Nota tecnica

(prodotto B)

nel quadro delle attività di:

"Armonizzazione con le rilevazioni degli anni 2001-2014 dei dati statistici ottenuti attraverso le rilevazioni sulla presenza straniera in Lombardia realizzate nell'ambito dell'Osservatorio Regionale per l'Integrazione e la Multietnicità (ORIM) negli anni 2015-2016",

attività svolta a integrazione

del contratto prot. RSF.2014.10584 del 29 settembre 2014 (CIG 57734555C8) per il servizio relativo alla progettazione, realizzazione ed elaborazione di tre indagini campionarie sulla popolazione immigrata in Lombardia nell'ambito delle attività dell'Osservatorio Regionale per l'integrazione e la multietnicità (ORIM) ***(Cod. SOC14003)"***

Il presente documento è stato redatto a cura del gruppo di lavoro all'uopo istituito presso CeSDES s.a.s. e coordinato da Marta Blangiardo, *Department of Epidemiology –Centre for Environment and Health*, Imperial College London (UK) con la partecipazione di Giorgia Papavero (Fondazione ISMU) e Alessio Menonna (Fondazione ISMU).

Sommario

Prima parte

1. Osservazioni preliminari per la conoscenza del materiale statistico ORIM

L'Osservatorio Regionale per l'Integrazione e la Multietnicità (ORIM) ha curato annualmente, sin dall'avvio della sua attività (2001), una rilevazione campionaria avente come universo di riferimento il complesso della popolazione straniera ultra 15enne proveniente dai c.d. "Paesi a forte pressione migratoria (PFPM)" e presente a qualunque titolo (regolare e non) in uno dei comuni della Regione Lombardia.

Tali indagini, svoltesi negli anni 2001-2016 con una numerosità campionaria che è variata da un minimo di 3500 a un massimo di 9000 unità annue per il complesso della regione, ha reso disponibile un materiale statistico formato da oltre 110mila casi individuali, per ognuno dei quali sono state acquisite numerose informazioni sui principali caratteri strutturali di tipo bio-demografico e socio-economico:

- Sesso
- Età
- Stato civile
- Istruzione
- Religione
- Reddito
- Consumi;

accanto a informazioni legate allo status e al progetto migratorio:

- Cittadinanza
- Condizione rispetto al soggiorno
- Posizione anagrafica
- Anzianità della presenza in Italia, in Lombardia e nella provincia di presenza
- Progetti per il futuro;

così come a numerose informazioni riguardanti le condizioni della presenza rispetto all'ambito:

- Familiare
- Abitativo
- Lavorativo.

Spesso tali informazioni sono state arricchite da approfondimenti specifici su temi che, di volta in volta, si ritenevano meritevoli di indagine (dall'accesso/uso di alcuni servizi essenziali, all'atteggiamento rispetto al voto, al sistema delle relazioni e reti amicali, alle abitudini di vita, ecc.).

***

Partendo dalla disponibilità del materiale statistico raccolto nel corso degli ultimi due anni, il lavoro che sta alla base del servizio in oggetto consiste nel renderlo fruibile in ambito Eupolis, in aggiunta a quanto già disponibile per gli anni 2001-2014, in una forma che si presti alla divulgazione, sia tra ricercatori che tra un più vasto pubblico interessato al tema.

A questo proposito, oltre a trattare il materiale al fine di assicurare omogeneità di contenuto nel tempo e chiarezza nelle modalità espositve, si è ritenuto utile fornire, anche in questa circostanza, elementi di accreditamento sul piano metodologico rispetto ai dati e alle procedure di acquisizione. Questo spiega quanto qui proposto al successivo punto 5, dove la nota sul campionamento – trasferibile su supporto (web) di accesso ai dati statistici nella sezione dei metadati – è stata redatta in lingua inglese anche in vista di una valorizzazione e di un uso del materiale ORIM in ambito internzionale. Non va infatti dimenticato che non esistono a livello europeo fonti statistiche, con micro dati sulla presenza straniera, che possano garantire un rigore scientifico e un ricchezza/consistenza campionaria comparabile a quanto offerto da ORIM in questi anni.

Ciò premesso, le attività svolte in questa sede da CeSDES al fine di rendere fruibili anche i dati statistici acquisiti con le rilevazioni ORIM 2015-2016 possono così riassumersi:

a) Verifica e controllo dei contenuti dei data base ORIM 2015-2016;

b) Piano di ricodifica delle variabili comuni orientato a dare comparabilità ai dati statistici;
c) Intervento di armonizzazione delle basi dati, ridefinizione del contenuto dei File SPSS; specifica dei pesi per il riporto all'universo e il calcolo della varianza campionaria delle stime;
d) Rilascio di una nota metodologica e di una guida alla lettura del materiale statistico con indicazioni per un suo corretto utilizzo (piano di campionamento, pesi di riporto all'universo, pesi di replicazione);
e) Rilascio dei materiali in forma di 2 archivi statistici in formato System File SPSS completi di metadati (variabili e value labels) e pesi, in versione interna e in versione accessibile a utenti esterni e divulgabili secondo l'approccio Open Data;
f) Fornitura della documentazione di supporto riguardante le rilevazioni (questionari anni 2015-2016)

2. Caratteristiche delle rilevazioni e organizzazione dell'indagine

Definizione del piano di campionamento

L'impostazione organizzativa e le scelte metodologiche ampiamente consolidate negli anni hanno basato l'acquisizione dei dati su un campone della popolazione target in grado di configurarsi come rappresentativo, pur con margini di errore/approssimazione diversi, tanto a livello regionale quanto per ogni singola provincia. Lo schema di campionamento ha costantemente previsto una preliminare assegnazione della N unità campionarie regionali alle province – distinguendo Milano città dal complesso di tutti gli altri comuni – sulla base di un criterio che ha sempre voluto garantire in ogni entità territoriale una soglia minima e un limite massimo di unità statistiche. Con riferimento agli estremi ci si è mossi da un massimo di 400 unità nella provincia di Sondrio e 3000 in quella di Milano, quando si poteva contare su un camprione regionale di 9000 unità (anni 2006-2009), a un minimo di 190 e 600, rispettivamente, quando si è potuto contare solo su una numerosità regionale di 3500 unità (Anno 2016).

Il totale di casi comunque assegnati a ogni ambito provinciale è stato quindi ripartito tra un opportuno campione di comuni identificati al suo interno con

appropriati criteri di rappresentatività, spesso anche rispetto alla lettura del territorio sulla base dei distretti socio-sanitari. Si sono così identificati da poco meno di 200 a oltre 400 comuni, a secodo degli anni, come unità campionarie di primo stadio entro cui procedere alla selezione del collettivo di stranieri da sottoporre a indagine (unità di secondo stadio) con una procedura di scelta probabilistiche nel rispetto delle regole del "campionamento per centri"[1].

A livello pratico ogni unità campionaria è stata intervistata – in forma diretta *face to face* – da personale specializzato, mediante la somministrazione di un questionario strutturato in quesiti a risposta chiusa riguardanti le sue principali caratteristiche, individuali, familiari e di contesto socio-economico (sesso, età, stato civile, cittadinanza, istruzione, religione, regolarità rispetto al soggiorno, residenza anagrafica, condizione abitativa, struttura familiare, attività economica, professione, reddito e consumi, ecc.). A quanto sopra si sono di cvolta in volta aggiunti, nel corso degli anni, questiti relativi a tematiche che hanno formato oggetto di approfondimento specifico.

