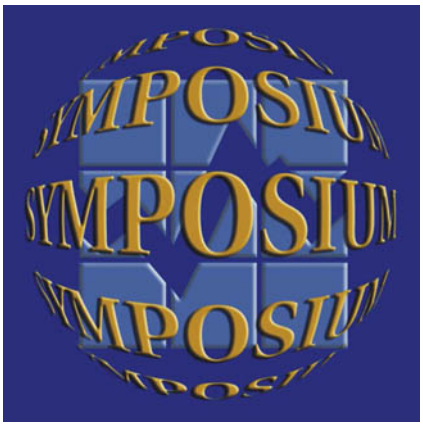
N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2004 : Méthodes innovatrices pour enquêter auprès des populations difficiles à joindre

2004



Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2004 de Statistique Canada
Méthodes innovatrices pour enquêter auprès des populations difficiles à joindre

ÉCHANTILLONNAGE DE CENTRES : STRATÉGIE D'ENQUÊTE AUPRÈS DES POPULATIONS DIFFICILES À ÉCHANTILLONNER

Fulvia Mecatti[1]

RÉSUMÉ

Au cours des 20 dernières années, l'immigration clandestine est devenue un problème important en Italie, ce qui devait aviver l'intérêt des organismes publics pour les enquêtes auprès de la population immigrante. Cette dernière est difficile à échantillonner, puisqu'il est habituellement impossible d'obtenir des listes ou des bases de sondage complètes et que les unités demandent ordinairement à garder l'anonymat, d'où l'impossibilité d'identifier les unités d'échantillonnage comme on a l'habitude de le faire. L'échantillonnage de centres, stratégie élaborée récemment en Italie, est destiné aux enquêtes auprès de la population immigrante. On répond alors aux questions suivantes : comment joindre les unités à des fins d'interview? comment estimer les paramètres d'intérêt? Il faut alors envisager une estimation de la moyenne d'une variable quantitative.

MOTS CLÉS : Bases de sondage chevauchantes, échantillonnage sur place, multiplicité, profil.

1. INTRODUCTION

On a conçu l'échantillonnage de centres (EC) en Italie pour les enquêtes menées auprès de la population immigrante.

L'immigration extraeuropéenne est devenue un grand problème en Italie ces 20 dernières années. L'explication principale en est peut-être que ce pays est passé rapidement d'une *terre d'émigration* (jusqu'à la fin des années 1960) à une *terre d'immigration* (depuis le début des années 1990). De plus, sa situation géographique offre un large accès à l'immigration principalement issue de l'Afrique du Nord, de l'Europe orientale et du Moyen-Orient, mais venant aussi d'Asie et d'Amérique latine. C'est un phénomène qui s'est rapidement et constamment amplifié dans les dernières décennies, souvent à l'encontre même des lois italiennes de l'immigration, si bien qu'une partie considérable de cette présence extraeuropéenne en sol italien aujourd'hui est illégale, est par conséquent hors de contrôle et présente des risques d'ordre social. Il faut aussi dire que deux amnisties récentes (en 1998 et 2002), qui visaient pourtant à contrôler la proportion de résidents clandestins et qui se limitaient à des catégories particulières, auraient pu favoriser un nouveau déferlement d'immigration clandestine. C'est ainsi que, de nos jours, les organismes publics d'intervention territoriale, migratoire et sociale éprouvent davantage le besoin de chiffrer et d'explorer un tel phénomène. Au niveau européen et au niveau national, on a entrepris dans la première moitié des années 1990 de faire périodiquement enquête sur la population immigrante.

Une enquête qui vise l'immigration tant officielle que clandestine s'adresse nettement à une population difficile à échantillonner au départ : on ne dispose pas dans ce cas d'une liste complète et précise pouvant servir de base de sondage; bien que finie et de taille N, la population cible est inconnue, ses unités ne sont pas identifiables et, en général, la théorie classique de l'échantillonnage ne s'applique pas. En ce qui concerne l'immigration officielle, un ensemble de listes partielles tirées de sources officielles qui possiblement se chevauchent, pourrait nous servir de base de sondage multiple. Quant à l'immigration clandestine, elle est cachée et élusive . On peut raisonnablement supposer qu'une partie considérable de ces immigrants sont sans abri et sans emploi, vivant dans des refuges ou dans des lieux exposés comme les parcs, les voitures ou des bâtiments désaffectés et se livrant à des activités illicites. Il se pose donc un problème de repérabilité et un grand souci de protection de l'anonymat des gens.

