Journal of Official Statistics, Vol. 27, No. 3, 2011, pp. 451–465

Centre Sampling Technique in Foreign Migration Surveys: A Methodological Note

Gianluca Baio[1], *Gian Carlo Blangiardo*[2], *and Marta Blangiardo*[3]

We address in this article the problem of performing a statistical survey of a group of individuals present in a certain population when information about the complete list of the members is missing or partially unknown. This problem is particularly relevant in immigration analysis, where many of the individuals are possibly illegal migrants and therefore not formally registered or accounted for in official statistics. We propose a sample method that integrates information provided by specific surveys and subjective knowledge available to the researcher about the geo-social reality of interest.

Key words: Incomplete list; migration surveys; aggregation centres.

1. Introduction

We address in this article the problem faced by an researcher when sampling a number of individuals from a population for which a list of members is completely missing or partially unknown and therefore the application of standard statistical sampling methods becomes impractical.

Our motivating example is the analysis of the presence of immigrants in a particular area, where many of the individuals are possibly illegal migrants (that is, they are not formally registered or accounted for by official statistics). However, this problem might be encountered in a variety of situations in the social sciences when the interest lies in the estimation of hidden populations (Salganik and Heckathorn 2004): examples include research to assess the number of injection drug users (McKnight et al. 2006) or to estimate the population of homeless (Dávid and Snijders 2002).

Since the 1990 s several methods have been proposed to estimate stocks and flows of irregular immigration in Europe; in this article we recall the main attempts, but we refer the reader to the work of Jandl (2008) and Jandl et al. (2008) for a detailed review.

The first estimate of migration stocks was provided by the International Labour Office in 1991, assuming that the proportion of illegal immigrants in Europe was between 10 and

[1] Università di Milano Bicocca, Via Bicocca degli Arcimboldi, IT-20126 Milan, Italy and University College London, UK. Email: gianluca@statistica.it
[2] Università di Milano Bicocca, Via Bicocca degli Arcimboldi, IT-20126 Milan, Italy. Email: giancarlo.blangiardo@unimib.it
[3] Imperial College London, South Kensington Campus, London SW7 2A2, UK. Email: m.blangiardo@imperial.ac.uk

15 percent of the officially recorded resident foreign population, as documented in Clarke (2000). Similar studies were carried out by the International Centre for Migration Policy Development (Widgren 1994) and by the Committee on Migration, Refugees and Demography, presented at the Conference on the situation of illegal migrants in the Council of Europe Member States (Paris, 13 December 2001). These studies assume that the number of illegal foreign persons is a fixed percentage of the total foreign population present in the area under study.

A different methodology was implemented to estimate flows of illegal immigrants using data from apprehension statistics; for instance Heckmann and Wunderlich (2000) estimated that there were 400,000 immigrants illegally smuggled into Europe, assuming that for each person caught there are two who pass unhindered, and in 2001 the International Centre for Migration Policy Development estimated that there were 286,000 illegal immigrant entering the (then) 15 countries of the European Union using the total border apprehension rates.

Papademetriou (2005), combining data from stock and flows, assessed that unauthorized immigrants represented at least 1% of the population of the by then 25 countries of the European Union and suggested that the figure was growing at annual rates into mid-hundreds of thousands.

Despite the attempts to produce reliable estimates, no agreed standard is available, and the different assumptions used in the literature lead to not completely comparable results. The European Project CLANDESTINO (Jandl et al. 2008) represents a very important contribution in this field, as it reviews the state of the art on the topic of illegal immigration in Europe, making a key distinction between indirect and direct methods to estimate illegal immigration.

The first group includes estimates that use existing data (e.g. from census/registers) or administrative statistics. The latter group implies the use of the target population directly and can be classified into: i) multiplier methods, starting from the proposition that the size of the unknown total population can be directly calculated from the size of a known subtotal by use of an appropriately estimated multiplier, ii) Capture-Recapture sampling schemes, originally developed for estimating animal populations (Petersen 1896) and recently applied to the estimation of the number of illegal immigrants in the Netherlands (Van der Leun et al. 1998); and iii) survey methods, where information is obtained on a sample of immigrants and inference is extended to the entire target population.

Focusing on iii), a key issue is how to choose the sampling scheme that should ensure that the sample is representative of the target population, in view of the partial or total missingness of the list from which to extract the sample. The *snowball* sampling scheme, introduced by Goodman (1961) has been used to deal with this issue (Natale 1998), starting with a set of statistical units that bring further units into the sample from among their acquaintances. The main problem with this sampling scheme is that the final sample is not randomly selected and could thus lead to biased estimates.