Attraverso la riorganizzazione del materiale statistico acquisito con le successive indagini campionarie è stato possibili procedere alle consuete analisi sia delle singole realtà territoriali sia del complesso del panorama regionale. A tale proposito si è costantemente fatto uso del sistema di doppia ponderazione delle unità campionate, una procedura in grado di garantire, da un lato, la rappresentatività di ogni sub-campione provinciale nei riguardi del suo corrispondente universo e, dall'altro, il rispetto del peso relativo di ogni provincia entro il panorama regionale. In pratica, si è fatto in modo che ogni unità territoriale (le province e la città di Milano) potesse contribuire a determinare i risultati regionali con un apporto proporzionale alla sua effettiva quota di immigrati (valutata sul totale regionale) e non sulla base del numero di interviste realizzate al proprio interno.

[1] Riguardo alla metodologia in tema di campionamento per la scelta delle singole unità da intervistare si veda: Blangiardo G. C., "Il campionamento per centri o ambienti di aggregazione nelle indagini sulla presenza straniera", in Aa.Vv., *Studi in onore di G. Landenna*, Giuffrè, Milano, 1996; e Blangiardo G. C., "Campionamento per centri nelle indagini sulla presenza straniera in Lombardia: una nota metodologica", in Aa.Vv., *Studi in ricordo di Marco Martini*, Giuffrè, Milano, 2004. Una versione aggiornata è in: Baio G., Blangiardo G. C. e Blangiardo M., "Centre sampling thecnique in foreign migration surveys: a methodological note", in *Journal of Official Statistics*, vol. 27, 3, 2011, pp. 451-465.

Aspetti organizzativi e svolgimento della rilevazione

Sul piano organizzativo un aspetto determinante nel garantire il successo di una rilevazione sulla presenza straniera svolta mediante il campionamento per centri sono la scelta degli i intervistatori. L'esperienza empirica ha largamente dimostrato come sia importante mettere in campo uno staff capace sia di ottenere la fiducia degli immigrati nella fase di contatto, sia di portare a termine in modo completo la rilevazione risolvendo i problemi linguistici e di comunicazione.

Spesso si è visto che la soluzione ottimale consiste nell'impiego di intervistatori che siano essi stessi stranieri, meglio ancora se ben inseriti nel circuito delle comunità presenti sul territorio (tale prerogativa sembra vitale per i cinesi). Tenuto conto che il compito di ogni intervistatore è quello di: a) prendere contatto/raggiungere il centro presso cui deve operare; b) identificare in esso la popolazione target e procedere "casual,mente" alla scelta dei soggetti da intervistare; c) ottenere la collaborazione dell'intervistato per dare seguito alla somministrazione del questionario, è importante che egli abbia sia le necessarie facilitazioni per accedere in ognuno dei luoghi previsti, sia un'adeguata formazione per poter svolgere, nel rispetto delle regole di campionamento, le attività richieste.

Uno dei punti su cui va svolta una opera di sensibilizzazione degli intervistatori è l'esigenza di rilevare correttamente, prima di entrare nei contenuti del vero e proprio questionario, il profilo di afferenza ai diversi centri da parte di ogni intervistato. Ciò viene realizzato sottoponendo all'intervistato l'elenco di tutti i centro presi in esame ai fini dell'indagine, e chiedendogli di precisare *"quali di questi luoghi/centri sul territorio egli frequenta in questo periodo"* .

Resta inteso che, anche in presenza di un buon livello di organizzazione e di formazione, sussiste il problema della collaborazione alla rilevazione da parte della popolazione contattata nei centri. A titolo indicativo la quota di rifiuti varia abitualmente dal 10-15% nei casi migliori a punte del 40% in situazioni più problematiche. In generale, come mostra il contenuto del prospetto 1, il tasso di rifiuto ha risentito in modo significativo del tipo di

centro presso cui è avvenuta la rilevazione: la caduta è minima (10-15%) quando il contatto avviene entro spazi chiusi o con accordi preventivi (centri di formazione, visite presso abitazione a soggetti estratti dal registro della popolazione), è in vece massima (40-45%) quando il contatto avviene in luoghi aperti al pubblico (centri commerciali, mercati).

Prospetto 1 Tasso di rifiuti per luogo di rilevazione (*)

Luogo di rilevazione	*N° di interviste effettuate*	*% di rifiuti ricevuti*
Centri che offrono servizi e assistenza	2.820	23,6
Centri di formazione	515	15,2
Luoghi di culto	236	33,3
Negozi etnici	452	36,2
Luoghi di svago	793	34,7
Centri commerciali	243	40,6
Ritrovi, luoghi di incontro all'aperto	2.013	38,7
Mercati in genere	372	45,0
Luoghi di lavoro o di reclutamento forza lavoro	235	21,7
Associazioni e centri culturali	299	28,6
Centri servizi	437	35,1
Abitazione privata (Registro della popolazione)	546	12,5
*Totale**	*8.961*	*30,8*

Fonte: Fondazione ISMU, Indagine sulle integrazione della popolazione straniera presente in Italia, 2008-2009*Sono esclusi i contatti per i quali non è stato rilevato il numero dei rifiuti a rispondere

3. Indicazioni operative per l'utilizzo del data base

Il data base contenente i dati delle 2 rilevazioni campionarie ORIM 2015-2016 viene reso disponibile nel formato SPSS attravrso 2 distinfi files.

Ogni file contiene in ogni riga del foglio "*vista dati*" la sequenza delle modalità delle variabili acquisite attraverso il corrispondente questionariorisulrati per ognuna delle unità campionate.

Nello stesso foglio trovano altresì collocazione sia le variabili che derivano da operazioni di ricodifica (es. età in classi), sia le variabili che identificano il sstema di ponderazione. A tale proposito si fa presente che vengono forniti due tipi di pesi:

1) i così detti "pesi provinciali" (PESIPROV), tali da consentire elaborazioni ciscoscritte a un dato ambito territoriale assicurando cone totale delle frequenze la numerosità campionaria che ha caratterizzato l'ambito in oggetto nel corso della rilevazione;

2) i così detti "pesi regionali" (PESIREG), costruiti a partire dai precedenti ma con un ulteriore rirpoporzionamento capace di assegnare al totale delle frequenze di ogni ambito territoriale un peso relativo che rispecchia la corrispondente collocazione nel panorama regionale. Di fatto, l'impiego dei PESIREG è necessario ogni qualvolta si producano elaborazioni diferite all'intera regione o anche solo a due o più ambiti sub-regionali (province o la città di Milano).

L'insieme delle variabili considerate nel file viene poi dettagliato analiticamente nel foglio "*vista variabile*" entro il quale sono cnsultabili le etichette, delle variabili e delle corrispondenti modalità, con le quali il programma rende disponibili le elaborazioni statistiche.

4. Glossario

Anzianità della presenza = intervallo di tempo tra la data dell'indagine e l'anno indicato dall'intervistato come arrivo (in Italia, in Regione Lombardia, nella provincia in cui è stato rilevato come presente.

Centri = ai fini del campionamento per centri (CS) si intendono per centri o luoghi di aggregazione quelle istituzioni anche non formali (Enti, Associazioni, gruppi, ecc) e quegli ambienti che raggruppano popolazione immigrata in quanto indirizzati a rispondere alle molteplici esigenze di vita di coloro che sono presenti sul territorio

Irregolari = soggetti presenti in Italia senza un valido titolo di soggiorno, in quanto non più valido ovvero mai posseduto

Residenti stranieri = soggetti in possesso di una stabile dimora presso un comune italiano con relativa iscrizione nel Registro della popolazione residente (o Anagrafe)

Pesi campionari = valore da assegnare a ogni unità campionata nel conteggio delle frequenze in occasione delle elaborazioni statistiche. Il peso può divergere dall'unità ogni qualvolta, per esigenze di rappresentatività, i casi relativi ad alcune categorie vanno rivalutati (pesi superiori a 1) o ridimensionati (pesi inferiori a 1).

