STATISTICA, anno LXII, n. 4, 2002

UNA PROPOSTA DI STIMA DELLA NUMEROSITÀ NEL CAMPIONAMENTO PER CENTRI

S. Migliorati, L. Terzera

1. INTRODUZIONE

La numerosità di alcune popolazioni non è sempre facilmente determinabile. Spesso è la natura stessa delle unità statistiche o l'assenza di liste complete ad impedirne la corretta enumerazione ovvero l'individuazione di una stima. In particolare, nell'ambito dello studio della presenza straniera in Italia, l'assenza di una lista completa ed esaustiva comporta una sottostima della numerosità, dovuta in modo particolare alla componente irregolare. E, d'altro canto, in tale contesto non è neppure possibile l'enumerazione essendo la popolazione di riferimento parzialmente elusiva.

Al fine di realizzare indagini sulla popolazione immigrata presente su specifiche aree, è stata proposta la tecnica di campionamento per centri (Blangiardo, 1996) che sfrutta la necessità degli stranieri di frequentare almeno un luogo di aggregazione di qualsivoglia natura (centri di culto, di svago, di servizi o altro) ben identificabile nell'area territoriale di riferimento. Tale tecnica presuppone la conoscenza dei centri frequentati dagli stranieri, pur risultando spesso ignota la numerosità dei frequentanti. Infatti, mentre in alcuni centri sono disponibili vere e proprie liste (per esempio, lista anagrafica, scuole di lingua, centri medici) o vi è la possibilità di una enumerazione dei presenti (per esempio, centri di assistenza con numero fisso di posti letto), in altri punti di aggregazione si è di fronte alla completa disinformazione (per esempio, *phone-center*, parchi, piazze).

Un'ulteriore complicazione deriva dal fatto che i centri si possono sovrapporre, nel senso che l'appartenenza di un soggetto ad un centro non è esclusiva e d'altro canto risulta ignota la numerosità delle unità che afferiscono a più centri.

La tecnica del campionamento per centri prevede la costruzione di un campione di ampiezza n attraverso l'estrazione di campioni casuali semplici in tutti i centri individuati, con numerosità campionaria pari ad n_l e tale che $\sum_{l=1}^{L} n_l = n$, dove L, noto, è il numero di centri sul territorio. Merita di essere sottolineato che

Ricerca finanziata tramite MURST COFIN99 *Metodi di Inferenza Statistica per Problemi Complessi.*

è la natura stessa di alcune popolazioni (tra cui quella immigrata, ma si pensi per esempio anche ai consumatori di un certo prodotto) a rendere il campionamento per centri strumento essenziale per reperire informazioni sulla popolazione oggetto d'attenzione o, comunque, quello che meglio risponde a esigenze di rappresentatività.

Benché numerose soluzioni al problema della stima della numerosità di una popolazione siano state proposte in letteratura, nessuna di queste sembra potersi adattare al caso dei centri. Ad esempio, gli innumerevoli lavori relativi ai *multiple frames*, ovvero alla presenza di due o più liste incomplete (Hartley, 1974; Skinner, 1991; Byczkowski *et al.*, 1998), presuppongono note le numerosità dei frequentanti i *frames* (assimilabili ai centri) ovvero le sovrapposizioni tra *frames* differenti; tuttavia, tali quantità sono ignote nel caso del campionamento per centri. Altrettanto può dirsi per le tecniche basate sull'unione di *list frames* e di *area frames* (Haines e Pollock, 1998) nonché per la tecnica nota come *network sampling* (Thompson, 1992). In generale, inoltre, in tutti gli approcci citati si prevede di estrarre un campione di *frame* e di osservare poi per i *frames* estratti tutte le unità esaustivamente; al contrario lo schema del campionamento per centri prevede di estrarre un campione da tutti i centri individuati, e tale scelta è da ricondursi al fatto che la natura di alcuni centri non li rende assimilabili a liste incomplete e, pertanto, ne preclude ogni possibilità di rilevazione esaustiva.

Nel presente lavoro si propone una stima per la numerosità della popolazione basata sull'approccio della verosimiglianza. Innanzitutto si ipotizza, per semplicità, la totale *detectability*, (Thompson, 1992, 1994), ovvero si suppone che l'estrazione delle unità dai centri avvenga nel momento di massimo affollamento, così che si possa ritenere che tutte le unità siano presenti al momento del campionamento. Sotto tale ipotesi viene individuata una stima di N nel caso dell'esistenza di due soli centri di aggregazione (paragrafo 2); il metodo è poi esteso a più centri (paragrafo 3). L'ipotesi, poco realistica, di completa *detectability* viene in seguito rimossa e per ovviare a tale mancanza si introduce un'opportuna famiglia di leggi di distribuzione a priori per le *detectability* (paragrafo 4). Pertanto, si propone una nuova stima di N basata sulla verosimiglianza integrata: il caso di due centri è l'oggetto del paragrafo 5, mentre nel paragrafo 6 viene fornita una sua generalizzazione al caso di più centri. Infine, nel paragrafo 7 si affronta uno studio simulativo basato su dati reali e sono discusse le *performances* degli stimatori proposti.

2. COMPLETA *DETECTABILITY*: STIMA DI N NEL CASO DI DUE CENTRI

Si consideri il caso dell'esistenza di due soli centri (C_1 e C_2) frequentati da una popolazione di numerosità ignota N. Sia N_l la numerosità di C_l ($l = 1, 2$) e, posto che ogni unità statistica può frequentare sia C_1 sia C_2, risulta $N_1 + N_2 \geq N$.

A ciascuno degli N soggetti è associato un cosiddetto "profilo" che informa circa la struttura di frequentazione dei centri essendo un vettore di dimensione pari al numero dei centri e composto da 0 e 1: 0 qualora il soggetto non frequenti il

centro e 1 altrimenti. Nel caso di due centri è possibile identificare tre profili: il primo ($\mathbf{u}_1$ = [1 0]) che caratterizza coloro che frequentano solo C_1, il secondo ($\mathbf{u}_2$ = [0 1]) è relativo a soggetti presenti solo in C_2 ed infine l'ultimo profilo ($\mathbf{u}_3$ = [1 1]) si riferisce all'insieme di unità che possono essere individuate in entrambi i centri. Ragionevolmente non viene previsto un quarto profilo, poiché si ipotizza che tutti i soggetti frequentino almeno un centro. Indicato con $N_{\mathbf{u}_r}$ il numero di soggetti con profilo $\mathbf{u}_r$ (r = 1,2,3), ne segue che:

$$N_1 = N_{\mathbf{u}_1} + N_{\mathbf{u}_3}\,, \qquad N_2 = N_{\mathbf{u}_2} + N_{\mathbf{u}_3}\,, \qquad \sum_{r=1}^{3} N_{\mathbf{u}_r} = N\,.$$

Poiché ogni individuo è caratterizzato da uno ed un solo profilo, risulta evidente che è possibile ottenere una stima per N tramite quella delle numerosità dei profili, tenendo presente che, in generale, risultano ignote tutte le numerosità sopra introdotte.