To overcome these limitations, we propose here a sampling method based on an augmented set of information about a number of "aggregation centres" that the target population of immigrants regularly visit. Our sampling scheme allows us to weigh the original biased sample in order to provide a consistent estimate of the overall migrant population characteristics. The actual performance of this method has been empirically tested over the last decade in Italy.

The article is structured as follows. In Section 2 we present the general methodology, discussing the main assumptions behind it. In Section 3 we describe how these assumptions can be relaxed and how to practically compute the weights to be assigned to every sampled unit. Finally, in Section 4 we present a worked out example to estimate the main features of the Egyptian population living in the Milan area (Italy). Section 5 is a discussion of the main points developed in the article.

2. Framework of Analysis

The general situation can be described as follows. We consider a local area under investigation, and we assume that the universe of foreign citizens present at the time of the survey is made up of N units. Typically, the number N is not known. Moreover, we assume that each of these individuals entertains some relationship with K aggregation centres or gathering places located in the area. Some examples are institutions, places of worship, entertainment, care, meeting or similar. Notice that any register of foreigners attending courses, services, etc. or the official Population Register in a municipality, or province can be considered as a "centre".

Clearly, if the value of N and the complete list of the universe were known, it would be possible to randomly select n statistical units, for example using a Simple Random Sampling (SRS) scheme. In this case, the probability of inclusion in the sample is defined for each individual and is for each draw uniformly equal to $1/N$. The properties of this estimator are well known in the statistical literature.

Unfortunately, the value of N and the overall composition of the universe are typically unknown in the case of foreign migration surveys (specifically with reference to the overall population of legal and illegal immigrants). For this reason, we need to conceive of an alternative sampling scheme, yet maintaining the desirable inferential properties of the standard SRS method (particularly in the frequentist framework). To this aim, we propose a Centre Sampling (CS) scheme, which essentially amounts to the following steps:

1. Sample with replacement n out of the K reference centres;
2. For each draw, one statistical unit is randomly chosen among the individuals who regularly access the selected centre.

Obviously, for the CS the individual probability of inclusion in the sample depends i) directly on the number of centres that the individual actually attends and that are selected in the first stage and ii) inversely on the number of individuals in the population who are attached to each of those centres.

For any individual in the overall population, let us define the vector $\mathbf{u}(i) = [u_1(i), u_2(i), \ldots, u_K(i)]$, where

$$u_k(i) = \begin{cases} 1 & \text{if the } i\text{th unit has regular access to centre } k \\ 0 & \text{otherwise} \end{cases}$$

(Alternatively, we could consider "how much time" is spent in each centre. In this case, attendance can be formally expressed by a value T ($0 \leq T \leq 1$) proportional to the time spent in the centre.)

The vector $\mathbf{u}(i)$ characterises the *profile* of the ith individual with respect to the K centres. The CS individual probability of inclusion for the ith unit in the population can be calculated as

$$p(i) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{N_k} u_k(i) \tag{1}$$

where N_k is the total number of individuals in the population who entertain relationships with centre k.

As is obvious from (1), knowledge of the (components of the) profile $\mathbf{u}(i)$ is fundamental for the determination of this probability. Unfortunately, we are not able to know this profile *ex-ante*. Nevertheless, for each of the n units who entered the sample (and completed the survey) we can gather information on the centres he or she actually attends through a specific additional part of the questionnaire, so that the corresponding n vectors $\mathbf{u}(r)$, for $r = 1,2, \ldots, n$, can be obtained. The probability of inclusion in the sample can therefore be estimated *ex-post* for each of the n sampled individuals.

The idea behind the CS scheme is to devise a set of weights such that the weighted sample obtained from the CS procedure has the same representativeness as a hypothetical simple random sample stratified with respect to the distribution of the profiles of attendance at the centres for the N units. As is easy to see, the representativeness achieved in each local environment through the use of the CS technique is essentially equivalent to that obtained when:

i) The universe is stratified on the basis of the attendance at the K centres (that is the profiles defined by $\mathbf{u}$);
ii) The n units are chosen proportionally, randomly, and with replacement from the N_q units in each of the strata, with

$$\sum_{q}^{2^K - 1} N_q = N$$

2.1. Identification of the Weights

In order for the CS scheme to be meaningful, we need to impose some constraint on the weights. First, let us define $N(\mathbf{u})$, the number of individuals in the overall population who possess a given profile $\mathbf{u} = (u_1, u_2, \ldots, u_K)$, that is a sequence of 0's and 1's, in terms of centres regularly visited, and the corresponding population proportion

$$\pi(\mathbf{u}) := \frac{N(\mathbf{u})}{N}$$

The idea behind the CS is that the n sampled units, suitably weighted, should give a sample frequency distribution consistent with the population distribution of $\pi(\mathbf{u})$.