Paesi a forte pressione migratoria (PFPM) = Insieme formato dall'unione di tutti i c.d."Paesi in via di sviluppo" (Less developed countries LDc nella classificazione internazionale) e di quelli dell'Est Europa, comprensivo dei nuovi membri entrati nell'Unione Europea con i successivi allargamenti a partire dal 2004.

Stranieri regolari ma non residenti = soggetti presenti in Italia con un valido titolo di soggiorno, ma non iscritti (o non ancora iscritti) presso l'anagrafe di un comune italiano

Tasso di rifiuto = percentuale di soggetti contattati dall'intervistatore che non si sono resi disponibili partecipare all'intervista.

Seconda parte

5. Centre sampling method
Metodologia di campionamento per Centri

Introduction

In this methodological note we present the theoretical framework of the Centre Sampling (CS) method, a sampling technique which has been developed in the early '90s particularly in reference to the study of the immigrant population (Blangiardo 1991; 1993; 1996; 1999; Baio et Al., 2011) and implemented by ISMU Foundation.

The increasing importance of the immigration phenomenon in Italy required the development of new data collection techniques that would allow to monitor the phenomenon and produce the relevant information from a policy perspective. Indeed, official data sources on immigrations suffer from several limitations that ultimately compromise the validity of the produced estimates.

The main source of data on the immigrant population can be distinguished in two categories:

i) administrative data i.e. the records of the permits of residence issued by the Ministry of Interior, and the population registry (so called "Anagrafe");

ii) survey data, focused on a specific target population (e.g. Labour Force Survey, EU-Silc)[2]..

The first type of data can be used to estimate the stock and the flow of immigrants, allowing few disaggregation by basic demographic characteristics (typically regional distribution, citizenship, gender, age). The

[2]) And every ten years the additional contribution of Census data.

second one can provide more detailed information on aspects related to the topics investigated through the survey, and it is suitable for comparisons between the native and the foreign sub-populations. However, these data sources are characterized by the following disadvantages.

i) They take into consideration only regular migrants: by definition, administrative data includes only the sub-group of immigrants with a valid permit of residence, or registered in the population registry, and official surveys typically sample from the resident population, excluding those individuals (native and foreign) that are not listed in the population registry (e.g. homeless and illegal migrants). Given the characteristics of the migration phenomenon in Italy, where till last few years a not negligible share of immigrants were undocumented, estimates produced from these data sources cannot represent a valid description of the immigrant population.

ii) The level of details of official data is too limited, and general surveys typically do not focus on specific features of any sub-groups in the sample. It follows that the available data sources are not suitable to provide information neither on the specific characteristics and structural aspects of the migration phenomenon, nor on the life conditions and integration patterns of the immigrant population.

Given these observations, the CS technique is a valid alternative sampling method within the Italian contest, as it is able to target the entire foreign population, both legal and illegal, and to collect a wide set of information.

It is worth to notice that there are other estimation methods that can be implemented to provide information on the whole immigrant population, or to complement the existing data sources on regular migration with specific figures on stocks and flows of undocumented migrants. A detailed discussion of alternative methods can be found in the CLANDESTINO Project (Triandafyllidou, 2009) which reviews the state of the art on the topic of illegal immigration in Europe, critically describes the features and the critiques of the existing methods, and assesses the quality of the available estimates obtained with different techniques (Jandl et al., 2008).

The high reliability of the figures produced adopting the CS method, and its wide applicability on other "hidden populations", are the underlying motivations to present a detailed discussion on this technique.

The CS method enables to carry out a probabilistic survey even in the situation where the list of statistical units representing the universe of reference is missing or partially incomplete, as in the case of surveys targeting all migrants without regard to their juridical status. It overcomes this obstacle by exploiting daily life social interactions within the immigrant population. Indeed, the CS technique is based on the need of all migrants, legal and illegal, to attend at least one local aggregation centre for social contacts, health care, religion, leisure or simply for everyday needs.

The full list of aggregation centres (institutions, places of worship or entertainment, care centres, meeting points, shops, telephone centres, etc.) can be fairly easily mapped by an informed researcher. Once a sufficiently wide and heterogeneous set of centres is identified, it is possible to randomly identify a sample of centres, and then randomly choose a sample of immigrants among the attendees of the selected centre. Notice that the number of interviews in a certain centre will depend on its size of attenders: in principle the more they are and the larger will be the sub-sample of selected units.

It is important to underline that the final sample of immigrants obtained following this procedure is structurally biased and cannot be representative of the whole reference population. Indeed, the inclusion probability of immigrant *i* selected in centre *l* depends positively on the number of centre he/she visits, and negatively on the number of regular attendees of centre *l*.

However, the probability of inclusion can be obtained *ex-post*, by asking the interviewed which centre, among the set considered in the sample, they usually attend. These "profiles of attendance" will be used to compute a set of weighting coefficients that correct for the inclusion probability of each immigrant in the sample and that eliminate the original bias. Eventually the weighted sample has the same representativeness of a simple random sample (SRS) drawn proportionally from the distribution of attendance profile in the universe of reference.

Starting from the representative sample by CS method, the stock of immigrants can be estimated reconciling the CS weighted data with the population registry data, augmenting the number of registered immigrants by the proportion of sample respondents declaring (irrespective of their legal status) to not be listed in the population registry. Such stock can be further detailed, calculating - through the proportion provided by the CS sample - the amount of the irregular component.

Other than providing a valid estimate of the size of the immigrant population (regular and not), the second advantage of the data collection carried out using the CS method is that it is specifically customized in order to give a detailed picture of the migration phenomenon. Indeed, relative to official data sources, the survey contains more precise and detailed information on the demographic and socio-economic characteristics of the immigrant population. Moreover, it collects valuable and unique information on specific aspects related to the migration phenomenon, for instance on integration processes, paths in and out the illegal status, remittances behaviors, or perceived discrimination.

In the following of this note we describe the methodological framework of the CS technique, the practical aspects related to its implementations, and the applications of this method since the mid ‘90s.

Methodological framework

The main features of the CS techniques are:

i) the assumption that all individuals in the reference population visit at least one aggregation centre;

ii) the ex-post determination of the inclusion probability of each sampled individual.

Therefore, the CS method can be easily extended to other “hidden” or “hard-to-reach” populations, as long as the following basic assumptions are satisfied: the centres are known (or have been identified in previous

surveys), they are a finite number, they can overlap and they cover the entire target population.

We consider a given local area under investigation. The universe of foreign citizens present at the time of the survey is made of H statistical units (typically the number H is unknown). Each unit can be reasonably assumed to be connected with one or more 'centres' or 'aggregation places' within the study region (e.g. cultural or religious institutions, social networks, etc.). Consequently, once a sufficiently large set of centres has been identified, the universe can be formalised by means of a simple list (such as that depicted in Figure 1) or, alternatively, it is possible to describe the reference population using a contingency table that combines the list of Figure 1 with the information on the centres visited by the individuals (Figure 2).

