Techniques d'enquête, juin 1998
Vol. 24, n° 1, pp. 21-31
Statistique Canada

Estimations à partir de bases de sondage de structure plusieurs à plusieurs

TERRI L. BYCZKOWSKI, MARTIN S. LEVY et DENNIS J. SWEENEY[1]

RÉSUMÉ

Pour les sondages, il doit idéalement y avoir correspondance de un à un entre les unités de la base de sondage et les éléments de la population-cible à l'étude. Dans bien des cas, toutefois, la base de sondage a une structure plusieurs à plusieurs, c'est-à-dire qu'une unité de la base de sondage peut être associée à de multiples éléments de la population-cible et, vice-versa, un élément de la population-cible peut être associé à de multiples unités de la base de sondage. C'est ce qui s'est produit dans le cadre d'une enquête sur les caractéristiques des immeubles, pour laquelle la base de sondage était constituée des adresses de voirie et les immeubles commerciaux formaient la population-cible. La base de sondage était complexe, car une même adresse pouvait correspondre à un seul immeuble, à plusieurs immeubles ou à une partie d'un immeuble. Nous présentons ici des estimateurs et des formules pour en calculer la variance, selon des plans d'échantillonnage aléatoire simple et stratifié, lorsque la base de sondage est de structure plusieurs à plusieurs.

MOTS-CLÉS: Bases de sondage imparfaites; erreurs de correspondance; enquête sur les caractéristiques des immeubles; pondération; échantillonnage aléatoire simple; échantillonnage aléatoire stratifié.

1. INTRODUCTION

Cette recherche fait suite à une étude qui avait été menée pour le compte d'une société de services publics, dans le but d'estimer diverses caractéristiques des immeubles commerciaux (population-cible) situés dans la région desservie par cette société. La liste des immeubles commerciaux ne pouvait être établie par dénombrement sur le terrain, en raison des coûts qu'aurait engendrés une telle méthode. On disposait toutefois d'une base de sondage constituée des adresses de voirie (c.-à-d. des adresses où il y avait un compteur), dont une des lacunes venait du rapport de plusieurs à plusieurs entre la base et la population-cible (immeubles commerciaux); ainsi, certaines unités de la base de sondage étaient associées à de multiples éléments de la population-cible et certains éléments de la population-cible étaient associés à un grand nombre d'unités de la base de sondage. En fait, plusieurs rapports entre les adresses et les immeubles étaient relativement complexes.

Cependant, un des avantages de cette base était qu'elle fournissait la consommation annuelle totale d'électricité pour chaque adresse de voirie; nous disposions donc d'une variable permettant la stratification efficace de la base des adresses. La superficie commerciale totale était une des caractéristiques importantes à mesurer; des études menées aux États-Unis ont en effet révélé que la consommation d'énergie est fonction à la fois de la taille du bâtiment et de l'activité qui y est menée. À titre d'exemple, la consommation est plus élevée dans les immeubles utilisés pour la prestation de soins de santé ou la vente d'aliments, alors qu'elle est plus faible dans les immeubles utilisés pour des cérémonies religieuses ou des assemblées publiques. Il existe également une corrélation entre la consommation d'énergie et la taille du bâtiment, même lorsque l'utilité du bâtiment n'est pas connue, comme c'était le cas ici (U.S. Department of Energy 1992).

Les bases de sondage imparfaites ont fait l'objet de nombreuses études, dont on peut en trouver des résumés détaillés dans Kish (1965), Wright et Tsao (1983) et Lessler et Kalsbeek (1992). D'autres ouvrages traitent de l'échantillonnage par multiplicité, où la base de sondage est conçue de manière à avoir une structure de plusieurs à plusieurs. Dans ce dernier cas, des imperfections sont introduites dans la base de sondage, afin de recueillir plus efficacement de l'information sur les occurrences qui sont rares dans une population (Birnbaum et Sirken 1965, Sirken 1972a,b et Casady et Sirken 1980). Hansen, Hurwitz et Madow (1953a,b) proposent un estimateur à utiliser avec les bases de sondage qui ont une structure plusieurs à un, c'est-à-dire où les éléments de la population sont représentés plusieurs fois dans la base. Cet estimateur est celui qu'a choisi le National Agricultural Statistics Service (NASS) pour ses enquêtes (Musser 1993), lorsque la base a une structure plusieurs à un. Bandyopadhyay et Adhikari (1993) proposent quant à eux des estimateurs pour un quotient, la moyenne d'une population et le total d'une population, lorsque le degré de répétition dans la base de sondage est inconnu. Cependant, ces estimateurs ne peuvent être utilisés que dans les cas d'échantillonnage aléatoire simple avec base de sondage de plusieurs à un.