[1] Département de la statistique, Université de Milan-Bicocca, Via Bicocca degli Arcimboldi, 8, Ed. U7, 20126 Milan, Italie. Fulvia.Mecatti@unimib.it.

L'échantillonnage de centres est une stratégie récemment avancée par laquelle on entend répondre aux questions que nous avons déjà posées : comment échantillonner les unités? comment estimer les paramètres d'intérêt? À la section 2, nous répondrons à la première et présenterons les outils de base d'un cadre d'échantillonnage de centres (EC). À la section 3, nous parlerons de la question de l'estimation des paramètres d'intérêt. À la section 4, nous envisagerons une estimation de la moyenne selon un échantillonnage aléatoire simple tiré parmi tous les centres. À la section 5, nous examinerons les plans de sondage à un et à deux degrés. À la section 6 enfin, nous livrerons des observations en conclusion.

2. CENTRE, PROFIL ET MULTIPLICITÉ

Au début des années 1990, soit au tout début des enquêtes auprès de la population immigrante, on a appliqué des plans d'échantillonnage à dépistage de liens comme l'échantillonnage en réseau et l'échantillonnage par effet de boule de neige où les unités sélectionnées sont priées de signaler toute autre unité avec laquelle elles auraient des liens quelconques (voir l'examen poussé de la question dans Sudman et Kalton, 1986). Ces enquêtes du passé ont fait voir une habitude caractéristique de la population immigrante italienne qui consiste à se concentrer localement ou territorialement pour combler des besoins religieux, sanitaires, sociaux, récréatifs, etc. Ces lieux sont connus et bien situés sur le territoire d'intérêt de sorte qu'un échantillonnage sur place peut être envisagé (Sudman et Kalton, 1986).

On trouvait le terme « *centre* » dans le premier projet d'échantillonnage de centres (Blangiardo, 1996). Il s'agissait soit d'une liste partielle soit d'un lieu de concentration territoriale des unités visées. Ainsi, une mosquée est un centre où on peut joindre les immigrants musulmans. Il en va de même des refuges pour les immigrants sans abri ou de tout registre officiel décrivant une présence étrangère sur le territoire national.

On identifie un ensemble de L centres de manière à bien couvrir la population visée en supposant que chacune de ses unités appartient à au moins un centre ou s'y rend régulièrement. Le tableau 1 énumère $L=13$ centres dont on sait qu'ils embrassent la population immigrante de Milan grâce à une enquête réalisée en 2002 par l'ISMU (Institut pour l'intégration et la multiethnicité). Nous considérons trois types de centres; le type 1 est formé pour l'essentiel de listes partielles puisées à des sources administratives; le type 2 comprend les centres sans liste où un certain dénombrement est possible (centres qui distribuent des bons ou qui assurent un nombre fixe de services d'alimentation ou d'hébergement, par exemple); le type 3 comprend les centres sans base de sondage et sans données de dénombrement.

Tableau 1 : Ensemble de centres couvrant la population immigrante à Milan en 2002

l	Centres	Type
1	Centres d'accueil	2
2	Centres d'assistance sociale	2
3	Centres de formation linguistique	1
4	Centres religieux	3
5	Centres médicaux	1
6	Centres d'aide juridique et professionnelle	1
7	Associations culturelles	2
8	Centres de services et de renseignements	2
9	Services publics	2
10	Centres de divertissement	2-3
11	Centres commerciaux et commerces ethniques	3
12	Centres en plein air	3
13	Maisons particulières	3

Pour reprendre le cas général, supposons que les tailles respectives de la population N et des centres N_l $(l=1\cdots L)$ sont inconnues. Comme les unités peuvent appartenir à plus d'un centre ou se rendre dans plusieurs, les centres

seront d'ordinaire en chevauchement, d'où $\sum_l N_l > N$. À cause des impératifs d'anonymat, on doit aussi supposer que la structure et les tailles en chevauchement sont inconnues et que, en règle générale, il ne peut y avoir de correspondance algébrique entre les unités et les centres.