Figure 1. A possible representation of the universe using a complete list

Sequence	*Individual details (name, address, ...)*
1	*a*
2	*b*
3	*c*
…	...
i	...
…	...
$H-1$	*y*
H	*z*

Figure 2. Representation of the universe in terms of a contingency table

Sequence	*Individual details*	*List of centres*[(a)]					
	W(*i*)	Centre 1	Centre 2	Centre 3	…	Centre $k-1$	Centre k
1	*a*	1	0	0	…	0	1
2	*b*	0	0	1	…	0	0
3	*c*	1	0	0	…	1	0
…	...	…	…	…	…	…	…
i	...	…	1	0	…	1	0
…	...	…	…	…	…	…	…
$H-1$	*w*	0	1	1	…	0	0
H	*z*	1	1	0	…	1	1
		Total H_1	Total. H_2	Total H_3	…	Total H_{k-1}	Total H_k

[(a)] For the i^{th} individual each column contains a '1' if they visit the centre, and a '0' otherwise. The column total indicates the number of statistical units who visit that centre

In practice, two methods can be envisaged to obtain a sample of N subjects out of the H available ones, which is representative of the entire population under study:

i) if a list like the one showed in Figure 1 is available, it is possible to randomly choose N names from it. In this way, we have a simple random sampling (SRS) scheme for which the typical estimator properties are well known;

ii) if the only available information is in the form of a list of centres visited by the immigrants, it will then be possible to randomly select N centres and then to randomly extract one individual out of the H_j that visit that specific centre (for j = 1, …, k).

The latter procedure can guarantee that the sample is representative of the population (exactly as the former would be) only if all the statistical units have the same probability of being selected in the sample. However, it is easy to see that the probability of inclusion of each individual in ii) is positively correlated with the number of centres visited and inversely correlated to the number of people who visit those same centres.

In particular, under the scheme of i) the probability that subject $w(i)$ is included in the SRS at each of the N draws (with repetition) is always equal to 1 / H, for any $w(i)$ and irrespective of the number of centres that they visits. On the contrary, under the scheme of ii) the same probability can be expressed as follows:

$$p(i) = \frac{1}{k} \sum_{j=1}^{k} \frac{1}{H(j)} u_j(i),$$

for each i = 1,2, …, H. This probability is a function of $u_j(i)$ which characterises the profile for subject i in terms of the centres they visit – i.e. u_j(i) = 1 if $w(i)$ visits centre j and 0 otherwise.

For each of the N units who entered the sample (and completed the survey) it is possible to gather additional information on the centres he or she actually attends, so that the correspondent n vectors $u(r)$ (for r = 1, 2, 3, …,

N) can be obtained (notice that here *i* indexes the individuals in the universe, made by *H* units, while *r* indexes the sample, made by *N* units). 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 $c(r)$, as functions of the vector $u(r)$, such that the sample obtained using the scheme in ii) has the same representativeness of a hypothetical SRS obtained under the scheme in i) but stratified with respect to the distribution of the profiles of attendance to the centres for the *H* units in the universe.

In conclusion, the representativeness achieved in each local environment through the use of the CS technique is essentially equivalent to that obtained when the universe is stratified on the basis of the attendance to the *k* centres (that is the profiles defined by *u*) and the *N* units to be interviewed are chosen proportionally and randomly (with replacement) from the total number of units in each of the $k^2 - 1$ (i.e. all the possible combinations of the *k* centres but the one in which the individual visits none) strata defined by $\sum_q N_q = N$.

Setting up of survey in each local area

a) Identifying the weights

According to the above discussion, in order to maintain the representativeness in each single local sample we need to determine a set of weights such that the N units, suitably weighted, provide a sampling frequency distribution that is consistent with the population distribution represented by $H(u) / H$. Here, $H(u)$ represents the number of individuals that present a profile described by the vector u, among the *H* units forming the overall population.

In effect, the weights are identical for all the $f(u)$ units in the sample that present the same profile u, and can be determined by the ratio

$$c(u) = \frac{H(u)/H}{f(u)/N},$$

defined as a function of the unknown quantity $H(u) / H$.

b) Estimating H(u) / H

First of all, we notice that if the universe is made by $H(u)$ units characterised by the given profile u in the population, then the probability of randomly selecting one individual possessing such profile among the H_j individuals who regularly visit the jth centre can be defined as $H(u) / H_j$, if the element u_j of the the vector u is equal to 1, and 0 otherwise.

If N_j random and independent units that visit the jth centre are selected using a Bernoulli method, the corresponding expected number of statistical units possessing profile u in the jth centre is given by the expression N_j [$H(u) / H$], for $j = 1, 2, \ldots, k$.

In general, if we consider all the N units sampled in the k centres (with $N = \sum_{j=1}^{k} N_j$), the expected absolute frequency of the units with profile u is expressed by:

$$\sum_{j=1}^{k} N_j [H(u)/H_j] u_j .$$

Consequently, the corresponding expected sample proportion is:

$$\sum_{j=1}^{k} \frac{N_j}{N} \frac{H(u)}{H_j} u_j .$$

If the sample is large enough, we can approximate this with the relative frequency $f(u)/N$, observed in the N sampled units:

$$\frac{f(u)}{N} \approx \sum_{j=1}^{k} \frac{N_j}{N} \frac{H(u)}{H_j} u_j .$$

Assuming that this latter condition holds and setting $p_j = H_j / H$ for simplicity, we obtain:

$$\frac{f(u)}{N} = \frac{H(u)}{H} \sum_{j=1}^{k} \frac{N_j}{N} \frac{H}{H_j} u_j = \frac{H(u)}{H} \sum_{j=1}^{k} \frac{N_j / N}{p_j} u_j$$

from which we derive:

$$\frac{H(u)}{H} = \frac{f(u)/N}{\sum_{j=1}^{k} u_j \frac{N_j / N}{p_j}}$$

Consequently, knowing the total number of selected unit N and the sample frequencies $f(u)$, and assuming as known (as a first approximation) the values of the p_j's — i.e. the relative frequencies with which the H units who form the population are distributed among each centre — this estimation procedure leads to the specification of the weights in the following form:

$$c(u) = \frac{H(u)}{H} / \frac{f(u)}{N} = \frac{1}{\sum_{j=1}^{k} u_j \frac{N_j / N}{p_j}} .$$

c) *Allocating the sample size into the k centres: impact on the computation of the weights*

As suggested earlier, according to the CS scheme, the selection technique for each of the N sample units amounts to the following two steps: *1)* random and independent selection (with replacement) of one of the k centres, with probability uniformly equal to $1 / k$; and *2)* random and independent selection of one of the H_j units attending the drawn centre, each with constant probability equal to $1/H_j$.

Accordingly, the number N_j of units sampled in each centre is a binomial random variable, taking the values 0, 1, 2, …, s, …, N with probability

$$\Pr(N_j = s) = \frac{N!}{(N-s)!s!}\left(\frac{1}{k}\right)^s\left(\frac{k-1}{k}\right)^{(N-s)}$$

with expected value and variance respectively equal to N / k and $[N(k-1)] / k^2$.

It is not too difficult to show that 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 centre exerts on the population. In other words, using the criterion of direct proportionality with respect to the ratios $p_j = H_j / H$:

$$N_j = N\frac{p_j}{\sum_{j=1}^{k} p_j}, \qquad \text{for } j = 1, 2, 3, \ldots, k.$$

It should be pointed out that the adoption of this criterion to allocate the sample units in each centre simplifies the computation of the weights $c(u)$. In fact, recalling that

$$c(u) = \frac{1}{\sum_{j=1}^{k} u_j \dfrac{N_j / N}{p_j}}$$

and introducing the notation

$$N_j = N\frac{p_j}{\sum_{j=1}^{k} p_j},$$

it will follow that

$$c(u) = \frac{1}{\sum_{j=1}^{k} \dfrac{u_j}{p_j}},$$

and, setting for simplicity $p^* = \sum_{j=1}^{k} p_j$, we obtain:

$$c(u) = \frac{p^*}{\sum_{j=1}^{k} u_j} .$$

Consequently, by allocating the N sample units to the k centres proportionally to the values of the p_j's , the weights for each vector u vary only with the quantity $\sum_{j=1}^{k} u_j$, i.e. the number of non-null elements in 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.