La documentation propose également deux méthodes pour estimer les caractéristiques de la population à partir

[1] Terri L. Byczkowski, Institute for Policy Research, Martin S. Levy et Dennis J. Sweeney, Department of Quantitative Analysis and Operations Management, University of Cincinnati, Cincinnati, OH 45221, U.S.A.

d'une base de sondage de structure plusieurs à plusieurs. Mentionnons d'abord l'estimateur de Horvitz-Thompson, (1952), qui fournit des estimations sans biais de la moyenne et du total d'une population lorsqu'il existe plusieurs probabilités de sélection. Musser (1993) montre comment calculer les bonnes probabilités d'inclusion pour les éléments de la population sélectionnés par échantillonnage aléatoire simple à partir d'une base de plusieurs à un. Cependant, la méthode de Musser peut également être appliquée au calcul des probabilités d'inclusion pour des éléments de population dans un échantillon aléatoire simple prélevé d'une base de sondage de structure plusieurs à plusieurs. En deuxième lieu, Lavallée (1995) a adapté la méthode à poids partagés – utilisée pour les enquêtes longitudinales – pour l'appliquer aux bases de sondage plusieurs à plusieurs.

Le but du présent article est de proposer une autre méthode pour estimer le total, les chiffres et la moyenne de la population, avec des bases de sondage de structure plusieurs à plusieurs, selon des plans d'échantillonnage aléatoire simple et stratifié. Nous calculons également les expressions de la variance de ces estimateurs. Les résultats que nous présentons n'ont pas seulement un intérêt intrinsèque; en effet, les expressions de la variance des estimateurs sont essentielles à l'étude des effets des imperfections de correspondance inhérentes aux bases de sondage plusieurs à plusieurs sur la précision de ces estimations.

Nous présentons à la section 2 les estimations obtenues par échantillonnage aléatoire simple sans remise (ÉASSR). Nous y décrivons également la méthode d'échantillonnage avec laquelle ces estimateurs sont applicables, puis présentons un résultat du biais et proposons des façons d'en exprimer la variance.

À la section 3, certains de ces résultats sont appliqués à l'échantillonnage aléatoire stratifié. Enfin, à la section 4, nous tirons quelques conclusions, discutons des limites de la méthode et proposons des suggestions pour de futurs projets de recherche.

2. BASES DE SONDAGE PLUSIEURS À PLUSIEURS POUR L'ÉCHANTILLONNAGE ALÉATOIRE SIMPLE

Il est utile de représenter sous forme de graphique la relation entre la base de sondage et la population-cible. Les unités d'échantillonnage dans la base de sondage et les éléments de la population-cible forment les deux ensembles de noeuds; des arcs relient les unités d'échantillonnage aux éléments de la population-cible. Ces arcs illustrent la structure du rapport entre la base de sondage et la population-cible. La figure 2.1 présente un exemple d'une base de sondage et d'une population-cible avec un rapport de plusieurs à plusieurs. Il y a 7 unités d'échantillonnage dans la base, 6 éléments dans la population-cible et 10 liens (arcs) entre les unités d'échantillonnage et les éléments de la population. Cette structure de plusieurs à plusieurs est donc représentée par un graphique formé de 13 noeuds et de 10 arcs. Dans la présente analyse, nous supposons que chaque élément de la population est relié aux unités de la base par au moins un arc et que chaque unité de la base est liée aux éléments de la population, là aussi par au moins un arc.

Établissons maintenant quelques notations. Il nous paraît commode d'identifier les unités de la base et les éléments de la population par leurs indices respectifs. Supposons que $F = \{1, 2, ..., N\}$ représente l'ensemble des indices pour N unités d'échantillonnage et que $T = \{1, 2, ..., M\}$ représente l'ensemble des indices pour les M éléments de la population-cible. Un arc peut être représenté comme étant une paire ordonnée, dont le premier élément provient de F et le deuxième, de T. On dit qu'un élément de la population k dans T est représenté par l'unité d'échantillonnage j dans F s'il y est relié par un arc désigné (jk). Ceci signifie que, lorsque j est dans l'échantillon, il existe une probabilité différente de zéro de recueillir des données à partir de l'élément de la population k. Nous représenterons par y_k la mesure de l'élément de la population k dans T qui nous intéresse.