Dans le but de surmonter l'absence d'identificateurs d'unités et de s'occuper du chevauchement, l'échantillonnage de centres combine les éléments de sondage à base multiple et d'échantillonnage en réseau en utilisant des règles de *profil* et de *multiplicité*.

À l'ensemble d'unités identifiées par $\{1 \cdots i \cdots N\}$ qui représente habituellement la population *P*, on substitue une matrice $N \times L$ où les unités sont en ligne et les centres en colonne et où on trouve la valeur 1 si l'unité *i* appartient au centre *l* et la valeur 0 dans les autres cas (comme on peut le voir à la figure 1). À noter que la somme de chaque colonne correspond à la valeur de taille des centres N_l.

Chaque ligne de la matrice (théorique) de la figure 1 définit un profil d'unité. Un profil est un vecteur de longueur *L* qui nous dit dans quels centres l'unité *i* peut être repérée ou non. Soit $\mathbf{u}_r$ le profil générique inclus dans la classe $\mathbf{U}$ (connue) de tous les profils possibles $\left(r = 1 \cdots 2^L - 1\right)$. Comme chaque unité a un profil unique, il y a correspondance de plusieurs à un entre les unités $\{1 \cdots i \cdots N\}$ ou encore entre la population de valeurs $y\ \{y_1 \cdots y_i \cdots y_N\}$, d'une part, et la classe $\mathbf{U}$ de tous les profils, d'autre part, si bien qu'une partition théorique de la population *P* devient possible comme à la figure 1. Ainsi, par l'utilisation des profils, une sorte d'identification est retrouvée et le problème de chevauchement est surmonté.

L'ensemble D_r d'unités ayant le même profil $\mathbf{u}_r$ délimite un *domaine* comme dans l'échantillonnage à base multiple et un *réseau* comme dans l'échantillonnage en réseau.

Figure 1 : Partition de la population par profilage

$$
\begin{array}{l}
1\ \textbf{centre} \quad \cdots \quad L \\
\textbf{unités}\ \left\lfloor \dot{N} \right\rfloor
\end{array}
\begin{bmatrix}
0 & \cdots & 1 & \cdots & 1 \\
 & & \vdots & & \\
1 & \cdots & 0 & \cdots & 1 \\
 & & \vdots & & \\
0 & \cdots & 1 & \cdots & 0
\end{bmatrix}
\longrightarrow \mathbf{u}_r \in \mathbf{U} \longrightarrow D_r = \{i \in P \text{ avec profil } \mathbf{u}_r\}
$$

$$
\begin{array}{cccccc}
 & \downarrow & & \downarrow & & \downarrow \\
 & N_1 & \cdots & N_l & \cdots & N_L
\end{array}
\qquad r = 1 \cdots 2^L - 1
\qquad \text{avec } \bigcup_r D_r = P \text{ et } D_r \bigcap D_t = \varnothing \quad \forall r \neq t
$$

La multiplicité du profil $\mathbf{u}_r$, dont les éléments u_{rl} prennent la valeur 0 ou 1, est définie par $m_r = \sum_l u_{rl}$. Ainsi, elle indique le nombre de centres auxquels appartient chaque unité $i \in D_r$. La théorie de l'estimation de l'échantillonnage de centres que nous présentons dans les sections qui suivent procède par calcul de multiplicité pour déterminer les probabilités d'échantillonnage de chaque unité, ce qui dépend nettement du nombre de centres auxquels elle appartient. Nous pondérerons les paramètres et les estimateurs par valeurs de multiplicité pour rendre le plus applicables possible les résultats types de la théorie de l'échantillonnage.