In the following Box 1, we show a worked example of the procedure so far.

BOX 1 Adjusted sample estimations for a generic quantitative attribute Y (e.g. the number of years since arrival in the resident country), assuming the values $p_j = H_j / H$ known.

Basic data

Universe (4 centres)					Sample size = 500		
	Distribution of profiles Centre ID				Relative frequency	Mean(*Y*) for profile P(*u*)	Grand mean
Profile P(*u*) [1]	**1**	**2**	**3**	**4**			
A	1	0	0	0	0.18	2.4	0.432
B	0	1	0	0	0.10	1.6	0.160
C	0	0	1	0	0.05	3.2	0.160
D	0	0	0	1	0.22	4.1	0.902
E	1	1	1	1	0.05	2.1	0.105
F	1	0	1	0	0.05	8.3	0.415
G	0	1	1	1	0.15	6.5	0.975
H	1	1	0	0	0.09	4.4	0.396
I	0	0	1	1	0.08	2.3	0.184
L	1	0	0	1	0.03	1.3	0.039
					1.00	Mean(*Y*)=	3.768
Coefficients p_j	0.40	0.39	0.38	0.53	p^*=1.70		

[1] For simplicity, we considered here a universe presenting only 10 of the $2^4 - 1$ possible type of contact with the 4 centres.

If, as stipulated here, the values of the p_j's are known, the sample is distributed among the 4 centres according to the condition $N_j = \frac{p_j}{\sum_{j=1}^{k} p_j}$, and

in this case we obtain the rounded values

$$N_1 = 118,\ N_2 = 115,\ N_3 = 112 \text{ and } N_4 = 156.$$

The computation of the coefficients $c(u)$ allows to obtain an unbiased estimation of the expected value for Y (recall that this is a quantitative variable of interest, measured on the surveyed individuals), as substantiated in the following table.

Using the weights to obtain an unbiased estimation of the average of Y

Profile P(u)	Expected structure of the sample **1**	**2**	**3**	**4**	Sample Frequency	mean(Y) for profile P(u)	Estimation of the overall mean	Weights	Adjusted sample frequency	Adjusted mean(Y)
A	53	0	0	0	53	2.4	127.1	1.70	90	216.0
B	0	29	0	0	29	1.6	47.1	1.70	50	80.0
C	0	0	15	0	15	3.2	47.1	1.70	25	80.0
D	0	0	0	65	65	4.1	265.3	1.70	110	451.0
E	15	15	15	15	59	2.1	123.5	0.43	25	52.5
F	15	0	15	0	29	8.3	244.1	0.85	25	207.5
G	0	44	44	44	132	6.5	860.3	0.57	75	487.5
H	26	26	0	0	53	4.4	232.9	0.85	45	198.0
I	0	0	24	24	47	2.3	108.2	0.85	40	92.0
L	9	0	0	9	18	1.3	22.9	0.85	15	19.5
Total	118	115	112	156	500		2078.5		500	1884.0
Sample mean (unadjusted mean of Y)							4.157			3.768

d) Relaxing the assumption on ex-ante knowledge of the p_j's

The statistical information available for the population does not generally allow the exact knowledge of the 'degree of importance' of each $p_j = H_j / H$ that characterises every centre with respect to the number of subjects

visiting it. Nevertheless, as suggested earlier, this parameter is fundamental in the estimation procedure for the CS weights.

The values of p_j, while not quantifiable in absolute terms given the available level of information, can be determined in relative terms. For instance, although it is not possible to express the values of p_l and p_m (that characterise the population frequency for two generic centres l and m), we can still estimate their ratio p_l / p_m (or equivalently we can work with the reciprocal p_m / p_l) by means of a procedure entirely based on the sampling information.

This procedure (presented in Blangiardo 1996) leads to estimators possessing the usual statistical properties of unbiasedness and consistency (Migliorati 1997). Consequently, it is possible to compute the final version of the weights $c(u)$ adjusting the bias introduced by the use of a set of fixed values, say q_j, considered known a priori (but subject to possible modifications in light of the information provided by the sample) instead of the unknown p_j's.

To sum up, in the absence of substantial information about the p_j's, the specification of the weights is by means of the following steps:

1. An arbitrary preliminary weight q_j is attributed to each of the k centres in order to approximate (as closely as possible, and also in the light of general information on that local area) the different unknown values of p_j.

2. The N sample units are distributed between the k centres according to the relationship:

$$N_j = N \frac{q_j}{q^*}, \text{ with } q^* = \sum_{j=1}^{k} q_j .$$

3. The computation of the weights $c(u)$ is then based on the following:

   - The expected frequency of units which show a given profile is calculated by the expression:

$$\sum_{j=1}^{k} \frac{Nq_j}{q^*} \frac{H(u)}{H_j} u_j \, .$$

- The value of q_j generally differs from the true (unknown) value of the correspondent p_j, the bias being quantified by a correction factor $d_j = q_j / p_j$. Introducing the substitution $q_j = p_j \, d_j$, we then have:

$$\sum_{j=1}^{k} \frac{Np_j d_j}{q^*} \frac{H(u)}{H_j} u_j$$

and since $p_j = H_j / H$, it will then follow that:

$$\sum_{j=1}^{k} \frac{Nd_j}{q^*} \frac{H(u)}{H} u_j \, .$$

- If again we approximate the expected relative frequency of the units who have the profile u with the corresponding observed sample relative frequency $f(u) / N$, we then have:

$$\frac{f(u)}{N} = \sum_{j=1}^{k} \frac{d_j}{q^*} \frac{H(u)}{H} u_j = \frac{H(u)}{H} \frac{1}{q^*} \sum_{j=1}^{k} d_j u_j \, .$$

Now, since we can estimate the quantity

$$\frac{H(u)}{H} = \frac{f(u) q^* / N}{\sum_{j=1}^{k} d_j u_j} \, ,$$

using the sample information, the weights can be then defined as:

$$c(u) = \frac{H(u)/H}{f(u)/N} = \frac{q^*}{\sum_{j=1}^{k} d_j u_j},$$

whose specification requires only the knowledge of the values of the q_j's, which are fixed *ex-ante* by the experimenter, and of the ratios $d_j = q_j / p_j$.

Since it is not possible to obtain a direct value for the p_j's, in order to compute these ratios a possibility is to define the relative importance of each centre with respect to a fixed 'baseline' centre, for instance one that is considered to be more reliable in terms of the data it provides. Let this centre be indexed by *b*: the relative importance is then defined as p_j / p_b, for a generic centre *j* (see Box 3 for a numerical example).

Using a similar rationale, we can compute the ratios q_j / q_b (using again the same centre *b* as baseline), and it is therefore possible to compute the ratios

$$d'_j = \frac{q_j / q_b}{p_j / p_b}$$

for which the following relations hold: $d'_j = (q_j / p_j)(p_b / q_b) = d_j (p_b / q_b)$.

Consequently, the weights can be determined for each of the *N* individuals in the sample as functions of the corresponding profile *u* and proportionally to a constant factor $B = q_b / p_b$, by means of the following relations:

$$c'(u) = \frac{q^*}{\sum_{j=1}^{k} d'_j u_j} = \frac{q^*}{\sum_{j=1}^{k} d_j u_j (p_b / q_b)} = \frac{(q_b / p_b) q^*}{\sum_{j=1}^{k} d_j u_j} = Bc(u) .$$

The factor *B*, that is common to all the weights, will be subject to adjustment during the final re-proportioning of the weights, in order to ensure that the sum of all the sampled units is equivalent to the chosen sample size.