Il campionamento per centri prevede l'estrazione di un campione casuale semplice di n_1 unità statistiche da C_1 e di n_2 unità da C_2. La rilevazione campionaria fornisce, quindi, le quantità $f_{\mathbf{u}_r}$ che rappresentano le frequenze campionarie di individui con profilo $\mathbf{u}_r$.

In situazione di completa assenza informativa circa tutte le numerosità considerate $(N, N_1, N_2, N_{\mathbf{u}_1}, N_{\mathbf{u}_2}, N_{\mathbf{u}_3})$ la funzione di verosimiglianza non consente di pervenire ad una stima dei valori assoluti delle medesime in quanto una delle equazioni di verosimiglianza è combinazione lineare delle rimanenti: risulta al più possibile individuare una stima per le numerosità relative, ovvero rapportate ad N, il che preclude la possibilità di pervenire ad una stima per la numerosità totale N che rappresenta, tuttavia, il principale oggetto d'interesse.

Per ovviare a tale problema, è realistico supporre che uno dei due centri sia una lista incompleta così che la sua numerosità possa ritenersi nota (è possibile identificare il centro con una lista come, per esempio, la lista anagrafica, gli iscritti ad una scuola di lingue e così via).

Nel caso in esame si ipotizza nota la numerosità N_1 delle unità frequentanti il centro C_1, ma poiché ciò non significa che lo siano anche le numerosità dei profili di tali unità, risulta innanzitutto opportuno determinare le stime $\hat{N}_{\mathbf{u}_1}$ e $\hat{N}_{\mathbf{u}_3}$ ottenute massimizzando la verosimiglianza relativa alle prove effettuate nel primo centro e pervenendo a:

$$\hat{N}_{\mathbf{u}_1} = N_1 f_{\mathbf{u}_1} / n_1; \quad \hat{N}_{\mathbf{u}_3} = N_1 f_{\mathbf{u}_3,1} / n_1 \tag{1}$$

dove con $f_{\mathbf{u}_r,l}$ si intende la frequenza campionaria del r-esimo profilo ottenuta dal campione estratto dal centro C_l $(r = 1,2,3;\ l = 1,2)$.

Si noti che gli stimatori che ne derivano risultano essere corretti e consistenti per le corrispondenti numerosità di profilo.

Per quanto riguarda la quantità $N_{\mathbf{u}_2}$, la corrispondente stima è deducibile dai dati campionari ottenuti nel secondo centro unitamente alle informazioni sul profilo $\mathbf{u}_3$ tratte da C_1. Così procedendo, la funzione di verosimiglianza, una volta sostituita a $N_{\mathbf{u}_3}$ la rispettiva stima fornita con la (1), risulta proporzionale alla quantità:

$$L\left(N_{\mathbf{u}_2}\middle|\hat{N}_{\mathbf{u}_3}, f_{\mathbf{u}_2}\right)\propto\left(\frac{N_{\mathbf{u}_2}}{N_{\mathbf{u}_2}+\hat{N}_{\mathbf{u}_3}}\right)^{f_{\mathbf{u}_2}}\left(\frac{\hat{N}_{\mathbf{u}_3}}{N_{\mathbf{u}_2}+\hat{N}_{\mathbf{u}_3}}\right)^{n_2-f_{\mathbf{u}_2}} \tag{2}$$

e massimizzando rispetto a $N_{\mathbf{u}_2}$ si perviene a:

$$\hat{N}_{\mathbf{u}_2}=\hat{N}_{\mathbf{u}_3}\frac{f_{\mathbf{u}_2}}{n_2-f_{\mathbf{u}_2}} \tag{3}$$

così che, ricorrendo alla (1) e alla (3), la stima per la numerosità N risulta:

$$\hat{N}=\hat{N}_{\mathbf{u}_1}+\hat{N}_{\mathbf{u}_2}+\hat{N}_{\mathbf{u}_3}=\frac{N_1}{n_1}\left[f_{\mathbf{u}_1}+f_{\mathbf{u}_3,1}\left(1+\frac{f_{\mathbf{u}_2}}{n_2-f_{\mathbf{u}_2}}\right)\right] \tag{4}$$

Si noti che le stime così ottenute sono sempre positive, rispettano le relazioni d'ordine esistenti tra le numerosità e paiono "intuitivamente" ragionevoli dato che ben sfruttano l'informazione riguardante la numerosità delle unità statistiche che frequentano entrambi i centri.

3. GENERALIZZAZIONE AL CASO DI TRE O PIÙ CENTRI

Nella realtà il caso di due soli centri si verifica raramente, è quindi necessario a questo punto aggiungere un terzo punto di aggregazione e successivamente cercare una generalizzazione a L centri.

Siano quindi C_1 C_2 e C_3 i centri, di numerosità N_1 N_2 ed N_3, così che le unità statistiche possono presentare uno tra sette profili possibili escludendo, anche in questo caso, l'eventualità di soggetti isolati non rintracciabili in alcun centro, ovvero il profilo costituito da soli "0". Nella Figura 1 sono rappresentati graficamente tutti i profili esistenti con tre punti di aggregazione.

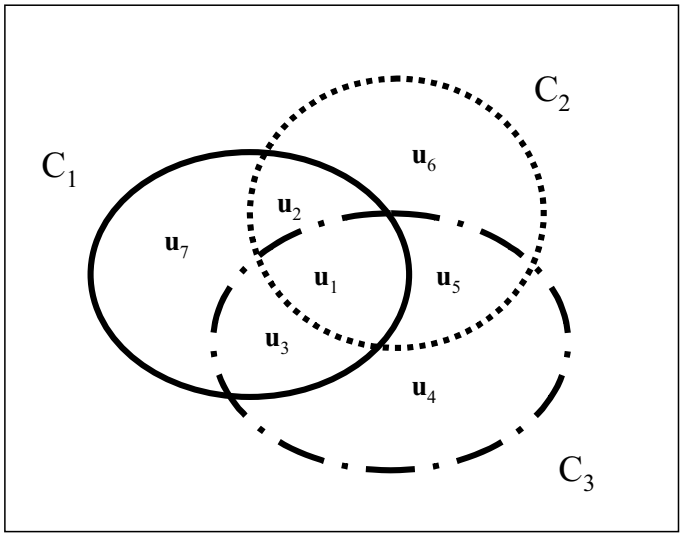

Figura 1 – Rappresentazione dei profili in presenza di 3 centri.

Anche in questo caso risulta possibile individuare una stima di N basandosi sulla semplice ipotesi che un centro, ad esempio il primo, sia una lista incompleta e tenendo presente la relazione: $N = \sum_{r=1}^{7} N_{\mathbf{u}_r}$.

Pertanto, posto N_1 noto a priori e procedendo in modo analogo a quanto fatto nel caso di due soli centri, la massimizzazione della funzione di verosimiglianza relativa al solo centro C_1 consente di pervenire ad una stima (corretta e consistente) delle numerosità dei profili che prevedono la frequentazione del centro considerato, ovvero:

$$\hat{N}_{\mathbf{u}_r} = N_1 \frac{f_{\mathbf{u}_r,1}}{n_1}, \qquad r = 1,2,3,7 \tag{5}$$

dove, ancora una volta, con $f_{\mathbf{u}_r,1}$ si intende la frequenza campionaria del r-esimo profilo relativa al campione estratto da C_1.