This constraint does hold if each sample unit that is associated with a profile $\mathbf{u}$ is weighted by a coefficient defined as the ratio

$$w(\mathbf{u}) := \frac{\pi(\mathbf{u})}{\hat{\pi}(\mathbf{u})} = \frac{N(\mathbf{u})/N}{n(\mathbf{u})/n} \tag{2}$$

where $n(\mathbf{u})$ is the number of sample units who possess profile $\mathbf{u}$ (and similarly $\hat{\pi}(\mathbf{u})$ is the sample proportion of such individuals).

Equation (2) requires the knowledge of the population proportion $\pi(\mathbf{u})$. But N and $N(\mathbf{u})$ are typically both unknown or not available, therefore this proportion must be estimated.

2.2. *Estimating the Proportion $\pi(u)$*

Suppose there are $N(\mathbf{u})$ units characterised by the given profile $\mathbf{u}$ in the population. Then, the probability of randomly selecting one individual possessing such a profile from those attached to the kth centre can be defined as

$$p_k(\mathbf{u}) = \begin{cases} N(\mathbf{u})/N_k & \text{if } u_k = 1 \\ 0 & \text{if } u_k = 0 \end{cases}$$

Hence, if n_k random and independent units that visit the kth centre are selected using a Bernoulli method, the corresponding expected number of statistical units possessing profile $\mathbf{u}$ in the kth centre is given by the expression.

In general, if we consider all the n units sampled in the K centres $\left(n = \sum_{k=1}^{K} n_k\right)$, the expected absolute frequency of the units with profile $\mathbf{u}$ is expressed by

$$\mathrm{E}[n(\mathbf{u})] = \sum_{k=1}^{K} n_k \frac{N(\mathbf{u})}{N_k} u_k \tag{3}$$

Consequently, the corresponding expected sample proportion is

$$\mathrm{E}[\hat{\pi}(\mathbf{u})] = \mathrm{E}\left[\frac{n(\mathbf{u})}{n}\right] = \sum_{k=1}^{K} \frac{n_k}{n} \frac{N(\mathbf{u})}{N_k} u_k$$

This quantity is not known, as it clearly depends on the unknowns $N(\mathbf{u})$ and N_k. However, it can be easily proven that

$$\mathrm{Var}\,[\hat{\pi}(\mathbf{u})] = \frac{1}{n^2} \sum_{k=1}^{K} n_k \frac{N(\mathbf{u})}{N_k} \left(1 - \frac{N(\mathbf{u})}{N_k}\right) u_k$$

which goes to 0 for sufficiently large n. In other words, if the sample is large enough, we can reasonably assume that the observed value of the sample proportion $n(\mathbf{u})/n$ can be used as a sensible (and convenient) estimation of its unknown expected value, that is

$$\frac{n(\mathbf{u})}{n} = \sum_{k=1}^{K} \frac{n_k}{n} \frac{N(\mathbf{u})}{N_k} u_k \tag{4}$$

Using (4) and setting $f_k = N_k/N$ for simplicity, we obtain

$$\hat{\pi}(\mathbf{u}) = \frac{n(\mathbf{u})}{n} = \frac{N(\mathbf{u})}{N}\sum_{k=1}^{K}\frac{n_k}{n}\frac{N}{N_k}u_k = \pi(\mathbf{u})\sum_{k=1}^{K}\frac{n_k/n}{f_k}u_k$$

from which we derive

$$\pi(\mathbf{u}) = \hat{\pi}(\mathbf{u}) \Bigg/ \sum_{k=1}^{K}\frac{(n_k/n)u_k}{f_k} \tag{5}$$

Consequently, knowing the total number of selected units n and the sample distribution of $n(\mathbf{u})$, and assuming as known (as a first approximation) the values of the f_k's — i.e. the relative frequencies with which the N units who form the population are distributed among the centres — the estimation provided in (5) leads to the specification of the weights in the following form

$$w(\mathbf{u}) = \frac{\pi(\mathbf{u})}{\hat{\pi}(\mathbf{u})} = \left(\sum_{k=1}^{K}\frac{(n_k/n)u_k}{f_k}\right)^{-1} \tag{6}$$

Notice that the weight is common for all individuals sharing the same profile $\mathbf{u}$.