<u>BOX 2</u> – Adjusted sample estimations for a generic quantitative attribute Y (e.g. the number of years since arrival in the resident country), assuming the values $p_j = H_j / H$ unknown.

(**Basic data – cfr. Box 1**)

Since the values of the p_j's are not known in this case, we choose suitable prior 'guesses' of the q_j's and we reproportion the sample size accordingly. For instance, if we let $q_1 = 0.8$; $q_2 = 0.5$; $q_3 = 0.6$ and $q_4 = 0.4$, since $N_j = N \frac{q_j}{\sum_{j=1}^{k} q_j}$, we can obtain the following sample sizes (suitable rounded, for the sake of simplicity): $N_1 = 174$; $N_2 = 109$; $N_3 = 130$ and $N_4 = 87$.
Under these premises, we can specify the expected structure of the sample as

	Expected structure of the sample				Sample frequency	mean(Y)	Overall mean estimation
Profile	**1**	**2**	**3**	**4**			
A	78	0	0	0	78	2.4	187.8
B	0	28	0	0	28	1.6	44.6
C	0	0	17	0	17	3.2	54.9
D	0	0	0	36	36	4.1	148.0
E	22	14	17	8	61	2.1	128.2
F	22	0	17	0	39	8.3	322.9
G	0	42	51	25	118	6.5	766.4
H	39	25	0	0	64	4.4	282.5
I	0	0	27	13	41	2.3	93.3
L	13	0	0	5	18	1.3	23.4
Total	174	109	130	87	500		2052.0
Sample mean – unadjusted estimation for mean(Y)							4.104

From this procedure, we obtain only an unadjusted estimation for the mean of *Y*.

However, we can correct the bias induced in this estimation using the following procedure:

1) Assuming that we are able to somehow provide a 'guess' for the ratios p_j / p_b (see Box 3 for a numerical example). If we use the 1st centre as the baseline b we then obtain the following values for p_j / p_1, ($j = 1, 2, 3, 4$): $p_1 / p_1 = 1.00$; $p_2 / p_1 = 0.98$; $p_3 / p_1 = 0.95$; and $p_4 / p_1 = 1.33$. We can now proceed to the computation of the ratios relative to the values of the coefficients q_j, that were preliminarily assigned in terms of 'draft values' as $q_1 = 0.8$; $q_2 = 0.5$; $q_3 = 0.6$; and

$q_4 = 0.4$. We then get the values: $q_1 / q_1 = 1.00$; $q_2 / q_1 = 0.63$; $q_3 / q_1 = 0.75$; and $q_4 / q_1 = 0.50$.

2) From the ratios $d'_j = \dfrac{q_j / q_1}{p_j / p_1}$, we get the values of d'_j necessary to the estimation of the coefficients $c'(u) = \dfrac{q^*}{\sum_{j=1}^{k} d'_j u_j}$, when in this case we get $q^* = \sum_{j=1}^{k} q_j = 2.3$.

In the following box 3 we present the values of the coefficients and how they can be applied to compute the mean of the variable *Y*.

Notice the impact of $B = (q_b / p_b) = (q_1 / p_1) = 0.8 / 0.4 = 2$: the sampling frequencies obtained are influenced by this scale factor. If increasing the total sample size accordingly (in this case from $N = 500$ to $N = 1000$) is not an option, we can still reproportion the actual weights (but notice also that this has no effect on the computation of the average value).

Moreover, it is worth noticing that by computing the *ex-post* value for *B* (recall that $B = \sum_{i=1}^{N} c'_i(u) \Big/ N$), it is always possible to compute the value of $p_b = B / q_b$ and, from the ratios of 1), those of the p_j's.

BOX 3 – Estimating the ratios p_l / p_m

Profile	Expected structure of the sample				Sample frequency	mean(Y) for profile P(*u*)	Weights	Adjusted sample frequency	Adjusted estimation for mean(Y)
P*(u)*	**1**	**2**	**3**	**4**					
A	78	0	0	0	78	2.4	2.30	180	432.00
B	0	28	0	0	28	1.6	3.59	100	160.00
C	0	0	17	0	17	3.2	2.91	50	160.00
D	0	0	0	36	36	4.1	6.10	220	902.00
E	22	14	17	8	61	2.1	0.82	50	105.00
F	22	0	17	0	39	8.3	1.29	50	415.00
G	0	42	51	25	118	6.5	1.27	150	975.00
H	39	25	0	0	64	4.4	1.40	90	396.00
I	0	0	27	13	41	2.3	1.97	80	184.00
L	13	0	0	5	18	1.3	1.67	30	39.00
	174	109	130	87	500			1000	3768.00
						Adjusted estimation of mean(Y)=			3.768

Starting from the sample data (in this specific case we have talked of the 'expected' structure of the sample), we can highlight the frequency of the subjects sampled in centre *l* that also declare to visiting sample *m*.

	Visiting *m* as well			
Sampled in *l*:	**1**	**2**	**3**	**4**
1	174	61	43	35
2	39	109	56	56
3	34	69	130	96
4	13	33	46	87

For example, in this case all the 174 individuals sampled in centre 1 obviously also state that they visit centre 1; 61 out of the 174 (in particular profiles E and H in Box 1) also visit centre 2. Moreover, 43 (profiles E and F) visit centres 1 and 3 and finally, 35 (profiles E and L) say that, besides centre 1, they are also connected with centre 4.

We indicate the number of people sampled in a given centre l and that are attached to centre m as well using the notation $N_{l,m|l}$. Now if we take the ratio of this number to the total number of individuals sampled from centre *l*

(N_l), we then obtain an estimation of the corresponding probability $H_{l,m} / H_l$ that a random individual having connections with l does have connections with m as well.

Similarly, the ratio of the number of people sampled in m that also visit centre l ($N_{l,m|m}$) to the number of people sampled in m (N_m) is an estimation of the probability $H_{l,m} / H_l$ that a random individual who visit m also does visit l.

These estimations can be easily computed starting from the following table of absolute frequencies suitably transformed into their relative counterpart.

Individuals sampled in:	Also visiting:			
	***m* = 1**	***m* = 2**	***m* = 3**	***m* = 4**
***l* = 1**	**1.00**	0.35	0.25	0.20
***l* = 2**	0.36	**1.00**	0.51	0.51
***l* = 3**	0.26	0.53	**1.00**	0.74
***l* = 4**	0.15	0.38	0.53	**1.00**

Bearing this in mind, it is possible to show that:

$$\frac{N_{l,m|l} / N_l}{N_{l,m|m} / Nm} = \frac{H_{l,m} / H_l}{H_{l,m} / H_m} = \frac{H_l}{H_m} = \frac{H_l / H}{H_m / H} = \frac{p_l}{p_m}.$$

The four ratios p_l / p_1 used in the example in Box 2 can be obtained as:

1. the ratio of $\frac{N_{1,1|1}}{N_1} = 1$ to $\frac{N_{1,1|1}}{N_1} = 1$ provides the obvious estimation of $p_1 / p_1 = 1$;

2. the ratio of $\frac{N_{1,2|1}}{N_1} = 0.35$ to $\frac{N_{1,2|2}}{N_2} = 0.36$ provides the estimation of $p_2 / p_1 = 0.98$;

3. the ratio of $\frac{N_{1,3|1}}{N_1} = 0.25$ to $\frac{N_{1,3|3}}{N_3} = 0.26$ provides the estimation of $p_3 / p_1 = 0.95$;

4. the ratio of $\frac{N_{1,4|1}}{N_1} = 0.20$ to $\frac{N_{1,4|4}}{N_4} = 0.15$ provides the estimation of $p_4 / p_1 = 1.33$.