La stima delle numerosità dei restanti profili può essere ottenuta utilizzando congiuntamente le informazioni campionarie relative ai centri C_2 e C_3 e le stime già individuate con la (5) per i profili che, oltre alla frequentazione del centro 1, prevedono anche quella di almeno uno dei due centri rimanenti (ovvero i profili $\mathbf{u}_1$, $\mathbf{u}_2$ e $\mathbf{u}_3$).

Tenendo pertanto presente che $N - N_{\mathbf{u}_7} = N_{\mathbf{u}_1} + N_{\mathbf{u}_2} + N_{\mathbf{u}_3} + N_{\mathbf{u}_4} + N_{\mathbf{u}_5} + N_{\mathbf{u}_6}$, la funzione di verosimiglianza, una volta sostituita la (5) alle corrispondenti numerosità di profilo ignote, può scriversi:

$$L\left(N_{\mathbf{u}_r}\left|\hat{N}_{\mathbf{u}_1},\hat{N}_{\mathbf{u}_2},\hat{N}_{\mathbf{u}_3},f_{\mathbf{u}_r};\ r=4,\ldots,6\right.\right)\propto\prod_{r=4}^{6}\left(N_{\mathbf{u}_r}\right)^{f_{\mathbf{u}_r}}\left(\frac{1}{\hat{N}_{\mathbf{u}_1}+\hat{N}_{\mathbf{u}_2}+\hat{N}_{\mathbf{u}_3}+\sum_{r=4}^{6}N_{\mathbf{u}_r}}\right)^{n_2+n_3}$$

È immediato verificare che le stime di massima verosimiglianza per le tre numerosità di profilo $N_{\mathbf{u}_4}, N_{\mathbf{u}_5}$ e $N_{\mathbf{u}_6}$, ovvero per le numerosità dei profili che prevedono la frequenza dei soli centri C_2 e C_3, sono le soluzioni del seguente sistema lineare:

$$\begin{bmatrix} 1 & -\dfrac{f_{\mathbf{u}_4}}{n_2+n_3-f_{\mathbf{u}_4}} & -\dfrac{f_{\mathbf{u}_4}}{n_2+n_3-f_{\mathbf{u}_4}} \\ -\dfrac{f_{\mathbf{u}_5}}{n_2+n_3-f_{\mathbf{u}_5}} & 1 & -\dfrac{f_{\mathbf{u}_5}}{n_2+n_3-f_{\mathbf{u}_5}} \\ -\dfrac{f_{\mathbf{u}_6}}{n_2+n_3-f_{\mathbf{u}_6}} & -\dfrac{f_{\mathbf{u}_6}}{n_2+n_3-f_{\mathbf{u}_6}} & 1 \end{bmatrix}\begin{bmatrix} N_{\mathbf{u}_4} \\ N_{\mathbf{u}_5} \\ N_{\mathbf{u}_6} \end{bmatrix}=\begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} \tag{6}$$

dove

$$c_r=\left(\hat{N}_{\mathbf{u}_1}+\hat{N}_{\mathbf{u}_2}+\hat{N}_{\mathbf{u}_3}\right)\frac{f_{\mathbf{u}_{r+3}}}{n_2+n_3-f_{\mathbf{u}_{r+3}}},\ (r=1,2,3).$$

Merita di essere sottolineato che se da un lato la matrice chiamata in causa coinvolge quantità campionarie così che la sua invertibilità deve essere verificata di volta in volta, dall'altro l'applicazione empirica a dati reali ha messo in luce come in generale non si ponga alcun problema di indeterminazione. Inoltre l'approccio proposto può essere generalizzato in modo immediato e naturale al caso di un generico numero L di centri.

A tal fine si supponga che $\mathbf{u}_r$ $(r=1,\ldots,2^{L-1}-1)$, siano i profili che prevedono la frequenza, non esclusiva, al centro C_1, ovvero che presentano un "1" in prima posizione ed almeno un altro "1" in una delle posizioni restanti; sia inoltre $\mathbf{u}_{2^L-1}$ il profilo che prevede la frequenza esclusiva del primo centro. Con riferimento a tali profili, la conoscenza della numerosità N_1 consente di pervenire alla stima di massima verosimiglianza per le corrispondenti numerosità, stima avente la struttura (5).

Viceversa, per i successivi $2^{L-1}-1$ profili $\mathbf{u}_r$ $(r=2^{L-1},\ldots,2^L-2)$, si ipotizzi la non frequenza del centro C_1, ovvero uno "0" in prima posizione.

La stima per le ignote numerosità rimanenti $N_{\mathbf{u}_r}$ $(r=2^{L-1},\ldots,2^L-2)$, risulta allora la soluzione del sistema lineare $\mathbf{A}\mathbf{N}_{\mathbf{u}}=\mathbf{c}$, dove $\mathbf{A}$ è una matrice quadrata di dimensione $2^{L-1}-1$, il cui generico elemento a_{rj} risulta

$$a_{rj} = \begin{cases} 1 & \text{se } r = j \\ -\dfrac{f_{\mathbf{u}_{r+2^{L-1}-1}}}{\sum_{l=2}^{L} n_l - f_{\mathbf{u}_{r+2^{L-1}-1}}} & \text{se } r \neq j \end{cases} \tag{7}$$

$\mathbf{N_u} = \left\lfloor N_{\mathbf{u}_{2^{L-1}}}, \ldots, N_{\mathbf{u}_{2^L-2}} \right\rfloor$ è il vettore delle numerosità ignote, e il generico elemento c_r del vettore $\mathbf{c}$ dei termini noti è

$$c_r = \left(\sum_{i=1}^{2^{L-1}-1} \hat{N}_{\mathbf{u}_i} \right) \frac{f_{\mathbf{u}_{r+2^{L-1}-1}}}{\sum_{l=2}^{L} n_l - f_{\mathbf{u}_{r+2^{L-1}-1}}}, \quad \left(r = 1, \ldots, 2^{L-1} - 1\right). \tag{8}$$

4. PARZIALE *DETECTABILITY*

Alcuni centri si caratterizzano, come precedentemente evidenziato, per la completa assenza d'informazioni riguardo alla numerosità dei presenti al momento del campionamento. In tali casi, benché sia opportuno effettuare le estrazioni nel momento di massimo affollamento, non è assicurata la completa visibilità dei soggetti. Ad esempio, il campionamento in una piazza può essere svolto un sabato o domenica pomeriggio, cioè nel momento di massima presenza, ma d'altro canto alcune unità possono essere comunque assenti durante il campionamento.