2.3. Allocating the Sample Size into the K Centres to Compute the Weights

In summary, the CS scheme is based on the assumptions that the sample is large enough and that we know the relative importance (in terms of popularity/attendance) of each centre. If these assumptions hold, the selection technique for each of the n sample units amounts to the following two steps: a) random and independent selection (with replacement) of one of the K centres, with probability uniformly equal to $1/K$; and b) random and independent selection of one of the N_k units attending the selected centre, each with constant probability equal to $1/N_k$.

Accordingly, the number of units sampled in each centre is a binomial random variable

$$\Pr(n_k = s) = \frac{n!}{(n-s)!s!}\left(\frac{1}{K}\right)^{s}\left(\frac{K-1}{K}\right)^{(n-s)}$$

(for $s = 0,1,\ldots,n$), with:

$$\mathrm{E}[n_k] = \frac{n}{K} \quad \text{and} \quad \mathrm{Var}[n_k] = \frac{n(K-1)}{K^2}$$

The efficiency of this sampling technique can be increased if each centre is associated with a constant number of statistical units equal to n/K, or even better when the n sample units are divided among the K centres proportionally to the "attraction" each of the centres exerts on the population. In other words, using the criterion of direct proportionality with respect to the ratios $f_k = N_k/N$,

$$n_k = n\frac{f_k}{\sum_{k=1}^{K} f_k} \tag{7}$$

(notice that since individuals can be attached to more than one centre, in general $\sum_k N_k > N$).

Using this approach to allocate the sample units in each centre simplifies the computation of the weights $w(\mathbf{u})$. In fact, if (7) holds, then substituting into (6) and defining for simplicity $f^* := \sum_k f_k$, we have

$$w(\mathbf{u}) = \frac{f^*}{\sum_{k=1}^{K} u_k} \tag{8}$$

Consequently, by allocating the n sample units to the K centres proportionally to the values of the f_k's, the weights for each vector $\mathbf{u}$ vary only with the quantity $\sum_{k=1}^{K} u_k$, i.e. the number of non null elements in $\mathbf{u}$. In other words, under these assumptions, the only relevant variable for the determination of the weights is the number of centres attended by each sample unit subject to weighting.

3. Relaxing the Assumption of *Ex-ante* Knowledge of the f_k's

Equation (8) allows the researcher to estimate the weights for the CS procedure as functions of the unknown parameters f_k's. However, as already pointed out, the statistical information available for the population does not generally allow the evaluation of these parameters. In the following, we propose two strategies to overcome this problem.

3.1. Using Preliminary Importance Rates for the Centres

In summary, when the values of the f_k's are not available, the calculation of the weights $w(\mathbf{u})$ may be performed by means of the following steps

1. A preliminary "*importance rate*" q_k is attributed to each of the K centres in order to approximate (as closely as possible, also in the light of general knowledge of the immigrant group in that local area) the different unknown values of f_k. This step is in fact a naïve application of Bayesian principles to encode the (subjective) prior information available to the researcher.
2. The n sample units are distributed among the K centres according to the relationship

$$n_k = n\frac{q_k}{q^*} \tag{9}$$

with $q^* = \sum_{k=1}^{K} q_k$. This is effectively a counterpart of (7).
3. The computation of the weights $w(\mathbf{u})$ is based on the following:
   - Substituting (9) into (3), the expected frequency of units with a given profile is calculated by the expression

$$\mathrm{E}[n(\mathbf{u})] = \sum_{k=1}^{K} n\frac{q_k}{q^*}\frac{N(\mathbf{u})}{N_k}u_k$$

   - The value of q_k generally differs from the true (unknown) value of the corresponding f_k, the bias being quantified by a correction factor $d_k = q_k/f_k$.

Introducing the substitution, we then have:

$$\mathrm{E}[n(\mathbf{u})] = \sum_{k=1}^{K} n \frac{f_k d_k}{q^*} \frac{N(\mathbf{u})}{N_k} u_k$$

and since $f_k = N_k/N = \pi(\mathbf{u}) N_k / N(\mathbf{u})$

$$\mathrm{E}\left[\frac{n(\mathbf{u})}{n}\right] = \sum_{k=1}^{K} \frac{d_k}{q^*} \pi(\mathbf{u}) u_k$$

- If again we approximate the expected relative frequency of the units who possess profile $\mathbf{u}$ with the corresponding observed sample relative frequency $\hat{\pi}(\mathbf{u})$, we have

$$\hat{\pi}(\mathbf{u}) = \sum_{k=1}^{K} \frac{d_k}{q^*} \pi(\mathbf{u}) u_k = \pi(\mathbf{u}) \frac{1}{q^*} \sum_{k=1}^{K} d_k u_k$$

Now, assuming the size of the bias d_k is known (at least to a certain degree of approximation), we can estimate the quantity

$$\pi(\mathbf{u}) = \frac{\hat{\pi}(\mathbf{u}) q^*}{\sum_{k=1}^{K} d_k u_k}$$

using the sample information (and the knowledge of the d_k's), and the weights can be defined as

$$w(\mathbf{u}) = \frac{\pi(\mathbf{u})}{\hat{\pi}(\mathbf{u})} = \frac{q^*}{\sum_{k=1}^{K} d_k u_k} \tag{10}$$

whose specification requires only the knowledge of the values of the q_k's, which are fixed *ex-ante* by the researcher, and of the ratios $d_k = q_k / f_k$.