3. ESTIMATION

Dans le contexte naturel où l'échantillonnage de centres a vu le jour tel que décrit dans l'introduction, l'intérêt s'est d'abord porté sur l'estimation de la population de taille *N* et plusieurs propositions ont déjà été faites à ce sujet dans des ouvrages scientifiques (voir Blangiardo, Migliorati et Terzera, 2004, pour un traitement de la question). L'attention s'avivant pour la question, on s'est intéressé à des caractéristiques de la population autres que la taille.

Soit *y* une variable d'enquête quantitative ou dichotomique. Comme *N* demeure inconnu, nous regardons d'abord l'estimation de la moyenne de population $\bar{Y}$. En considérant les profils $\mathbf{u}_r \in \mathbf{U}$ au lieu des unités non identifiées, le paramètre $\bar{Y}$ peut ainsi s'exprimer : $\bar{Y} = \sum_r \sum_{i \in D_r} y_i / N$. En intégrant la multiplicité, nous obtenons :

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

où $\alpha_l = N_l / N$ désigne le poids du centre *l* et où

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

indique la moyenne de ce même centre ajusté par la multiplicité. À noter que, en utilisant des règles de profil et de multiplicité, le chevauchement entre centres se retrouve dans la moyenne ajustée $\tilde{\tilde{Y}}_l$, si bien que la relation habituelle d'association se vérifie comme dans l'équation (1) tout en ne se vérifiant pas – à cause du chevauchement – dans le cas de la moyenne (non ajustée) du centre $\bar{Y}_l = \sum_{i \in l} y_i / N_l$. De plus, dans les enquêtes d'échantillonnage de centres qui ont normalement lieu auprès des immigrants italiens, on dispose habituellement de données auxiliaires de pondération pour les centres. Bien qu'on ignore les valeurs absolues de taille *N* et N_l, on peut supposer que le poids du centre α_l est connu. Dans ce cas, les résultats généraux de l'échantillonnage stratifié s'appliquent : si un estimateur sans biais $\bar{y}_l$ est donné pour la moyenne ajustée du centre $\tilde{\tilde{Y}}_l$, alors $\bar{y} = \sum_l \alpha_l \bar{y}_l$ sera un estimateur sans biais de la moyenne de population $\bar{Y}$. Par hypothèse d'indépendance entre centres, la variance est $V(\bar{y}) = \alpha_l^2 V(\bar{y}_l)$. Par ailleurs, comme l'échantillonnage peut se faire parmi les centres se chevauchant seulement et non parmi les profils, l'analogie avec l'échantillonnage stratifié ne nous aide plus et des adaptations théoriques spéciales s'imposent.

Dans des applications réelles d'échantillonnage de centres, les centres sont habituellement hétérogènes pour ce qui est de la variable d'enquête dans le cas de la population immigrante et il est donc suggéré d'échantillonner tous les centres indépendamment. On sait aussi par expérience que les profils sont observables quand on préserve l'anonymat des unités. Ainsi, on demande à chaque unité sélectionnée de déclarer non seulement la valeur d'enquête y_i, mais aussi son profil, c'est-à-dire les centres auxquels elle appartient en dehors de celui où elle a été échantillonnée.

Prenons l'échantillon du centre *l* où n_l unités sont sélectionnées selon un plan de sondage quelconque. Il peut donc y avoir partition des données de l'échantillon de chaque centre *l* par rapport à la classe $\mathbf{U}$ de tous les profils. Soit $d_r^{(l)}$ l'ensemble des unités échantillonnées dans le centre *l* et partageant le profil $\mathbf{u}_r$. Comme les unités ne sont identifiées que par leur profil, les probabilités d'inclusion réfèrent aux profils. Soit $\delta_{ri}^{(l)}$ l'indicateur d'appartenance à l'échantillon de l'unité *i* au profil $\mathbf{u}_r$ dans l'échantillon du centre *l*; c'est-à-dire qu'il s'agit d'une variable indicatrice aléatoire qui prend la valeur 1 si l'unité $i \in D_r$ figure dans l'échantillon du centre *l* et la valeur 0 autrement. $E(\delta_{ri}^{(l)}) = \pi_{ri}^{(l)}$ désigne alors la probabilité d'inclusion du premier ordre compte tenu du plan de sondage. Voici un estimateur sans biais, dans un plan de sondage général, de la moyenne ajustée $\tilde{\tilde{Y}}_l$ telle que définie en (2) (Mecatti et Migliorati, 2003) :