Si presume, quindi, che nel generico centro l, $\left(l = 1, \cdots, L\right)$, al momento dell'indagine siano presenti N_l' soggetti anziché N_l, con $N_l' \leq N_l$. Di conseguenza anche le numerosità dei differenti profili potranno risultare inferiori a quelle reali, e, essendo le stime di queste ultime le quantità utilizzate per la stima della numerosità totale, è evidente la necessità di introdurre un elemento correttivo. A tale riguardo, sembra particolarmente utile l'impiego della *detectability* (Thompson, 1992, 1994), cioè della probabilità che un'unità statistica sia osservabile al momento del campionamento. Risulta, inoltre, naturale assumere tale *detectability* costante sugli individui dotati del medesimo profilo. Ciò significa, per esempio, che due immigrati iscritti in anagrafe e individuabili in un *phone-center* abbiano la stessa *detectability* poiché possiedono lo stesso profilo nel caso di due soli centri.

In accordo con quanto sopra siano:

$$\vartheta_{\mathbf{u}_r} = \frac{N'_{\mathbf{u}_r}}{N_{\mathbf{u}_r}}, \; r = 1, \ldots, 2^L - 1 \qquad \vartheta_{N_l} = \frac{N'_l}{N_l}, \; l = 1, \ldots, L \tag{9}$$

le *detectability* per le unità caratterizzate dal profilo $\mathbf{u}_r$ e, rispettivamente, per le unità che frequentano il centro *l*, dove $N'_{\mathbf{u}_r}$ indica l'effettivo numero di soggetti con profilo $\mathbf{u}_r$ presenti al momento del campionamento.

Si osservi che la *detectability* del generico *l*-esimo centro, ϑ_{N_l}, risulta essere una media ponderata delle *detectability* dei profili.

Ai fini dell'individuazione di una stima della numerosità della popolazione, alle quantità reali ignote ($N_l, N_{\mathbf{u}_r}$) devono sostituirsi quelle visibili al momento del campionamento ($N'_l, N'_{\mathbf{u}_r}$). La conseguente comparsa delle quantità $\vartheta_{\mathbf{u}_r}$ e ϑ_{N_l} nella funzione di verosimiglianza, quantità che possono vedersi alla stregua di parametri di disturbo, pone un problema di eliminazione dei medesimi. Tuttavia, non risulta possibile un'eliminazione tramite massimizzazione, ovvero con il ricorso ad una verosimiglianza profilo, in quanto l'informazione campionaria che verrebbe utilizzata ai fini della stima delle *detectability* renderebbe impossibile la successiva individuazione di una stima di massima verosimiglianza per le ignote numerosità di profilo.

Pertanto, risulta opportuno supporre che $\vartheta_{\mathbf{u}_r}$ e ϑ_{N_l} siano determinazioni di altrettante variabili casuali, per le quali occorre fare ricorso all'elicitazione di opportune leggi di distribuzione a priori, così che l'eliminazione dei parametri di disturbo risulta possibile tramite il ricorso a verosimiglianze integrate.

Con riferimento alle distribuzioni a priori, al fine di preservare il più possibile le caratteristiche della popolazione considerata, si possono adottare le seguenti assunzioni. Innanzitutto, indicata con m_r $(r = 1,\ldots,2^L - 1)$ la molteplicità dell'*r*-esimo profilo, ovvero la somma di "1" presenti nel profilo medesimo, non è per nulla restrittivo supporre che vi sia uguaglianza in distribuzione fra *detectability* $\vartheta_{\mathbf{u}_r}$ relative a profili con la stessa molteplicità. In secondo luogo appare del tutto realistico supporre che vi sia indipendenza fra le variabili casuali $\vartheta_{\mathbf{u}_r}$.

Di conseguenza, dalle definizioni e dalle assunzioni sopra esposte discendono sia la dipendenza sia la diversità in distribuzione di ϑ_{N_l}.

Con riferimento alle leggi a priori, dato che la visibilità di un individuo è chiaramente decrescente al crescere del numero di centri frequentati, risulta naturale porre le leggi medesime in funzione delle molteplicità m_r. In particolare, queste ultime possono essere introdotte nell'estremo inferiore del supporto della generica variabile casuale $\vartheta_{\mathbf{u}_r}$, supporto che deve comunque essere un sottoinsieme dell'intervallo [0,1]. Una formulazione generale dell'estremo inferiore per $\vartheta_{\mathbf{u}_r}$ che soddisfa alle caratteristiche di cui sopra è la seguente:

$$k_1 + k_2\left(\frac{m_{\max} - m_r}{m_{\max} - m_{\min}}\right) \tag{10}$$

dove k_1 è la percentuale di affollamento minima comune a tutti i profili, k_2 rappresenta il "tasso di affollamento", cioè la variazione di affollamento minimo effettivamente previsto per il profilo considerato in funzione della molteplicità del profilo stesso e $m_{\max}$ e $m_{\min}$ sono, rispettivamente, il valore massimo e quello minimo assumibili dalla molteplicità m_r. È evidente che all'aumentare del numero di centri previsti dal profilo $\mathbf{u}_r$, l'estremo inferiore del supporto per i differenti profili varia uniformemente tra k_1 e $k_1 + k_2$. Ad esempio, nel caso di tre centri e posto ragionevolmente $k_1 = 0.6$ (ovvero si suppone presente al momento del campionamento almeno il 60% dei soggetti con profilo di molteplicità massima) e $k_2 = 0.2$ (cioè almeno il (60+20)% dei soggetti con profilo di molteplicità minima si suppone presente al momento del campionamento), l'estremo inferiore del supporto di $\vartheta_{\mathbf{u}_r}$ (r =1, …, 7) sarà pari 0.8 se la molteplicità risulta 1, a 0.7 se la molteplicità è pari a 2, ed infine con molteplicità massima, ovvero $m_r = 3$, l'estremo inferiore sarà uguale a 0.6.

Il modo in cui sono state definite le *detectability* condiziona la scelta delle distribuzioni a priori. Risulta, infatti, conveniente utilizzare delle distribuzioni Beta di parametri $\alpha > 1$ e $\beta = 1$ troncate nell'estremo fornito con la (10); in tale modo si assegna una più elevata probabilità ad intervalli di valori della *detectability* prossimi all'estremo superiore del supporto, e ciò risulta del tutto coerente con il fatto che la rilevazione avviene nel momento di maggiore affollamento dei centri. Nel presente lavoro sono state prese in considerazione delle distribuzioni Beta di parametri $\alpha = 2, \beta = 1$, vale a dire leggi di distribuzione triangolari troncate.

5. PARZIALE *DETECTABILITY*: STIMA DI *N* NEL CASO DI DUE CENTRI

Fissando l'attenzione sul caso di due centri, così come è emerso nei paragrafi 2 e 3 per il caso di completa visibilità, non è possibile procedere alla determinazione della stima in completa assenza di informazioni. Supponendo allora che la numerosità di uno dei due centri sia nota (per esempio quella di C_1) ovvero che il medesimo sia una lista incompleta, si rende necessario introdurre leggi di distribuzione a priori solo per $\vartheta_{\mathbf{u}_2}$ e per $\vartheta_{\mathbf{u}_3}$. Di conseguenza, posto, ad esempio, $k_1 = 0.6$ e $k_2 = 0.2$, la distribuzione congiunta per $\vartheta_{\mathbf{u}_2}$ e $\vartheta_{\mathbf{u}_3}$ risulta essere:

$$\varphi\left(\vartheta_{\mathbf{u}_2}, \vartheta_{\mathbf{u}_3}\right) = c\,\vartheta_{\mathbf{u}_2}\vartheta_{\mathbf{u}_3}; \qquad \frac{4}{5} < \vartheta_{\mathbf{u}_2} < 1; \quad \frac{3}{5} < \vartheta_{\mathbf{u}_3} < 1 \tag{11}$$

dove c è un'opportuna costante di normalizzazione.