3.2. Using One of the Centres As Baseline

The method just shown is based on the assumption that the experimenter is able to define a set of preliminary values q_k's, as close as possible to the true value of the f_k's (i.e. is in the position of "controlling" the bias introduced). Since the latter values are unknown, this method is not the most efficient.

To overcome this disadvantage, we make use of the following procedure (Blangiardo 1996). Let us define, the number of units in the population who regularly visit both centres h and j, for $h,j = 1, \ldots, K$. Then, it is easy to show that

$$r_{hj} =: \frac{f_h}{f_j} = \frac{N_h/N}{N_j/N} = \frac{N_{hj}/N_j}{N_{hj}/N_h}$$

A suitable sample estimator for r_{hj} is given by

$$\hat{r}_{hj} =: \frac{n_{hj}/n_j}{n_{hj}/n_h}$$

where r_{hj} is the observed sample frequency of individuals who regularly visit both centres h and j. This estimator has the usual frequentist statistical properties of unbiasedness and consistency (cf. Migliorati 1997).

We can now modify (10) to correct the bias by means of the following scheme. First, we identify a *base* centre, b. This choice is based on the subjective knowledge that the researcher has about the set of centres available for the analysis (and again could be encoded in the form of a suitable probability distribution, i.e. under a more formal Bayesian approach). Each value of the q_k's is divided by q_b (the fixed preliminary importance rate for the base centre), giving a new series of values $\theta_k = q_k/q_b$, which express the relative importance of each centre with respect to the base centre. Next, we define

$$\delta_k =: \frac{\theta_k}{r_{kb}} = \frac{q_k/q_b}{f_k/f_b}$$

Obviously, this is an unknown quantity (as it is a function of the ratio r_{kb}). However, as suggested above, it can be entirely estimated by means of the information contained in the sample using $\hat{r}_{kb}$ instead of r_{kb}

$$\hat{\delta}_k = \frac{\theta_k}{\hat{r}_{kb}}$$

Some easy algebra shows that we can reexpress the values of the correction factors d_k as

$$d_k = \hat{\delta}_k \frac{q_b}{f_b} \tag{11}$$

and substituting (11) into (10), we are able to compute

$$\hat{w}_r(u) =: \frac{q^*}{\sum_{k=1}^{K} \hat{\delta}_k u_k} = w_r(u)\frac{q_b}{f_b}$$

for $r = 1,2,\ldots,n$. The weights $\hat{w}_r(u)$ are completely estimable by the sample data and are equivalent to the the weights $w_r(u)$, up to a constant factor q_b/f_b. This factor, which is common to all the weights, will be accounted for when the final adjustment of the n coefficients occurs, to guarantee that the condition

$$\sum_{r=1}^{n} w_r(u) = n$$

holds, i.e. in order to obtain the equivalence between the sum of all the weights in each survey area and the relevant sample size.

4. Example: Estimating the Main Characteristics of the Egyptian Population in Milan

In order to give a practical explanation of the steps required to use the CS method in a survey on migration, we present in this section an example of its application to estimate the main features of the Egyptian population in Milan (Blangiardo 2000). The first step was to consider a hypothetical universe in which each unit could be classified with respect to his/her relationship with a set of centres or gathering places, i.e. the following: mosque, copt-orthodox church, language centres, Egyptian restaurants, social-service centres, islamic butcher's shops, places of entertainment, kindergartens, and advisory offices.

Note that these centres have been conveniently identified by means of a preliminary analysis of the local environment and represent a collection of heterogeneous places that almost all the Egyptians in Milan are likely to have visited at least once. In order to further extend the coverage of the set of centres, the Population Register was also included. All the Egyptians registered on the date of the survey as stable residents were regarded as "visiting" the centre in question.

