}_{\mathbf{u}_r}$ appear naturally in the likelihood function, presenting themselves as nuisance parameters. However, since it is not possible to eliminate them through maximization, they have been treated as random variables for which appropriate a priori distributions have been elicited. In particular, Beta distributions truncated in the lower extreme have been used, the latter a function of the profile multiplicity m_r , so as to guarantee greater visibility for individuals attending a very small number of centers.

The elimination took place through integration, after having introduced appropriate transformations, leading to an integrated likelihood. The estimates of as yet unknown profile sizes have finally been deduced by the maximization of the integrated likelihood, which can only come about numerically. A comparative evaluation of the proposed estimates has been carried out in the above-mentioned article.

Further proposals for estimating the population size together with the expression of the variance of the estimator and of its unbiased sample estimator can be found in Pratesi, Rocco (2002, 2003), although the authors assume some hypotheses different from the standard CS ones (the observation of all the units in the selected centers in the one-phase design (2002) and the knowledge of the center sizes N_l in the two-phase one (2003)).

4.4 Small area issues

The estimate of the population size N proposed in section 3.2 calls for the adjustment (10) of the sampling proportion n'_{Ai}/n'_i which refers to people belonging to the i^{th} nationality and present in the registry records and which represents an estimate of the corresponding relative frequency $\mathcal{G}_{Ai}$ in the population. Notice that adjustment (10) allows to make such a proportion closer to the proportion ${}_j n'_A / {}_j n'$ that refers to the whole macro area j the nationality of concern belongs to and which is an estimate of the corresponding relative frequency ${}_j \mathcal{G}_A$ in the population.

Clearly this is a “small area” or “domain” estimation problem (Rao, 2003; Kish, 1987) as the issue derives from the fact that the sub-population considered (i.e. the i^{th} nationality) is quite small, thus inducing a small sub-sample.

Alternative ways to make such an adjustment have been proposed (Marasini and Migliorati, 2003). These make it possible to identify composite estimators which make a synthesis between the information deriving from the sub-sample (related to the i^{th} nationality) on one side and the information from the whole sample (related to the whole macro area) on the other. In such a way both the specificity of the domain and the reliability of the larger sample are taken into account.

A first procedure based on Chebyshev inequality leads to adjustment (10). Instead, two Bayesian approaches (one empirical and one hierarchical) make it possible to determine new point and interval estimates of $\mathcal{G}_{Ai}$. They both provide the information pouring between the smaller and the larger sample through the hyper-parameters of the prior

distribution given to $\mathcal{G}_{Ai}$ and ${}_j\mathcal{G}_A$. In an empirical analysis the hyper-parameters which enter the posterior mean of $\mathcal{G}_{Ai}$ are specified numerically by means of some sampling quantities obtained from the macro area sample. In a hierarchical approach they are replaced by the posterior modes of the hyper-parameters of the prior for ${}_j\mathcal{G}_A$. Interesting remarks on the comparison between the different estimators can be made through simulative studies.

5. Open problems and future developments

If on one side the present work affords a clearer picture of the characteristics and potential of CS technique, on the other it reveals various aspects partially or even totally unexplored and interesting for future investigations.

A first example can be found in further applications of the most recent methodological aspects, which would make it possible to check more carefully the operational possibilities of such methods, partly by comparing them with the first proposals.

Again, a particularly interesting development could spring from the introduction of an estimate of relative profile frequencies (as set out in section 4.2) in the estimates of parameters obtained supposing such quantities to be known, as for example the one relating to the mean (section 4.1), as of today, calculated supposing that quantities α_l are known.

Besides, having additional information concerning the detectability at our disposal (easily obtainable through sample data concerning a measure of “regularity” of attendance at the various centers) the estimate of the overall size, set out in section 4.3, could at the same time be improved and simplified in the computational aspects.

Notice, finally, that the potential of the CS technique is not limited to the immigrant population alone, since it can easily be applied to many situations where the population in question presents elusive characteristics, rareness and thus a lack of frames.