Posto che la funzione di verosimiglianza relativa al secondo centro, condizionatamente al valore assunto dalle *detectability* $\vartheta_{\mathbf{u}_2}$ e $\vartheta_{\mathbf{u}_3}$, può esprimersi come:

$$L\left(N_{\mathbf{u}_2}, N_{\mathbf{u}_3} \mid \vartheta_{\mathbf{u}_2}, \vartheta_{\mathbf{u}_3}; f_{\mathbf{u}_2}\right) \propto \left(\frac{N_{\mathbf{u}_2}\vartheta_{\mathbf{u}_2}}{N_{\mathbf{u}_2}\vartheta_{\mathbf{u}_2} + N_{\mathbf{u}_3}\vartheta_{\mathbf{u}_3}}\right)^{f_{\mathbf{u}_2}} \left(1 - \frac{N_{\mathbf{u}_2}\vartheta_{\mathbf{u}_2}}{N_{\mathbf{u}_2}\vartheta_{\mathbf{u}_2} + N_{\mathbf{u}_3}\vartheta_{\mathbf{u}_3}}\right)^{n_2 - f_{\mathbf{u}_2}} \tag{12}$$

risulta comodo riparametrizzare ponendo $x = \dfrac{N_{\mathbf{u}_2}\vartheta_{\mathbf{u}_2}}{N_{\mathbf{u}_2}\vartheta_{\mathbf{u}_2} + N_{\mathbf{u}_3}\vartheta_{\mathbf{u}_3}}$ e $y = \vartheta_{\mathbf{u}_2}$, così che la (11) può essere riespressa come:

$$\varphi(x, y) = c\,\frac{{N_{\mathbf{u}_2}}^2}{{N_{\mathbf{u}_3}}^2}\,\frac{1-x}{x^3}\,y^3; \qquad x_{\min} \le x \le x_{\max}; \quad 0.8 \le y \le 1 \tag{13}$$

dove $x_{\min} = \dfrac{(4/5)N_{\mathbf{u}_2}}{(4/5)N_{\mathbf{u}_2} + N_{\mathbf{u}_3}}$, $x_{\max} = \dfrac{N_{\mathbf{u}_2}}{N_{\mathbf{u}_2} + (3/5)N_{\mathbf{u}_3}}$.

La verosimiglianza integrata risulta pertanto:

$$L_{\text{int}}\left(N_{\mathbf{u}_2}, N_{\mathbf{u}_3}\right) \propto \int_{0.8}^{1} \int_{x_{\min}}^{x_{\max}} L\left(N_{\mathbf{u}_2}, N_{\mathbf{u}_3} \mid x, y; f_{\mathbf{u}_2}\right)\varphi(x, y)\,dx\,dy \;\propto$$

$$\propto \frac{{N_{\mathbf{u}_2}}^2}{{N_{\mathbf{u}_3}}^2}\left\{B\left(f_{\mathbf{u}_2} - 2, n_2 - f_{\mathbf{u}_2} + 2; x_{\max}\right) - B\left(f_{\mathbf{u}_2} - 2, n_2 - f_{\mathbf{u}_2} + 2; x_{\min}\right)\right\} \tag{14}$$

dove $B(\alpha, \beta; x) = \int_0^x u^{\alpha-1}(1-u)^{\beta-1}\,du$ indica la funzione Beta incompleta.

Si noti che, affinché la (14) risulti ben posta, deve essere $f_{\mathbf{u}_2} \ge 3$, ma tale condizione non pare assolutamente restrittiva. Inoltre, introducendo la stima per $N_{\mathbf{u}_3}$ ottenuta dalle informazioni campionarie relative al primo centro e fornita con la (1), la (14) assume la forma:

$$L_{\text{int}}\left(N_{\mathbf{u}_2}\right) \propto \frac{{N_{\mathbf{u}_2}}^2}{{\hat{N}_{\mathbf{u}_3}}^2}\left\{B\left(f_{\mathbf{u}_2} - 2, n_2 - f_{\mathbf{u}_2} + 2; x_{\max}\right) - B\left(f_{\mathbf{u}_2} - 2, n_2 - f_{\mathbf{u}_2} + 2; x_{\min}\right)\right\} \tag{15}$$

Una massimizzazione della medesima può ovviamente avvenire solo per via numerica, in tale modo si perviene ad una stima $\hat{\hat{N}}_{\mathbf{u}_2}$ per $N_{\mathbf{u}_2}$. Facendo inoltre ricorso alla (1), si ottiene la nuova stima $f_{\mathbf{u}_1} > 11$ per N:

$$\hat{\hat{N}} = \hat{N}_{\mathbf{u}_1} + \hat{\hat{N}}_{\mathbf{u}_2} + \hat{N}_{\mathbf{u}_3}. \tag{16}$$

6. PARZIALE *DETECTABILITY*: STIMA DI *N* NEL CASO DI TRE O PIÙ CENTRI

Se $L = 3$ la stima della numerosità totale può avvenire in modo analogo a quello relativo a due soli centri. Supposta nota la numerosità del centro C_1, è possibile, tramite la verosimiglianza, ottenere delle stime per le numerosità di tutti i profili che prevedono la frequenza del primo centro ($\hat{N}_{\mathbf{u}_1}, \hat{N}_{\mathbf{u}_2}, \hat{N}_{\mathbf{u}_3}, \hat{N}_{\mathbf{u}_7}$, cfr. figura 1).

Le numerosità dei restanti profili verranno determinate tenendo conto delle corrispondenti *detectability*, che risultano tra loro indipendenti con leggi di distribuzione triangolari troncate, così che la legge a priori congiunta è esprimibile come:

$$\varphi\left(\vartheta_{\mathbf{u}_r};\ r = 1,\ldots,6\right) = c\prod_{r=1}^{6}\vartheta_{\mathbf{u}_r} \tag{17}$$

dove *c* è un'opportuna costante di normalizzazione e, coerentemente con le ipotesi emesse nel paragrafo 4 in merito agli estremi del supporto di ciascuna *detectability* e ponendo ancora $\hat{N}_{\mathbf{u}_1}, \hat{N}_{\mathbf{u}_2}, \hat{N}_{\mathbf{u}_3}$ e $k_2 = 0.2$, si ha: $0.6 < \vartheta_{\mathbf{u}_1} < 1$ dato che la molteplicità del profilo $\mathbf{u}_1$ è massima e pari a 3, $0.7 < \vartheta_{\mathbf{u}_r} < 1$ per r = 2, 3, 5 essendo la molteplicità dei profili $\mathbf{u}_2$ $\mathbf{u}_3$ e $\mathbf{u}_5$ pari a 2 e $0.8 < \vartheta_{\mathbf{u}_r} < 1$, per r = 4, 6, essendo la molteplicità dei profili $\mathbf{u}_4$ $\mathbf{u}_6$ pari a 1.