Consequently, the analysis considered a set of $K=10$ centres. The selection of a sample of $n = 307$ statistical units from the subsample of Egyptians living in Milan, and the subsequent construction of the weights to be assigned to each sample unit according to the methodology described above, was formalised according to the following procedure:

(a) On the basis of the *ex-ante* information about the attendance intensity of the 10 selected centres, and assuming the Population register as the base centre, it is first required to determine the values of the θ_k's, i.e. the preliminary estimates (in relative terms with respect to the base centre) of the unknown f_k's. Moreover, the corresponding values θ_k/θ^* must be computed, where $\theta^* = \sum_k \theta_k$. The number of units to be selected in each centre is consequently calculated as shown in Table 1.
(b) Each selected unit was asked to fill in an additional questionnaire about his/her attendance at all the reference centres, from which his/her corresponding attendance profile was built as described by the vector **u**. For instance, the information we obtained from the first eight units interviewed in the mosque (denoted as Centre 1) can be reported as in Table 2.
(c) At the end of the survey, the frequencies of the overall profiles of the n_k units interviewed in a centre were counted and the underlying ratios n_{hj}/n_j were computed, as shown in Table 3.
(d) With these premises and still considering (as a completely arbitrary choice) the *ex-ante* selection of centre number 8 (Population register) as the base centre, the

Table 1. A scheme of calculation of the number of units to be contacted and interviewed for each of the reference centres

Center	Code	Ex-ante value of $\theta_k = q_k/q_8$	$\theta_k/\theta^*(\%)$	Sample size $n\theta_k/\theta^* = nq_k/q^*$
C_1 = Mosque	1	0.20	9	28
C_2 = Copt-Orthodox church	2	0.15	6	18
C_3 = Language centres	3	0.15	6	18
C_4 = Egyptian restaurants	4	0.24	10	31
C_5 = Social-service centres	5	0.08	3	9
C_6 = Islamic butchers' shops	6	0.34	14	43
C_7 = Places of entertainment	7	0.06	3	9
C_8 = Population register	8	1.00	42	129
C_9 = Kindergartens	9	0.12	5	16
C_{10} = Advisory offices	10	0.05	2	6
		$\theta^* = 2.39$	100	$n = 307$

Table 2. Excerpt of the questionnaire used to gather information about the profile u *for the individuals in the sample*

		Profile of the centres visited by the interviewed subjects (1 if yes; 0 if no)									
Center	Unit	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
1	1	1	0	0	0	0	0	0	1	0	0
1	2	1	0	1	1	0	1	0	0	0	0
1	3	1	0	0	1	0	1	0	1	0	0
1	4	1	0	0	0	0	0	0	0	0	0
1	5	1	0	0	0	0	0	0	1	0	0
1	6	1	0	0	0	0	0	0	1	0	0
1	7	1	0	0	0	0	1	0	1	0	0
1	8	1	0	0	0	0	0	0	1	0	0
	...					...				...	
	...					...				...	

quantities $\hat{r}_{h8}$ were obtained, which are final estimates of the ratios $r_{h8} = f_h/f_8$ as shown in Table 4 (to avoid confusion with Table 3, we label each centre with l)

(e) Using these values, we were also able to compute the ratios $\hat{\delta}_h = \theta_h/r_{hb}$, as reported in Table 5.
Moreover, the corresponding coefficients can be computed. The values related to the first 8 units are calculated and shown in Table 6.

The procedure is obviously replicated for all the units interviewed and the sample is then weighted with the final coefficients of Table 6. The weighted sample can be considered as representative of the corresponding population and qualitative analysis could be conducted on it. The questionnaire included specific questions to investigate whether the sampled individuals were registered in the official Population register and whether they held a regular working visa. Consequently, it was possible to estimate the total population in the area and to specify the total number of illegal migrants by simply reproportioning the results.

Table 3. Proportion of units interviewed in centre h who declare that they regularly visit centre j as well

Code of centre h	Code of centre j 1	2	3	4	5	6	7	8	9	10
1	1.00	0.00	0.04	0.09	0.00	0.43	0.00	0.70	0.00	0.00
2	0.00	1.00	0.12	0.12	0.00	0.35	0.12	0.41	0.00	0.00
3	0.12	0.06	1.00	0.12	0.00	0.65	0.06	0.41	0.00	0.00
4	0.15	0.11	0.30	1.00	0.00	0.48	0.07	0.41	0.00	0.00
5	0.00	0.00	0.11	0.00	1.00	0.00	0.00	0.33	0.00	0.11
6	0.13	0.11	0.11	0.16	0.39	1.00	0.05	0.61	0.00	0.03
7	0.00	0.00	0.14	0.14	0.00	0.29	1.00	0.57	0.00	0.00
8	0.16	0.27	0.13	0.17	0.05	0.22	0.06	1.00	0.04	0.06
9	0.00	0.00	0.00	0.15	0.00	0.31	0.62	0.92	1.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.00	1.00