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

L'estimateur (3) est une combinaison linéaire des estimateurs du type Horvitz-Thompson pour les profils et pondéré par la multiplicité, ce qui prouve d'emblée l'absence de biais. La variance est exacte :

$$V(\bar{y}_l) = \frac{1}{N_l^2}\left\{\sum_r \frac{1}{m_r^2}\left[\sum_{i\in D_r} y_i^2\, u_{rl}\ \frac{1-\pi_{ri}^{(l)}}{\pi_{ri}^{(l)}} \ + \ \sum_{i\neq j}\sum_{\in D_r} y_i y_j u_{rl}\ \frac{\pi_{rij}^{(l)} - \pi_{ri}^{(l)}\pi_{rj}^{(l)}}{\pi_{ri}^{(l)}\pi_{rj}^{(l)}}\right]\right.$$

$$\left. + \ \sum_{r\neq t}\sum \frac{1}{m_r m_t}\sum_{i\in D_r}\sum_{j\in D_t} y_i y_j u_{rl} u_{tl}\ \frac{\pi_{ritj}^{(l)} - \pi_{ri}^{(l)}\pi_{tj}^{(l)}}{\pi_{ri}^{(l)}\pi_{tj}^{(l)}}\right\}, \tag{4}$$

où $\pi_{rij}^{(l)} = E\left(\delta_{ri}^{(l)} \cdot \delta_{rj}^{(l)}\right)$ désigne les probabilités conjointes d'inclusion d'une paire d'unités différentes ayant le même profil $i \neq j \in D_r$ et $\pi_{ritj}^{(l)} = E\left(\delta_{ri}^{(l)} \cdot \delta_{tj}^{(l)}\right)$, les probabilités conjointes d'inclusion d'une paire d'unités $i \in D_r$ et $j \in D_t$ aux profils différents $\mathbf{u}_r \neq \mathbf{u}_t$. À noter que les deux termes mis entre crochets dans (4) correspondent à la variabilité à l'intérieur des profils selon la variance type de l'estimateur de Horvitz-Thompson, alors que le troisième terme en (4) est dû à la variabilité additionnelle entre les différents profils.

4. ESTIMATION SOUS UN PLAN D'ÉCHANTILLONNAGE ALÉATOIRE SIMPLE

On s'est servi efficacement de l'échantillonnage de centres pour faire enquête auprès de la population immigrante en Italie par l'échantillonnage aléatoire simple (EAS) de chaque centre visé (voir Eurostat, 2000, par exemple). Un tel échantillonnage peut normalement se faire lorsqu'on dispose d'une liste d'unités des centres ou qu'un dénombrement de ces unités est possible (centres de types 1 et 2 au tableau 1, par exemple). Dans le cas des centres sans liste ni dénombrement (type 3 du tableau 1, par exemple), on peut échantillonner sur place en choisissant au préalable une période de visite d'un centre selon l'hypothèse que les N_l unités appartenant au centre *l* seront toutes présentes *presque avec certitude*. Une telle hypothèse peut se vérifier si on tient compte des habitudes connues de la population immigrante italienne et/ou des particularités du centre. Plus précisément, on choisit un jour et/ou une heure précis comme « moment d'occupation maximale du centre » pour la visite sur place et on fait des interviews avec un groupe de personnes, généralement les n_l unités que l'intervieweur peut rencontrer à l'entrée ou à l'intérieur. C'est ainsi qu'on peut oublier les facteurs d'échantillonnage sur place comme les sous-périodes, les visites multiples d'un même centre et la détectabilité des unités (Sudman et Kalton, 1986). De plus, lorsque les conditions suivantes sont réunies :