Generalizzando la (12) è possibile ottenere la funzione di verosimiglianza relativa ai profili che prevedono la frequentazione di almeno uno dei due centri C_2 e C_3:

$$L\left(N_{\mathbf{u}_r}\middle|\vartheta_{\mathbf{u}_r}; f_{\mathbf{u}_r}; r = 1,\cdots,6\right) \propto \prod_{r=1}^{6}\left(\frac{N_{\mathbf{u}_r}\vartheta_{\mathbf{u}_r}}{\sum_{j=1}^{6} N_{\mathbf{u}_j}\vartheta_{\mathbf{u}_j}}\right)^{f_{\mathbf{u}_r}}. \tag{18}$$

Posto $x_r = \dfrac{N_{\mathbf{u}_r}\vartheta_{\mathbf{u}_r}}{\sum_{j=1}^{6} N_{\mathbf{u}_j}\vartheta_{\mathbf{u}_j}}$, r = 1, …, 5; $x_6 = 1 - \sum_{i=1}^{5} x_r$, ed $y = \vartheta_{\mathbf{u}_1}$

dalla (17) discende che:

$$\varphi\left(x_1, x_2, x_3, x_4, x_5, y\right) \propto \frac{N_{\mathbf{u}_1}{}^{10}}{\left(N_{\mathbf{u}_2} N_{\mathbf{u}_3} N_{\mathbf{u}_4} N_{\mathbf{u}_5} N_{\mathbf{u}_6}\right)^2} \frac{x_2 x_3 x_4 x_5\left(1 - \sum_{r=1}^{5} x_r\right)}{\left(x_1\right)^{11}} y^{11} \tag{19}$$

con $0.6 \le y \le 1$; $x_{\min(r)} \le x_r \le x_{\max(r)}$, $(r = 1,\cdots,5)$ dove $x_{\min(r)}$ e $x_{\max(r)}$ si deducono dai supporti di $\vartheta_{\mathbf{u}_r}$ $(r = 1,\cdots,6)$. Ad esempio risulta:

$$x_{\min(1)} = \frac{0.6N_{\mathbf{u}_1}}{0.6N_{\mathbf{u}_1} + \sum_{j=2}^{6} N_{\mathbf{u}_j}}$$

$$x_{\max(1)} = \frac{N_{\mathbf{u}_1}}{N_{\mathbf{u}_1} + 0.7(N_{\mathbf{u}_2} + N_{\mathbf{u}_3} + N_{\mathbf{u}_5}) + 0.8(N_{\mathbf{u}_4} + N_{\mathbf{u}_6} + N_{\mathbf{u}_7})}$$

Ciò premesso, una volta sostituite le stime $\hat{N}_{\mathbf{u}_1}, \hat{N}_{\mathbf{u}_2}, \hat{N}_{\mathbf{u}_3}$ ottenute tramite le informazioni campionarie relative al primo centro, la verosimiglianza integrata risulta:

$$L_{\text{int}}\left(N_{\mathbf{u}_4}, N_{\mathbf{u}_5}, N_{\mathbf{u}_6}\right) \propto \int_{x_{\min(1)}}^{x_{\max(1)}} \cdots \int_{x_{\min(5)}}^{x_{\max(5)}} \frac{\hat{N}_{\mathbf{u}_1}^{\ 10}}{\left(\hat{N}_{\mathbf{u}_2} \hat{N}_{\mathbf{u}_3} N_{\mathbf{u}_4} N_{\mathbf{u}_5} N_{\mathbf{u}_6}\right)^2} \left(x_1\right)^{f_{\mathbf{u}_1}-11} \left(x_2\right)^{f_{\mathbf{u}_2}+1} \left(x_3\right)^{f_{\mathbf{u}_3}+1} \left(x_4\right)^{f_{\mathbf{u}_4}+1} \left(x_5\right)^{f_{\mathbf{u}_5}+1} \left(1 - \sum_{r=1}^{5} x_r\right)^{f_{\mathbf{u}_6}+1} dx_1 \ldots dx_5 \tag{20}$$

Si noti che nella funzione integranda che compare nella (20) è possibile riconoscere il nucleo di una variabile casuale di Dirichlet purché sia soddisfatto il vincolo, non restrittivo, che sia $f_{\mathbf{u}_1} \ge 11$. Poiché tale vincolo deriva dall'aver posto $y = \vartheta_{\mathbf{u}_1}$, potrebbe essere utile porre y pari al $\vartheta_{\mathbf{u}_r}$ associato al profilo la cui frequenza campionaria è maggiore.

Le stime delle numerosità di profilo $N_{\mathbf{u}_4}, N_{\mathbf{u}_5}, N_{\mathbf{u}_6}$ ancora ignote possono essere dedotte dalla massimizzazione della verosimiglianza integrata che non può che avvenire per via numerica.

Merita, infine, di essere sottolineato che la procedura proposta per il caso di 3 centri può riproporsi in modo analogo per un generico numero L di centri. In particolare, la (17) e la (18) assumono rispettivamente le nuove forme:

$$\varphi\left(\vartheta_{\mathbf{u}_r};\ r = 1,\ldots,2^L - 2\right) = c \prod_{r=1}^{2^L-2} \vartheta_{\mathbf{u}_r} \tag{21}$$

$$L\left(N_{\mathbf{u}_r} \middle| \vartheta_{\mathbf{u}_r}; f_{\mathbf{u}_r}; r = 1,\cdots,2^L - 2\right) \propto \prod_{r=1}^{2^L-2} \left(\frac{N_{\mathbf{u}_r} \vartheta_{\mathbf{u}_r}}{\sum_{j=1}^{2^L-2} N_{\mathbf{u}_j} \vartheta_{\mathbf{u}_j}} \right)^{f_{\mathbf{u}_r}} \tag{22}$$

e, con trasformazioni analoghe a quelle del caso di tre centri, la (19) può essere generalizzata come segue:

$$\varphi\left(x_r, y;\, r=1,\ldots,2^L-3\right) \propto \frac{N_{\mathbf{u}_1}{}^{2^{L+1}-6}}{\prod_{r=2}^{2^L-2} N_{\mathbf{u}_r}^2} \frac{\prod_{r=2}^{2^L-3} x_r \left(1-\sum_{r=1}^{2^L-3} x_r\right)}{\left(x_1\right)^{2^{L+1}-5}} y^{2^{L+1}-5}. \tag{23}$$