Table 4. Final estimates of the relative importance of each centre, with respect to the baseline

Centres codes (l)	Proportion of the units interviewed in centre 8 who also declared attendance for centre l (row 8 in the table above) (A)	Proportion of the units interviewed in centre l who also declared attendance for centre 8 (column 8 in the table above) (B)	Computed value of $\hat{r}_{l8} = A/B$
1	0.16	0.70	0.23
2	0.27	0.41	0.67
3	0.13	0.41	0.32
4	0.17	0.41	0.41
5	0.05	0.33	0.16
6	0.22	0.61	0.37
7	0.06	0.57	0.11
8	1.00	1.00	1.00
9	0.04	0.92	0.05
10	0.06	0.67	0.09

5. Discussion

In this article we have proposed a new methodology to deal with statistical surveys in the case where the complete list of the members of a target population is unknown. The sampling procedure consists in gathering additional information from a set of individuals, which is then used to build suitable weights to re-proportion the sample. The bias introduced by the sampling procedure can be then corrected, as shown above.

In particular, we assume that the researcher has information about a number of aggregation centres that are regularly visited by the immigrants. We have shown in this article that if the researcher can estimate the relative importance of each centre (possibly with respect to a baseline centre), it is possible to compute suitable weights associated with individuals sharing the same profile in terms of attendance at each of the relevant centres.

Table 5. The adjusted values for $\hat{\delta}_k$

Centres	θ_k	$\hat{r}_{h8}$	$\hat{\delta}_k = \theta_h/\hat{r}_{h8}$
C_1 = Mosque	0.20	0.23	0.89
C_2 = Copt-orthodox church	0.15	0.67	0.23
C_3 = Language centres	0.15	0.32	0.47
C_4 = Egyptian restaurants	0.24	0.41	0.58
C_5 = Social-service centres	0.08	0.16	0.50
C_6 = Islamic butchers' shops	0.34	0.37	0.92
C_7 = Places of entertainment	0.06	0.11	0.57
C_8 = Population register	1.00	1.00	1.00
C_9 = Kindergartens	0.12	0.05	2.40
C_{10} = Advisory offices	0.05	0.09	0.57
	$\theta^* = 2.39$		

Table 6. Attendance profiles and values of the weights for the first eight sampled units

Id	Profile **u** for the $K = 10$ centres										Nonadjusted weights	Final weights[a]
r	1	2	3	4	5	6	7	8	9	10	$\hat{w}_r(u)$	$w_r(u)$
1	1	0	0	0	0	0	0	1	0	0	1.2650	0.76
2	1	0	1	1	0	1	0	0	0	0	0.8371	0.50
3	0	0	0	1	0	1	0	0	0	0	1.5940	0.95
4	1	0	0	1	0	1	0	1	0	0	0.7053	0.42
5	0	1	0	0	0	0	0	1	0	0	1.9492	1.17
6	0	1	0	1	0	0	0	0	0	0	2.9691	1.78
7	0	1	1	0	0	1	1	0	0	0	1.0941	0.65
8	0	1	1	0	0	0	0	1	0	0	1.4118	0.84
. . .											. . .	. . .
. . .											. . .	. . .
$\hat{\delta}_k$	0.89	0.23	0.47	0.58	0.50	0.92	0.57	1.00	2.40	0.57	FTC	307

[a] For each case the nonadjusted weights are obtained by diving the value of $\theta^* = 2.39$ by the total of the product of the values in each row and the underlying values of $\hat{\delta}_k$, corresponding to the centre to which the column refers. For instance, the weight of the first case is given by: $1.2650 = 2.39/(0.89 \times 1 + 0.23 \times 0 + 0.47 \times 0 + \cdots + 1.00 \times 1 + 2.40 \times 0 + 0.57 \times 0)$. The results were then adjusted and written in the column of final total coefficients (FTC).

The methodology can be applied to any territorial unit. In the case of a large metropolitan area, the centres can be actual physical places (i.e. the *central* mosque). On the other hand, when more dispersed areas are considered (e.g. a whole region), the centres can represent categories (i.e. mosques).

This methodology has actually been used for the past ten years on real data collected by the ISMU Foundation (Milan) to estimate the stocks of immigration, with particular reference to (but not exclusively in) the Italian region of Lombardia. Reports at the regional as well as at the local municipality level are routinely produced by ISMU (Blangiardo et al. 2008; Cesareo 2009).