(i) on suppose un ordre aléatoire des N_l unités présentes presque avec certitude dans le centre *l* au moment de la visite de l'intervieweur,

(ii) la première unité rencontrée ou interviewée peut être considérée comme ayant été choisie au hasard,

(iii) le reste $(n_l - 1)$ des unités interviewées sont considérées comme ayant été sélectionnées systématiquement avec un pas unitaire,

l'ensemble de n_l unités sélectionnées crée un échantillon systématique circulaire d'après la technique de distance multiple de Singh et Singh (voir, par exemple, Hedayat et Sinha, 1991, p. 238). Comme l'échantillonnage systématique circulaire équivaut à l'échantillonnage aléatoire simple, les probabilités d'inclusion sont les mêmes selon ces deux plans de sondage (Särndal, Swensson et Wretman, 1992, p. 77) et l'ensemble de n_l unités échantillonnées dans le centre *l* selon notre description peut être considérée comme issue d'un échantillonnage aléatoire simple. Ajoutons que la condition suffisante $n_l \geq N_l/2 + 1$ pour que les probabilités conjointes d'inclusion soient toutes positives en découle directement, à savoir que la taille de l'échantillon d'un centre doit être *suffisamment* grande.

Sous un plan EAS, les probabilités d'inclusion se présentent ainsi :

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

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

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

On obtient un estimateur sans biais de la moyenne $\bar{Y}$ sous un plan EAS en substituant en (3) la première équation en (5) et donc en associant tous les centres (Mecatti, Migliorati et Thompson, 2001) :

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

En substituant les équations (5) en (4) après quelques simplifications, la variance (exacte) de l'estimateur (6) se dégage :

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

où $Y_r=\sum_{i\in D_r}y_i$.

Enfin, un estimateur de variance sans biais sous un plan EAS nous est donné par

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

De (8), on dérive facilement une estimation conservatrice de la variance en négligeant les corrections de population finie $\left(1-n_l/N_l\right)$.

À titre d'exemple, considérons les données réelles pour *n*=1 100 immigrants ayant fait l'objet en 2002 d'une sélection selon un plan EAS à Milan dans chacun des *L*=13 centres énumérés au tableau 1. Nous avons un ensemble de 8 191 profils possibles. Nous nous intéressons à la variable d'enquête *y*, qui est celle de l'âge. Au tableau 2, on trouvera les valeurs de pondération α_l des centres qui sont déduites des données de 2001, les tailles d'échantillon réparties proportionnellement de telle sorte que $n_l=n\alpha_l\big/\sum_l\alpha_l$ et les moyennes estimées des centres ajustées par la multiplicité selon l'équation (3). À noter que les valeurs de ces moyennes ajustées ne nous renseignent pas sur les moyennes non ajsutées correspondantes à cause du chevauchement; elles offrent néanmoins de l'intérêt pour les analyses sociodémographiques. Ainsi, on constate que les plus jeunes ont besoin de services d'accueil (centre 1) et de divertissement (centre 10) et que les adultes seraient assez intégrés à la société pour se livrer à des activités culturelles (centre 7).

Tableau 2 : Exemple de données réelles (Milan, 2002) : estimation des moyennes des centres ajustées par la multiplicité

Centre l	1	2	3	4	5	6	7	8	9	10	11	12	13
α_l	0,0992	0,0769	0,3772	0,3449	0,2084	0,1489	0,0893	0,3549	0,1737	0,2754	0,1663	0,4045	0,0099
n_l	40	31	152	139	84	60	36	143	70	111	67	163	4
$\bar{y}_l$	8,64	15,90	16,85	13,01	10,62	15,54	24,79	10,15	14,11	6,29	10,95	8,40	6,92

L'estimation de la moyenne d'âge de population selon l'équation (6) donne $\bar{y}_{SRS} = 32{,}74$. L'estimation conservatrice de la variance que l'on obtient en (8) en négligeant les corrections de population finie pour les centres du type 3 donne $\hat{v}(\bar{y}_{SRS}) = 0{,}3226$. Si on néglige l'ensemble des corrections de population finie, on obtient $\hat{v}(\bar{y}_{SRS}) = 0{,}5131$.