It is easy to verify that the same procedure would lead to the construction of the ratios p_l / p_m using a different baseline centre, as showed in the following table.

p_l / p_m	Using as baseline			
	***m* = 1**	***m* = 2**	***m* = 3**	***m* = 4**
***l* = 1**	1,00	1,03	1,05	0,75
***l* = 2**	0,98	1,00	1,03	0,74
***l* = 3**	0,95	0,97	1,00	0,72
***l* = 4**	1,33	1,36	1,39	1,00

Moreover, irrespective on the baseline chosen, all the ratios represent adjusted estimations of the corresponding values, derived from the use of the 'true' parameters p_j's.

References

G.C.Blangiardo (1991), "*Appendice metodologica*", in: AA.VV, "*L'immigrazione extracomunitaria in Lombardia: il ruolo delle politiche regionali*", IReR-Regione Lombardia, Milano.

G.C.Blangiardo (1993), "*Un nuovo metodo di campionamento per le indagini sulla presenza straniera in Italia*", Quaderni del Dipartimento per lo Studio delle Società Mediterranee, Università di Bari, 3.

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

G.C.Blangiardo (1999), "*Methodological note*", in: AA.VV., *"Push and pull factors of international migration. Country report Italy"*, Eurostat, Luxembourg.

F. Mecatti. S. Migliorati (2001). "*Center sampling: theory and estimation*". Tech. Rep. 01-06. Dpt of Statistics. Pennsylvania State University.

G. C. Blangiardo. S. Migliorati e L. Terzera (2004). "*Center sampling: from applicative issue to methodological aspects*". Atti del Convegno della XLII Riunione Scientifica della Società Italiana di Statistica. Bari. 9-11 giugno 2004.

Baio G., Blangiardo G. C. e Blangiardo M. (2011), "*Centre sampling thecnique in foreign migration surveys: a methodological note*", in Journal of Official Statistics, vol 27, 3, pp. 451-465.

Triandafyllidou A. Ed. (2009), *Undocumented Migration: Counting the Uncountable. Data and Trends Across Europe*, http://www.gla.ac.uk/media/media_147171_en.pdf.

6. Surveys on migration through the Centre sampling method/ Indagini sulle migrazioni attraverso il metodo di campionamento per centri

Main framework

In the last 20-25 years, the Centre sampling method (CS) (Baio et Al., 2011) has been repeatedly applied in several surveys on the presence of foreign migrants, mainly in Italy. The first experiences in the early 1990s, mostly developed at the local level in Lombardia, have been instrumental to test the method on small samples and to validate the organisational aspects of the survey procedure. Later, the application of CS-based surveys has been extended to many new research projects, on several areas.

Among these, a first important example is the '*Indagine coordinata sulla presenza straniera in Italia*' (Coordinated survey on the foreign migrant presence in Italy, 1993-1994), which was developed by seven universities in the whole Italian territory. The method has been used in international research – e.g. '*Push and Pull factors of international migration – country report: Italy*', (Blangiardo 1999) in a comprehensive research developed by Eurostat and coordinated by the Netherland Interdisciplinary Demographic Institute (NIDI).

In the meantime, mostly thanks to the contribution of the Fondazione ISMU, the method was established within some local areas in the Lombardy Region of Italy: from the city of Milan (1996-2000), to its province (1997-2000), and to other provinces within the Region (Lodi, Mantua, Varese, Cremona, Lecco 2000-2001).

In 2001 the Regional Observatory for Integration and Multi-ethnicity was created in Lombardy, for which ISMU was given the responsibility of

management. Consequently, the application of the CS method was extended to all the 11 provinces (12 from 2005) and the sample size was increased to more than 8,000 units in the first five surveys (2001-2005), and has than overcome the 9,000 units in 2006-2009 (back to 8,000 in 2010-2011, 7,000 in 2012, 4,000 in 2013-2014 and 3,500 in 2015). Some more surveys were conducted in the period 2006-2008 on the foreign population in some provinces of Northern Italy (specifically Biella, Cuneo, Alessandria and Venice).

In the recent years, the CS scheme has also been applied in four large national surveys. The first, realised in 2005 for the Italian Ministry of Work and Social Policies, was based on a sample of more than 30,000 units and has covered 40 of the 103 Italian provinces (Blangiardo and Farina, 2006). The second in (which was started at the end of 2008 and was just concluded in 2009) involved 20 research groups in 33 different geographical areas with a sample of 12,000 units with the aim of gathering information on the theme of social integration (Cesareo and Blangiardo, 2009). The third survey in 2009, aimed at the analysis of the work paths in the migrant population, was based on a sample of more than 13,000 units, selected in 18 different geographical areas (ISMU, Censis and IPRS, 2010). The last survey, carried on in summer 2012, was designed to investigate both the households employing foreign workers in care activitiesemploying foreign and migrants employed in such activities. The size of CS sample was 1,500 for each sub sample and the universe was the whole country with specific attention to four regions of southern Italy.

Finally, we mention some recent experiences in which the CS scheme was used to produce representative samples in some specific European surveys: the 2007 survey developed by the European Commission and coordinated by the Department of Sociology of the University of Trento (particularly the research *Localmultidem* focused on Filipinos, Egyptian and Ecuadorians carried out in Milan by Ismu Foundation) and the Immigrant Citizens Survey, coordinated by King Baudouin Foundation & Migration Policy Group (for EU Commission) where CS method was used, other than in the Italian case, in the Portuguese and the Hungarian sample survey.

All in all, from the 1990s, almost forty surveys on the foreign migrants' universe (suitably defined according with the specific aims of the research) have been directly designed and developed using the CS scheme. The total number of statistical units included in those samples (summarised in the following table) is about 200,000.

Summary of the surveys developed using the CS scheme between 1991 and 2015

Subject and territorial reference	*No. sample*	*Year*
Foreign migrants living in Metropolitan area of Milan	500	1991
Foreign migrants living in Metropolitan area of Milan	500	1992
Foreign migrants living in the municipality of Monza	200	1992
Foreign migrants living in the municipality of Brescia	300	1992
National Academic Research Group on foreign migrants living		
Milan, Bologna, Ancon, Turin, Rome, Latina, Naples	3000 total	1993-1994
Foreign migrants living in Metropolitan area of Milan	1000	1996
Foreign migrants living in the province of Milan	2000 per year	1997-2000
Egyptians & Ghanaians in some Italian municipalities	1000	1997
Foreign migrants living in the province of Lodi	500 per year	1999 & 2001
Foreign migrants living in the province of Mantua	500 per year	2000 & 2001
Foreign migrants living in the province of Lecco	500 per year	2000 & 2001
Foreign migrants living in the province of Varese	500	2000
Foreign migrants living in the province of Cremona	500	2000
Foreign migrants living in Lombardy region	8000 per year	2001-05, 2010-
Foreign migrants living in Lombardy region	9000 per year	2006-2009
Foreign migrants living in Lombardy region	7000	2012
Foreign migrants living in Lombardy region	4000	2013-2014
Foreign migrants living in Lombardy region	3500	2015
Foreign migrants living in Lombardy region	3303	2016
Foreign migrants Southern and 10 provinces in Centre – Nord	30000	2005
Foreign migrants living in the province of Biella	500	2006
Foreign migrants living in the province of Cuneo	500	2007
Egyptians, Filipinos, Ecuadorians in Milan (Locamultidem	900	2007
Foreign migrants living in the province of Venice	800	2007
Foreign migrants living in the province of Alessandria	540	2008
Integration of foreign migrants living in Italy (Integration	12000	2008-2009
National sample of 18 local areas (provinces)	13000	2009
Immigrant Citizens Survey Italy, Portugal and Hungary	3000	2011
National Survey on household employing foreigners in care	1500	2012
National Survey on foreign migrants employed in care	1500	2012

How to survey the centres?