The choice of the centres selected for the analysis is obviously crucial, as they should have a sufficiently high degree of heterogeneity to include as many different immigrant life styles as possible. However, it seems reasonable to suppose that the researcher (or rather the team of researchers, possibly including statisticians, sociologists and demographers) might have sufficient knowledge about the specific area under investigation to make a sensible choice with respect to the number and characteristics of the selected centres.

An important assumption is that the main characteristics to be investigated (for instance, age, sex or other socio-demographic traits like legal immigration status) are represented in all their features in at least one of the centres included in the analysis. As an example, if the sample of individuals interviewed does not include people in the age group 20–34, then the weighted sample will be biased and it will not be able to produce reasonable inference on all the age groups (unless additional, external information is available).

On the other hand, if two researchers specify different sets of centres, but in each case the underlying characteristics of the target population are observed in the two samples that they derive, then, on average, the results of the centre sampling will be consistent.

Finally, possible developments of this work include the formal inclusion of prior information on the centres in the form of probability distributions in a Bayesian framework. This would allow inclusion of uncertainty on the relative importance of each centre and its propagation to the final estimation of the weights.

6. References

Blangiardo, G.C. (1996). Il campionamento per centri o ambienti di aggregazione nelle indagini sulla presenza straniera. In Studi in onore di G. Landenna. Milano (Italy): Giuffrè, 21–29. [In Italian]

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, European Printing Office, Brussels, Belgium, 107–117.

Blangiardo, G.C., Fasani, F., and Speciale, B. (2008). Consumption, Saving and Remittance Behaviour of Undocumented Migrants in Italy. In Sviluppo demografico ed economico nel mediterraneo, L. Di Comite et al. (eds). Bari, Italy: Cacucci Editore.

Cesareo, V. (2009). The 14th Italian Report on Migrations 2008. Monza, Italy: Polimetrica Publisher.

Clarke, J. (2000). The Problems of Evaluating Numbers of Illegal Migrants in the European Union. In Regularisations of Illegal Immigrants in the European Union, P. Bruycker (ed.). Brussels, Belgium: Bruylant, 13–22.

Dávid, B. and Snijders, T.A.B. (2002). Estimating the Size of Homeless Population in Budapest, Hungary. Quality and Quantity, 36, 291–303.

Goodman, L.A. (1961). Snowball Sampling. Annals of Mathematical Statistics, 32, 148–170.

Heckmann, F. and Wunderlich, T. (2000). Transatlantic Workshop on Human Smuggling. A Conference Report. Georgetown Immigration Law Journal, 15, 167–182.

Jandl, M. (2008). Methodologies for the Estimation of Stocks of Irregular Migrants. http://www.unece.org/stats/documents/2008.03.migration.htm. Accessed August 2010.

Jandl, M., Vogel, D., and Iglicka, K. (2008). Report on Methodological Issues. http://clandestino.eliamep.gr/wp-content/uploads/2009/10/clandestino_re%port-on-methodological-issues_final12.pdf. Accessed August 2010.

McKnight, C., Des Jarlais, D., Bramson, H., Tower, L., Abdul-Quader, A., Nemeth, C., and Heckathorn, D. (2006). Respondent-Driven Sampling in a Study of Drug Users in New York City: Notes from the Field. Journal of Urban Health, 83, 54–59.

Migliorati, S. (1997). Alcune considerazioni sul campionamento per centri. Statistica Applicata – Journal of the Italian Statistical Society, 3. [In Italian]

Natale, M. (1998). Determining Irregular Foreigners in the Italian Population. In La mesure de la migration clandestine en Europe, D. Delaunay and G. Tapinos (eds).: Eurostat Working Papers. Available at www.uni-mannheim.de/edz/pdf/eurostat/01/KS-AP-01-006-FR-I-FR.pdf.

Papademetriou, D.G. (2005). The Global Struggle with Illegal Migration: No End in Sight. http://www.migrationinformation.org/feature/display.cfm?ID=336. Accessed August 2010.

Petersen, C. (1896). The Yearly Immigration of Young Plaice into the Limfjord from the German Sea. Working Paper 6, Report of the Danish Biological Station to the Ministry of Fisheries.

Salganik, M. and Heckathorn, D.D. (2004). Sampling and Estimation in Hidden Populations using Respondent-Driven Sampling. Sociological Methodology, 34, 193–240.

Van der Leun, J.P., Engbersen, G., and Van der Heijden, P. (1998). Illegaliteit en criminaliteit. Schattingen, aanhoudingen en uitzettingen. Working paper, Rotterdam Erasmus University. [In Dutch]

Widgren, J. (1994). Multilateral Co-operation to Combat Trafficking in Migrants and the Role of International Organisations. 11th IOM Seminar on Migration, Geneva, Switzerland.

Received November 2009
Revised September 2010