Nous décrirons maintenant la technique d'échantillonnage en vertu de laquelle les estimateurs proposés ici peuvent s'appliquer. Supposons que n unités de la base sont sélectionnées dans F, par ÉASSR. Le nombre d'*éléments de la population* inclus dans l'échantillon et mesurés dépend toutefois de la nature du lien qui existe entre les unités de la base de sondage et les éléments de la population.

Dans le cas de l'ÉASSR, quatre scénarios sont possibles lorsqu'une unité est sélectionnée. Dans le premier scénario, l'unité de la base ne correspond qu'à un seul élément de la population (structure un à un). Ici, l'enquêteur recueillerait des données uniquement sur l'élément de la population qui correspond à l'unité sélectionnée de la base (voir unité 1 de la base de sondage, à la figure 2.1).

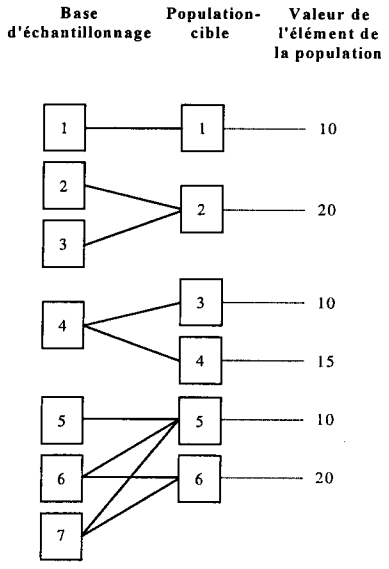

Figure 2.1 Exemple de la correspondance entre la base de sondage et la population-cible

Dans le deuxième scénario, plusieurs unités de la base de sondage correspondent à un élément de la population (structure plusieurs à un). À la figure 2.1, les unités 2 et 3 de la base de sondage correspondent à un seul élément de la population, le 2. Dans ce dernier cas, si les unités 2 ou 3, ou les deux, de la base de sondage sont incluses dans l'échantillon, on obtient alors de l'information sur l'élément 2 de la population. Il est donc possible que cet élément de la population (2) apparaisse jusqu'à deux fois dans l'échantillon – et aussi comme enregistrement, dans le fichier de données utilisé pour faire les estimations.

Dans le troisième scénario, une unité de la base de sondage correspond à plus d'un élément de la population (structure un à plusieurs). Toujours à la figure 2.1, l'unité 4 correspond aux éléments de la population 3 et 4. Ici, un seul élément de la population (3 ou 4) est choisi par une méthode de *randomisation* indépendante du choix des unités de la base de sondage. Cette méthode a été justifiée par des motifs économiques, car la collecte des données nécessitait de longues interviews sur place menées par des personnes ayant une formation technique. Nous présumons ici que les probabilités de randomisation sont égales. Cependant, toute autre probabilité différente de zéro pourrait également être utilisée (par ex. probabilité proportionnelle à la taille).

Le quatrième scénario prévoit une structure de plusieurs à plusieurs, laquelle est illustrée par les unités 5, 6 et 7 de la base de sondage et les éléments 5 et 6 de la population, à la figure 2.1. Comme ces cas complexes sont en fait une combinaison des scénarios 2 et 3 qui précèdent, les mêmes règles d'échantillonnage s'appliquent. Ainsi, si l'unité 5 est choisie, c'est l'élément de la population 5 qui est mesuré. Si l'unité 6 est choisie, un seul des deux éléments de la population 5 ou 6 est choisi au hasard et mesuré.

2.1 Total d'une population

2.1.1 Estimateur du total d'une population

Une base de sondage de plusieurs à plusieurs donne lieu à diverses probabilités de sélection. Les estimateurs proposés ici sont basés sur une méthode de pondération et constituent un prolongement de ceux présentés par Hansen et coll. (1953a p. 62-64). Les estimateurs proposés par ces auteurs, de même que les formules pour en calculer la variance, se limitent à la structure plusieurs à un; nous avons prolongé ces estimateurs pour les appliquer à une structure plusieurs à plusieurs.