In tale modo è possibile sostituire le stime $\hat{N}_{\mathbf{u}_1},\ldots,\hat{N}_{\mathbf{u}_{2^{L-1}-1}}$ ottenute dalle informazioni campionarie relative al centro C_1 e pervenire quindi alla verosimiglianza integrata:

$$L_{\text{int}}\left(N_{\mathbf{u}_r};\, r=2^{L-1},\ldots,2^L-2\right) \propto \int_{x_{\min(1)}}^{x_{\max(1)}} \cdots \int_{x_{\min(2^L-3)}}^{x_{\max(2^L-3)}} \frac{\hat{N}_{\mathbf{u}_1}{}^{2^{L+1}-6}}{\prod_{r=2}^{2^{L-1}-1} \hat{N}_{\mathbf{u}_r}^2 \prod_{r=2^{L-1}}^{2^L-2} N_{\mathbf{u}_r}^2} \tag{24}$$

$$\left(x_1\right)^{f_{\mathbf{u}_1}-2^{L+1}+5} \prod_{r=2}^{2^L-3} x_r^{f_{\mathbf{u}_r}+1} \left(1-\sum_{r=1}^{2^L-3} x_r\right)^{f_{\mathbf{u}_{2^L-2}}+1} dx_1 \ldots dx_{2^L-3}$$

7. RISULTATI DI ALCUNE SIMULAZIONI

Al fine di valutare le stime proposte nei paragrafi precedenti e tenendo presente che la complessità di quelle relative all'assenza di completa visibilità ne impone una individuazione numerica, si è proceduto con uno studio simulativo comparato.

La popolazione di riferimento prescelta è quella costituita da 1000 immigrati extracomunitari presenti nell'area milanese e campionati con la tecnica del campionamento per centri nel corso di un'indagine svolta nel 2000 (AA.VV., 2000).

In primo luogo si è ritenuto opportuno aggregare i 23 centri, individuati nella ricerca sul campo, in modo da potersi ricondurre ai casi più semplici di due o tre centri. Il criterio utilizzato per imputare un centro originale è stato quello della possibilità, o meno, di disporre di una lista incompleta o, per lo meno, di potere effettuare un'enumerazione.

In particolare, nel caso di tre centri C_1 è costituito dai punti di aggregazione con disponibilità di liste (per esempio, consultori, scuole di lingua), C_2 rappresenta i centri in cui sia possibile avere l'ordine di grandezza dei frequentanti, (costituito, prevalentemente, da punti di aggregazione gestiti da religiosi), ed infine C_3 è formato dall'insieme dei centri in cui sia impossibile qualunque enumerazione (in gran parte centri di svago o negozi).

Per quanto riguarda il caso di due centri, C_1 rappresenta l'insieme dei punti di aggregazione in cui sia disponibile qualche informazione sulla numerosità della frequenza (è costituito dai centri C_1 e C_2 del caso precedente), mentre C_2 coincide con il centro C_3 definito poco sopra.

La simulazione è consistita nell'estrazione di p=500 campioni dai 2 o 3 punti di aggregazione presenti secondo la tecnica di campionamento per centri e per diverse combinazioni di ampiezze campionarie e di valori delle *detectability*. Inoltre, accanto alla popolazione reale fornita dalla rilevazione sono state considerate differenti popolazioni simulate in cui le 1000 unità presentano profili diversi da quelli originali in modo da prefigurare scenari il più possibile "problematici" ai fini delle stime individuate.

Su ciascuno dei p campioni estratti si è proceduto al calcolo delle stime proposte nei precedenti paragrafi. Infine, le medie aritmetiche dei p valori ottenuti per tutte le suddette quantità hanno fornito una stima Monte Carlo per i valori attesi dei corrispondenti stimatori.

In particolare, la Tabella 1 riporta, per il caso di 2 centri, la stima relativa ad una situazione di completa *detectability* fornita con la (4).

TABELLA 1

Risultati simulazioni per la stima di N =1000 nel caso di due centri con completa visibilità

Ipotesi	Parametri della simulazione							$E(\hat{N})$
	N_1	N_2	$N_{\mathbf{u}_1}$	$N_{\mathbf{u}_2}$	$N_{\mathbf{u}_3}$	n_1	n_2	
A	782	904	96	218	686	25	25	1018
B	782	904	96	218	686	100	100	1005
C	782	904	96	218	686	200	200	999
D	600	600	400	400	200	100	100	1012
E	400	800	200	600	200	100	100	1023

I risultati delle simulazioni evidenziano come, in media, $\hat{N}$ risulta una buona approssimazione di N in tutti i casi considerati. In particolare, nel caso delle ipotesi A, B, e C è stata mantenuta invariata la popolazione iniziale, ma è stata cambiata l'ampiezza campionaria; ciò ha consentito di osservare una migliore *performance* di $\hat{N}$ al crescere di n. Nelle ipotesi D ed E la popolazione iniziale è stata variata in modo tale da ottenere una differente distribuzione delle numerosità dei profili privilegiando $\mathbf{u}_2$, cioè il profilo dei soggetti che frequentano solo il centro per cui non si ha a disposizione alcuna informazione, ponendosi in tal modo nella condizione più sfavorevole. In tale situazione si nota come al crescere della numerosità di $\mathbf{u}_2$ si registri una leggera sovrastima per $\hat{N}$ e ciò coerentemente col fatto che all'aumentare di $N_{\mathbf{u}_2}$, e fissato N pari a 1000, diminuisce l'apporto di informazioni che si suppone avere a priori tramite la conoscenza della numerosità del primo centro.

La tabella 2 si riferisce al caso di tre centri sempre nella situazione di completa *detectability*, e consente di valutare in media il comportamento dello stimatore ottenuto come soluzione del sistema (6).

TABELLA 2

Risultati simulazioni per la stima di N =1000 nel caso di tre centri con completa visibilità

	Parametri della simulazione													$E(\hat{N})$
	N_1	N_2	N_3	$N_{\mathbf{u}_1}$	$N_{\mathbf{u}_2}$	$N_{\mathbf{u}_3}$	$N_{\mathbf{u}_4}$	$N_{\mathbf{u}_5}$	$N_{\mathbf{u}_6}$	$N_{\mathbf{u}_7}$	n_1	n_2	n_3	
A	496	521	904	214	21	226	218	246	40	35	17	17	17	1048
B	496	521	904	214	21	226	218	246	40	35	67	67	67	1024
C	496	521	904	214	21	226	218	246	40	35	133	133	133	1019
D	496	521	904	214	21	226	218	246	40	35	52	54	94	1015
E	496	521	904	214	21	226	218	246	40	35	104	108	188	1014
F	490	490	490	70	110	110	200	110	200	200	67	67	67	1000
G	220	620	620	50	50	50	260	260	260	70	67	67	67	1035

Si osserva che, anche in questo caso, la stima $\hat{N}$ risulta una buona approssimazione del valore di N e tende a migliorare al crescere della numerosità campionaria (ipotesi A, B, C). Con le ipotesi D ed E si nota che un ulteriore effetto positivo sulla capacità approssimativa di $\hat{N}$ può ottenersi allocando proporzionalmente n nei tre centri. Quest'ultima ipotesi non appare difficilmente realizzabile nella realtà poiché, agli operatori sul campo può essere noto a priori l'ordine di grandezza dell'affluenza ai diversi centri.