5. AUTRES PLANS DE SONDAGE

Outre le plan d'échantillonnage aléatoire simple de chaque centre, deux autres plans de sondage sont présentés.

5.1 Échantillonnage de centres à un degré

Pour obtenir une couverture suffisante de la population, il sera peut-être nécessaire de considérer un grand nombre L de centres avec de faibles valeurs de taille N_l. Dans ce cas, il semble plus convenable de sélectionner $n<L$ centres selon un plan de sondage donné, puis de procéder à une observation complète des unités des centres choisis. C'est ce que nous appellerons un échantillonnage de *centres à un degré*. Soit l^* un centre sélectionné. Selon l'échantillonnage de centres utilisant des règles de profil et de multiplicité, on connaît la valeur de taille N_{l^*} et la moyenne réelle $\tilde{\tilde{Y}}_{l^*}$ ajustée par la multiplicité selon l'équation (2), puisqu'au centre l^* appartiennent toutes les unités $i \in D_r$, ayant le même profil $\mathbf{u}_r$ avec $u_{rl^*} = 1$. Ainsi, un estimateur sans biais de la moyenne de population dans un plan de sondage général nous est donné par (Mecatti et Migliorati, 2003)

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

où π_{l^*} désigne les probabilités d'inclusion (du premier ordre) du centre sélectionné. De plus, la théorie standard de l'échantillonnage en grappes à un degré s'applique (voir, par exemple, Hedayat et Sinha, 1991, p. 204-205). Si les centres sont sélectionnés selon un plan EAS, par exemple, c'est-à-dire que $\pi_l = n/L$ pour tous les centres $l = 1 \cdots L$, l'estimateur (9) a la variance

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

et un estimateur de variance sans biais sous un plan EAS sera

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

5.2 Échantillonnage de centres à deux degrés

Si L est une valeur élevée et que plusieurs centres ont aussi des valeurs élevées de taille N_l, il pourrait être peu pratique de procéder à une observation complète des unités dans les centres choisis. Il serait aussi préférable de sous-échantillonner dans les centres choisis pour des raisons budgétaires ou lorsque les unités appartenant au même centre devraient être essentiellement homogènes pour ce qui est de la variable d'enquête.

L'échantillonnage de centres utilisant des règles de profil et multiplicité est applicable en combinant les résultats des sections 3 et 5.1. C'est ce que nous appellerons un échantillonnage de *centres à deux degrés*. Au premier degré, on sélectionne $n<L$ centres selon un plan de sondage donné avec des probabilités d'inclusion (du premier ordre) π_l. Soit l^* désignant un centre sélectionné. Au second degré, on échantillonne n_{l^*} unités dans le centre sélectionné l^* selon un plan de sondage peut-être différent. Un estimateur de la moyenne du centre ajustée par la multiplicité $\tilde{\bar{Y}}_{l^*}$ et sans biais selon le plan de sondage au second degré nous est donné par (Mecatti et Migliorati, 2003)

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

où $\pi_{ri}^{(l^*)}$ désigne les probabilités d'inclusion (du premier ordre) selon le plan de sondage au second degré pour l'unité $i \in D_r$ appartenant au centre l^* sélectionné au premier degré. De là, la théorie standard de l'échantillonnage en grappes à deux degrés devient applicable (voir, par exemple, Hedayat et Sinha, 1991, p. 209). Un estimateur sans biais de la moyenne de population est

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

avec la variance

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

où π_{lh} désigne les probabilités conjointes d'inclusion des centres $l \neq h$ et $V(\bar{y}_l)$, désigne la variance de l'estimateur (12) selon la forme générale donnée en (4). Enfin, si on pose $\pi_{lh} > 0$ pour tous les centres $l \neq h = 1 \cdots L$, un estimateur de variance sans biais nous sera donné par

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

6. OBSERVATIONS ET CONCLUSIONS

Concluons par quelques observations pour de futures recherches sur la technique d'échantillonnage de centres (EC).