We present in this section some suggestions derived from our practical experience on how to design the survey in order to optimise the application of the CS strategy. The rules for the optimal application of the CS scheme are based on some fundamental points, which we review in the following.

a) Identifying the centres

Typically, the survey is preceded by a careful analysis of the relevant territory for the universe under study. If the researcher accounts for the environmental and socio-economic context, as well as for the conditions and behaviours that regulate the daily living of the target population, it is then possible to identify a given number of 'centres' to represent the main aggregation points of the individuals under study. In any case, it is fundamental that the set of the selected centres is sufficiently heterogeneous and such that every subject in the universe is, at least theoretically, reachable in at least another centre.

Moreover, the concept of 'centre' is not limited to physical places in which the population can be personally present: on the contrary, more 'formal' environments can be considered, such as for instance the list of members of a particular ethnic cultural association, church or religious group, members of unions or even the official population registry.

In practice, it is common to start the investigation by grouping the centres in macro categories, although they are then detailed as soon as specific information becomes available. The following list represents an example of possible categories of centres selected in some of the most recent surveys on foreign immigration in Italy (Fondazione ISMU 2009).

1. Centres offering services and assistance (first aid, job-centres, health clinics, canteens, public offices, …);
2. Development centres (language schools, professional development institutions, schools, university…);

3. Worship places (churches, mosques, temples, ...);
4. Ethnic shops (kebab shops, Halal butchers, …);
5. Entertainment places (cinema, clubs, gyms, bars, restaurants);
6. Shopping centres;
7. Open areas / aggregation points (stations, squares, parks, lakes, ...);
8. Markets (local markets, flower markets, farmers's markets);
9. Work places or job centres (construction sites; laboratories; restaurants and hotels; farms, …);
10. Cultural and social clubs
11. Services centres (phone centres, money transfer centres, ...);
12. Population registry.

In the actual development of the survey, it is necessary to proceed to a preliminary identification of the physical (or formal) places corresponding to each category. In the absence of suitable prior knowledge, the identification can be made by a pilot survey, specifically aimed at this aspect and possibly using known techniques such as 'snow ball sampling'. In other words, by means of the indication on 'which centres are actually visited' by the individuals sequentially interviewed, and starting from a limited set of centres, it is possible to extend this into a more comprehensive list of possibilities, until a sufficient heterogeneous range is reached.

When the geographic area under study is represented by a city or a metropolitan area, the map of centres provides their exact location, so that the relative assignment of the sample size and the organisation of the survey are not problematic (notice that in this case there is a single stage in the sampling procedure: the set of subjects is randomly selected by each of the centres identified for the analysis).

On the contrary, when the survey is characterised by a two stage process (where the first stage is the random selection of some local areas -e.g. towns, or municipalities- within the more general geographical entity under

study, e.g. a province or a region), the procedure becomes more complex and somehow less rigorous. In fact, each single first-stage unit can contribute to the survey only with the categories of centres that are actually present within its territory. Therefore, the sample units associated with each first stage unit are distributed among its centres proportionally to the 'higher level' distribution (i.e. the population frequencies of the province, or the region). This has the effect of accounting for the degree of attractiveness in the macro area for the category to which the centre belongs.

b) Setting the sample size

As suggested earlier, the way in which the units are associated with the set of centres varies with the type of sampling scheme (one vs two-stages). In the first case, our investigation suggests to assign the N sample units to the centres proportionally to the population frequency that is attached to each centre. However, this piece of information is rarely available (or reliable); nevertheless, it is possible to show that it suffices to use some prior estimations on the degree of 'overcrowding' in the different centres. For example, suppose we are concerned with four centres, say A, B, C and D. If we can reasonably assume that the population attached to A is half that attached to B, and that C and D are characterised by twice as many people as B (i.e. if A = 1, then B = 2 and C = D = 4), then a total initial sample size of N = 2200 will be divided among the four centres as 200 units in A, 400 units in B and 800 units in both C and D.

When the survey is conducted in a two-stage approach (the first stage being the random selection of a sample of local areas, suitable representative of the overall macro area), then the definition of the sample sizes is obtained in two steps: first, the N units are divided in the local areas selected as first-stage units, typically, it is sensible to partition the N units proportionally to the overall population under study. Then, the units associated with each local area are distributed among the centres in that area proportionally to their degree of importance, evaluated at the macro area level, for instance on the basis of the population attached with the category to which that centre belongs.

c) Organisational aspects

One of the main aspects in the organisation and actual development of any survey (and particularly so for the CS technique) is the choice and training of the interviewers. Empirical evidence has showed us that it is fundamental to use staff that is capable of gain the interviewees' trust in the first-contact stage, of completing the survey in all its aspects, taking care of all the linguistic and communication problems.

Often, the optimal solution is the use of foreign interviewers, better still if they are well in the loop of communities present in the relevant territory. In fact, the main role of the interviewer is to *a)* contact or directly visit the centre in which they need to operate; *b)* identify in the centre the target population and randomly select the subjects to be interviewed; *c)* obtain the collaboration of the interviewee and administer the questionnaire. In order to fulfil these requirements, it is important that the interviewer has the necessary facilities to get in each of the places they need to visit, an adequate education to develop, according to rules and regulations, the required activities.

The correct identification of the relationships between the interviewees and the centres is fundamental, and this should be stressed when training the interviewers. Pragmatically, this correctness can be ensured by showing the complete list of all the centres to the individuals who are then asked to specify '*which places / centres they are visiting / have been attached to lately*'.

Of course, even when the interviewers are well trained, there is still the issue of (partial or complete) non-response. As an indication, in our experience the non-response rate varies between 20% and 40%, in very complex situations. In general, as showed in the next table, the non-response rate depends closely on the type of centre in which the interview takes place: it is lowest (15%) when the contact is in closed spaces or after agreed appointments (eg interviews obtained in private homes), while it is highest (40-45%) when the contact is in public open spaces (such as shopping centres or markets).

Location of the interview	*Number of*	*% non-*
Centres offering services and assistance	2.820	23.6
Development centres	515	15.2
Worship places	236	33.3
Ethnic shops	452	36.2
Entertainment places	793	34.7
Shopping centres	243	40.6
Open areas / aggregation points	2.013	38.7
Markets	372	45.0
Work places or job centres	235	21.7
Cultural and social clubs	299	28.6
Services centres	437	35.1
Private home (drawn from the	546	12.5
*Total**	*8.961*	*30.8*

Source: Fondazione ISMU, Survey on foreign migrants present in Italy, 2008-2009

*Excludes the individuals for whom the total number of non-responses was not recorded

References

Baio G., Blangiardo G. C. e Blangiardo M. (2011), "*Centre sampling thecnique in foreign migration surveys: a methodological note*", in Journal of Official Statistics, vol 27, 3, pp. 451-465

G.C.Blangiardo (1999), "*Methodological note*", in: AA.VV., *"Push and pull factors of international migration. Country report Italy"*, Eurostat, Luxembourg.

G. C. Blangiardo P. Farina (ed.) (2006). "*Il Mezzogiorno dopo la grande regolarizzazione*", vol.3' Franco Angeli, Mulano.

V. Cesareo G.C. Blangiardo (ed.) (2009), "*Indici di intergrazione*", Franco Angeli, Milano.

ISMU, Censis, IPRS (2010), Immigrazione e Lavoro, Quadrni Ismu, 1/2010, Milano.