Dans un plan d'ÉASSR d'effectifs n, supposons que $J_1, ..., J_n$ représentent des variables aléatoires de telle sorte que $J_i = j$ si le i-ième prélèvement mène à la sélection de l'unité j dans F. Donc $\Pr(J_i = j) = 1/N$ pour j dans F et $i = 1, ..., n$. Supposons maintenant que $K_1, ..., K_n$ représentent des variables aléatoires de sorte que $K_i = k$ si le i-ième prélèvement dans F est suivi de la sélection de k dans T. Nous pouvons maintenant passer au prélèvement d'un échantillon aléatoire d'arcs $\{(J_1K_1), ..., (J_nK_n)\}$ dont la distribution de probabilité conjointe est déterminée à la fois par le plan d'échantillonnage ÉASSR et la randomisation subséquente (s'il y a lieu) pour choisir un élément dans T. (J_iK_i) a une probabilité marginale représentée par $\Pr\{(J_iK_i) = (jk)\} = (1/N)s_{jk}$, où s_{jk} est la probabilité conditionnelle calculée par $s_{jk} = \Pr(K_i = k \mid J_i = j)$. En d'autres mots, s_{jk} est la probabilité conditionnelle de sélectionner l'élément de population k dans T lorsque l'unité de la base de sondage j dans F est choisie. Ces probabilités conditionnelles, désignées probabilités d'arc, sont illustrées pour la figure 2.1 au tableau 2.1.

Tableau 2.1
Probabilités d'arc pour la figure 2.1

Arc jk	1,1	2,2	3,2	4,3	4,4	5,5	6,5	6,6	7,5	7,6
s_{jk}	1	1	1	1/2	1/2	1	1/2	1/2	1/2	1/2

Pour k dans T, supposons que U_k représente l'ensemble des unités dans F dont les arcs mènent à k dans T. Supposons également que $s_k = \sum_{j \in U_k} s_{jk}$. Si nous reprenons le langage de Hansen et coll. (1953a p. 62-64), sur lequel s'appuie notre raisonnement, s_k est le *facteur de pondération* de l'élément de la population k dans T. Les facteurs de pondération pour la figure 2.1 sont indiqués au tableau 2.2.

Tableau 2.2
Calcul des facteurs de pondération (s_k) des éléments de la population pour la figure 2.1

k	1	2	3	4	5	6
(s_k)	1	2	1/2	1/2	2	1

Les probabilités d'arc et les facteurs de pondération sont utilisés pour calculer les probabilités marginales de K_i, nommé $\Pr(K_i = k) = \sum_{j \in U_k} (1/N)s_{jk} = (1/N)s_k$, à savoir lorsque k est dans T et $i = 1, ..., n$. De toute évidence, le calcul des probabilités d'arc est l'étape cruciale dans l'élaboration des facteurs de pondération appropriés pour les données recueillies. Ce calcul dépend de la détermination adéquate de la structure de graphe pour chaque unité d'échantillonnage choisie: dans un sous-graphe connexe maximal (CM). Un sous-graphe connexe désigne un sous-ensemble de noeuds reliés par une série d'arcs. Maximal signifie qu'aucun noeud à l'extérieur du sous-ensemble n'est relié à un noeud qui appartient au sous-ensemble. Il y a 4 sous-graphes CM à la figure 2.1; chacun représente une structure base – population différente, à savoir les structures un à un, plusieurs à un, un à plusieurs et plusieurs à plusieurs.

Il n'est pas nécessaire de connaître la structure du graphe entier pour définir les estimateurs. Il suffit seulement de connaître la structure des sous-graphes CM auxquels appartiennent les unités *échantillonnées* de la base.

Nous faisons les observations suivantes au sujet de s_k et s_{jk} : i) $s_k = W$ indique que la probabilité de sélectionner l'élément de population k au i-ième prélèvement est de W fois celle d'un élément dont le facteur de pondération est un; ii) $0 < s_k \le N$, $k = 1, ..., M$; iii) $0 < s_{jk} \le 1$, $j \in U_k$ et $k = 1, ..., M$; iv) avec la structure un à plusieurs, $s_{jk} = s_k$; v) avec la structure plusieurs à un, $s_{jk} = 1$ pour tous les k et vi) $\sum_{k=1}^{M} \sum_{j=1}^{N} s_{jk} = N$.