Notons d'abord que, dans la théorie d'estimation déjà présentée, on pose l'hypothèse que les valeurs de pondération α_l des centres sont connues. Si on tient compte des particularités de la population immigrante italienne pour laquelle cette technique a été conçue, le traitement n'est pas difficile avec cette hypothèse, mais il faut par ailleurs considérer que la chose pourrait être peu pratique dans le cas d'autres populations difficiles à échantillonner. Dans des ouvrages scientifiques, on a proposé des méthodes d'estimation des α_l en utilisant les mêmes données d'échantillon lorsque N et N_l sont inconnus (Migliorati, 2002). Ainsi, on peut appliquer les résultats des sections

qui précèdent en utilisant des poids estimés dans les estimateurs proposés. Pour y arriver, il faudra pousser la recherche tant théorique qu'empirique pour étudier, par exemple, l'incidence des poids estimés sur l'absence de biais de l'estimateur de la moyenne et de l'estimateur de la variance.

Mentionnons ensuite que plusieurs analogies se dégagent entre l'échantillonnage de centres et l'échantillonnage à base double ou multiple. Si tous les centres disposent de listes et qu'on en connaît les valeurs de taille N_l, ces centres deviennent essentiellement des bases de sondage et les ensembles D_r d'unités au même profil sont des domaines selon la définition classique (Hartley, 1974). Dans ce cas, un estimateur sans biais du total $Y = N\bar{Y}$ découle directement de l'équation (3)

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

Il convient de noter que l'estimateur (16) est un estimateur sans biais de la taille N inconnue de la population par simple remplacement des valeurs y par l'unité. De plus, dans un échantillonnage à base duale, c'est-à-dire où L=2, il correspond à l'estimateur de Hartley pour base duale avec le choix le plus simple de pondération des données pour le domaine chevauchant, à savoir $1/2$. On a déjà comparé l'estimateur (16) aux grandes techniques concurrentes de l'échantillonnage à base duale (Mecatti, 2002). Les résultats de simulation indiquent que cet estimateur est une solution de rechange possible tant pour les propriétés d'inférence que pour les aspects pratiques. Toutefois, comme les formules (3) et (16) valent pour tout nombre de bases de sondage L>2, un complément de recherche s'impose compte tenu des possibilités de la technique d'échantillonnage de centres dans un contexte de base multiple.

REMERCIEMENTS

Je remercie Jon N.K. Rao pour les discussions et observations utiles qu'il a apportés au sujet de la technique d'échantillonnage de centres. Je remercie également Wesley Yung d'avoir examiné une version antérieure de ce document et d'y être lui aussi allé d'un certain nombre de suggestions utiles.

RÉFÉRENCES

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

Blangiardo, G. C., Migliorati, S. et Terzera, L. (2004), "Center Sampling: from Applicative Issues to Methodological Aspects", *Recueil de la XLII^{ème} réunion de la Société statistique de l'Italie*, pp. 377-388.

Eurostat (2000), "Push and Pull Factors of International Migration", *Country Report-Italy, 3/2000/E/n.5* Bruxelles: Office des publications officielles des Communautés européennes.

Hartley, H. O. (1974), "Multiple Frame Methodology and Selected Applications", *Sankhy$\bar{a}$*, Series C, 36, pp. 99-118.

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

Mecatti, F. (2002), "Center Sampling and Dual Frame surveys: comparisons among estimators for the total", document présenté à la XLI^{ème} réunion de la Société statistique de l'Italie, Milan, Italie.

Mecatti, F. et Migliorati, S. (2003), "Center Sampling: a strategy for elusive population surveys", document non publié.

Mecatti, F., Migliorati, S. et Thompson, S.K. (2001), "Center Sampling: Theory and Estimation", Rapport technique #01-06, Département de Statistique, Pennsylvania State University.

Migliorati, S. (2002), "New Developments in Center Sampling", document présenté à la XLI^{ème} réunion de la Société statistique de l'Italie, Milan, Italie.

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

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