Una distribuzione più "problematica" della popolazione sui profili (ipotesi F) si ottiene ipotizzando che un numero relativamente ridotto di unità statistiche presentino profilo $\mathbf{u}_1$ o $\mathbf{u}_2$ o $\mathbf{u}_3$, cioè frequentino oltre a C_2 o C_3 anche C_1 dove la numerosità è posta nota. Non sembra che tale ridistribuzione influisca sulla *performance* dello stimatore. Ciò nonostante, se un panorama di questo tipo viene esasperato (ipotesi G) riducendo sia il numero di frequentatori di C_1 sia la numerosità dei profili in comune con tale centro, lo stimatore proposto evidenzia una tendenza alla sovrastima.

Infine, nella tabella 3 vengono posti a confronto, nel caso di due centri, i valori attesi degli stimatori $\hat{N}$, $\hat{\hat{N}}$ forniti, rispettivamente, con la (4) e la (16), nel caso di parziale visibilità degli individui che frequentano C_2.

TABELLA 3

Risultati simulazioni per la stima di N =1000 nel caso di due centri con parziale visibilità

Ipotesi	Parametri della simulazione							$E(\hat{N})$	$E(\hat{\hat{N}})$
	$N_{\mathbf{u}_1}$	$N_{\mathbf{u}_2}$	$N_{\mathbf{u}_3}$	n_1	n_2	$\vartheta_{\mathbf{u}_2}$	$\vartheta_{\mathbf{u}_3}$		
A	96	218	686	25	25	0.8	0.6	1089	1049
B	96	218	686	50	50	0.8	0.6	1071	1011
C	96	218	686	100	100	0.8	0.6	1078	1004
D	96	218	686	100	100	0.95	0.95	1004	963
E	96	218	686	100	100	0.8	0.95	964	939
F	96	218	686	100	100	0.95	0.6	1136	1031
G	96	218	686	100	100	Casuale	Casuale	1030	978

Nelle ipotesi A, B e C, in cui le determinazioni assunte dalle *detectability* sono state poste pari al valor minimo assumibile, ricreando quindi la situazione di minor visibilità possibile, lo stimatore $\hat{\hat{N}}$, in media, risulta quello che meglio

approssima la quantità ignota e tale caratteristica tende ad accentuarsi ulteriormente al crescere di *n*. Viceversa, nel caso di quasi totale visibilità (ipotesi D), in media $\hat{\hat{N}}$ pare avere la tendenza a sottostimare, ma ciò è evidente poiché è stato ottenuto come media su tutti i possibili valori delle $\vartheta_{\mathbf{u}_r}$ (r = 2,3), mentre $\hat{N}$ in questa situazione risulta evidentemente lo stimatore più appropriato.

Le ipotesi E ed F riguardano le situazioni in cui la visibilità di un profilo è quasi totale, mentre è minima quella relativa all'altro. In tali contesti se è minima $\vartheta_{\mathbf{u}_3}$, cioè la visibilità delle unità che frequentano entrambi i centri, $\hat{\hat{N}}$ è ancora il miglior stimatore, tuttavia se ad essere minimo è $\vartheta_{\mathbf{u}_2}$ tutti gli stimatori presi in considerazione sembrano sottostimare *N*.

Un'ulteriore situazione che realisticamente può essere prospettata è quella in cui le determinazioni assunte dalle *detectability* vengono estratte casualmente in ogni campione, ovvero si suppone che in ciascuno dei 500 campioni estratti la visibilità possa essere differente; in tale situazione (ipotesi G) $\hat{N}$ e $\hat{\hat{N}}$ appaiono entrambi avere una buona capacità approssimativa.

Dipartimento di Statistica SONIA MIGLIORATI
Università degli Studi di Milano-Bicocca

Dipartimento di Statistica LAURA TERZERA
Università degli Studi di Milano-Bicocca

RIFERIMENTI BIBLIOGRAFICI

AA.VV., (2000), *L'immigrazione straniera nell'area milanese*, "Rapporto statistico dell'Osservatorio Fondazione Cariplo-I.S.MU.", Provincia di Milano.

M.D. BANKIER, (1986), *Estimators based on several stratified samples with* applications *to multiple frame surv*eys, "Journal of the American Statistical Association", 81, pp. 1074-1079.

T.L. BYCZKOWSKI, M.S. LEVY, D.J. SWEENEY, (1998), *Estimation in sample surveys* using *frames with a many-to-many structure*, "Survey Methodology", 24, 1, pp. 21-30.

G.C. BLANGIARDO, (1996), *Il campionamento per centri o ambienti di aggregazione nelle* indagini *sulla presenza straniera,* in "Studi in onore di Giampiero Landenna", Giuffré, Milano, pp. 15-30.

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

W.G. COCHRAN, (1977), *Sampling techniques*, 3rd ed., John Wiley & Sons, New York.

EUROSTAT, (2000), *Push and pull factors of international migration. Country report – Italy,* 3/2000/E/n.5 Bruxelles: European Communities Printing Office.

D.E. HAINES, K.H. POLLOCK, (1998), *Combining multiple frames to estimate population size and totals*, "Survey Methodology", 24, 1, pp. 79-88.

H.O. HARTLEY, (1974), *Multiple frame methodology and selected applications*, "Sankhya", ser. C., 36, pp. 99-118.

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

L. SANATHANAN, (1972), *Estimating the size of a multinomial population*, "Annals of Mathematical Statistics", 43, 1, pp. 142-152.

C.J. SKINNER, (1991), *On the efficiency of raking ratio estimation for multiple frame* surveys, "Journal of the American Statistical Association", 86, pp. 779-784.

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

S.K. THOMPSON, (1992), *Sampling*, Wiley, New York.

S.K. THOMPSON, (1994), *Detectability in conventional and adaptive sampling*, "Biometrics", 50, pp. 712-724.

RIASSUNTO

Una proposta di stima della numerosità nel campionamento per centri

Obiettivo del lavoro è l'individuazione di una stima della numerosità per popolazioni per le quali non esiste una lista completa, o anche solo più liste incomplete, e le cui unità sono parzialmente elusive. Per tali popolazioni (quale è, ad esempio, quella immigrata) la tecnica di campionamento per centri si rivela strumento essenziale. In tale ambito viene proposta una prima stima basata sull'approccio della verosimiglianza supponendo la totale *detectability* delle unità statistiche. Tale ipotesi, scarsamente realistica, viene quindi rimossa introducendo un'opportuna famiglia di leggi di distribuzione a priori per le *detectability* e proponendo quindi un'ulteriore stima della numerosità basata sulla verosimiglianza integrata. Infine, le suddette stime vengono confrontate tramite uno studio simulativo basato su dati reali.

SUMMARY

A new estimate of the population size in center sampling

The aim of the paper is to give an estimate of the population size when neither a complete list nor some incomplete lists are available and the population is partially undetectable. In such a framework the center sampling technique proves an essential tool. A first estimate relies on the likelihood approach assuming the complete detectability of the units. Then, this unrealistic hypothesis is removed introducing a prior distribution on detectability, so that a second estimate is given resorting to integrated likelihood. Finally, the previous estimates are compared through a simulation study.