References

- Berg S. (1983) Random contact process, snowball sampling and factorial series distributions, *Journal of applied probability*, 20, 31-46.
- Blangiardo G.C. (1996) Il campionamento per centri o ambienti di aggregazione nelle indagini sulla presenza straniera, in: *Studi in onore di Giampiero Landenna*, Giuffrè Ed., Milano.
- Blangiardo G.C. (1997) L’immigrazione straniera nell’area milanese, Rapporto Statistico dell’Osservatorio I.S.MU. Anno 1996, *Quaderni I.S.MU.*, 4.
- Blangiardo G.C. (2000), Sample design and implementation, Appendix: Methodological note on sampling technique, in *Push and pull factors of international migration. Country report – Italy*, Eurostat, 3/2000/E/n.5 Bruxelles: European Communities Printing Office, 16-22;107-117.
- Blangiardo G.C., (a cura di), (2002), L’immigrazione straniera in Lombardia, Rapporto statistico dell’Osservatorio Regionale per l’integrazione e la multietnicità 2001, Fondazione I.S.MU., Milano.

- Blangiardo G.C., (a cura di), (2003), L'immigrazione straniera in Lombardia, Rapporto statistico dell'Osservatorio Regionale per l'integrazione e la multiethnicità 2002, Fondazione I.S.MU., Milano.
- Byczkowski T.L., Levy M.S., Sweeney D.J. (1998) Estimation in sample surveys using frames with a many-to-many structure, *Survey Methodology*, 24, 1, 21-30.
- Cornelius W.A. (1982) Interviewing undocumented immigrants: methodological reflections based on fieldwork in Mexico and US, *International Migration Review*, 6, 2, 378-441.
- Goodman L.A. (1961) Snowball sampling, *Annals of Mathematical Statistics*, 32, 148-170.
- Haines D.E., Pollock K.H. (1998) Combining multiple frames to estimate population size and totals, *Survey Methodology*, 24, 1, 79-88.
- Hartley H.O. (1974) Multiple frame methodology and selected applications, *Sankhya*, ser. C., 36, 99-118.
- Kish L. (1987) *Statistical Design for Research*, Wiley, New York.
- Lohr S.L., Rao J.N.K. (2000) Inference from dual frame surveys, *Journal of the American Statistical Association*, 95, 271-280.
- Marasini D., Migliorati S. (2003) Small area estimation for immigrant people, submitted.
- Mecatti F. (2002a) La stima della media nel campionamento per centri, *Statistica*, 2, 285-297.
- Mecatti F. (2002b) Center Sampling and Dual Frame surveys: comparisons among estimators for the total, *Atti della XLI Riunione Scientifica SIS*, Milano, submitted to SMA-JISS.
- Mecatti F., Migliorati S. (2001) Center sampling: theory and estimation, *Tech. Rep. 01-06*, Dpt of Statistics, Pennsylvania State University.
- Mecatti F., Migliorati S. (2003a) Center Sampling: a strategy for elusive population surveys, submitted.
- Mecatti F., Migliorati S. (2003b) Confronti fra stimatori per la media nel campionamento per centri, *Statistica*, in print.
- Migliorati S. (1997) Alcune considerazioni sul campionamento per centri, *Statistica Applicata* 9, 3, 369-386.
- Migliorati S. (2002) New developments in center sampling, *Atti della XLI Riunione Scientifica SIS*, Milano, submitted to SMA-JISS.
- Migliorati S., Terzera L. (2002) Una proposta di stima della numerosità nel campionamento per centri, *Statistica*, LXII, 4, 737-753.
- Pratesi M., Rocco E. (2002) Centre sampling for estimating elusive population size, *Statistica*, LXII, 4.
- Pratesi M., Rocco E. (2003) Two-phase centre sampling for estimating elusive population size, in: *Atti del Convegno S.CO. 2003 "Modelli Complessi e Metodi computazionali Intensivi per la Stima e la Previsione"*, Treviso.
- Rao J.N.K. (2003) *Small Area Estimation*, Wiley, New York.
- Skinner C.J. (1991) On the efficiency of raking ratio estimation for multiple frame surveys, *Journal of the American Statistical Association*, 86, 779-784.
- Sudman S., Kalton G. (1986) New developments in the sampling of special populations, *Annual Review of Sociology*, 12, 401-429.
- Thompson, S. K. (2002) *Sampling*, Wiley, New York, 2nd Ed.