Maintenant, supposons que $x_1, ..., x_M$ représentent les valeurs pondérées associées aux indices dans T, c'est-à-dire supposons que $x_k = y_k/s_k$. Définissons les variables aléatoires $x_{K_1}, ..., x_{K_n}$, associées respectivement aux prélèvements 1 à n à partir de F, de sorte que x_{K_i} prend la valeur x_k si $K_i = k$. Nous pouvons alors écrire

$$\mathrm{E}(x_{K_i}) = \sum_{k=1}^{M} x_k \Pr(K_i = k) = \frac{1}{N}\sum_{k=1}^{M} \frac{y_k}{s_k} s_k = \frac{Y}{N}, \qquad (2.1)$$

où $Y = \sum_{k=1}^{M} y_k$ est le véritable total de la population. Nous utilisons comme estimateur du total de la population, selon un plan ÉASSR avec structure plusieurs à plusieurs,

$$\hat{Y} = \frac{N}{n}\sum_{i=1}^{n} x_{K_i}. \qquad (2.2)$$

Si l'on utilise (2.1), il s'ensuit que

$$\mathrm{E}(\hat{Y}) = \mathrm{E}\left(\frac{N}{n}\sum_{i=1}^{n} x_{K_i}\right) = \frac{N}{n}\sum_{i=1}^{n} \mathrm{E}(x_{K_i}) = \frac{N}{n} n \frac{Y}{N} = Y.$$

Nous obtenons alors

Théorème 2-1: L'estimateur (2.2) du total d'une population, utilisé avec un plan d'ÉASSR, est sans biais.

À partir de la figure 2.1, nous présentons maintenant un exemple simple de l'utilisation de cet estimateur. Supposons qu'un échantillon aléatoire simple formé de quatre unités a été sélectionné de la base de sondage illustrée par la figure 2.1 (2, 3, 4 et 7), ce qui a mené par la suite à la sélection des éléments de population 2, 4 et 5. L'estimateur du total de la population,

$$\hat{Y} = \frac{N}{n}\sum_{i=1}^{4} x_{K_i}, \text{ a la valeur } \frac{7}{4}\left[\frac{20}{2} + \frac{20}{2} + \frac{15}{(1/2)} + \frac{10}{2}\right] = \frac{385}{4}.$$

L'estimateur qui précède peut également être utilisé pour obtenir les chiffres de population. Nous pourrions ainsi estimer la taille de la population-cible en supposant que $y_k = 1$ pour tous les k. Nous pourrions également estimer le nombre d'éléments de la population qui possèdent certaines caractéristiques, en supposant que $y_k = 1$ pour les éléments de la population qui présentent les caractéristiques qui nous intéressent et que $y_k = 0$ pour ceux qui n'ont pas ces caractéristiques.

2.1.2 Variance de l'estimateur du total d'une population

Il convient en premier lieu de définir certains autres termes et notations utilisés dans cette section. Supposons que P représente l'ensemble de toutes les paires non ordonnées d'arcs. Une paire non ordonnée d'arcs est dite *inadmissible* si les deux éléments ne peuvent être inclus dans un échantillon. Supposons que $Q = \{j \text{ dans } F\text{: plus d'un arc émerge de } j\}$. Alors $R' = \{[jk, jk']: j \in Q \text{ et } k \neq k'\}$ représente l'ensemble des paires non ordonnées inadmissibles d'arcs. L'ensemble des paires non ordonnées *admissibles* d'arcs est la série complémentaire $R^* = P \setminus R'$.

Pour illustrer ceci, examinons la figure 2.1. En vertu de la méthode d'échantillonnage utilisée, si l'unité 4 de la base de sondage est sélectionnée, un seul des deux éléments de population 3 ou 4 peut être inclus dans l'échantillon. Par conséquent, {[4,3][4,4]} est une paire non ordonnée inadmissible d' arcs. Les autres paires d'arcs non ordonnées et inadmissibles à la figure 2.1 sont {[6,5][6,6]} et {[7,5][7,6]}. Par conséquent, R' = {[4,3][4,4], [6,5][6,6], [7,5][7,6]}.

Théorème 2-2: La variance de l'estimateur (2.2) est

$$\mathrm{V}(\hat{Y}) = \frac{N}{n}\left[\sum_{k=1}^{M} \frac{y_k^2}{s_k} + 2\frac{(n-1)}{(N-1)}\sum \sum_{[jk,j'k']\in R^*} \left(\frac{y_k s_{jk}}{s_k}\frac{y_{k'} s_{j'k'}}{s_{k'}}\right)\right] - Y^2, \qquad (2.3)$$

où la double somme englobe toutes les paires non ordonnées *admissibles* d'arcs $[jk, j'k']$.

Preuve:

$$\mathrm{V}(\hat{Y}) = \mathrm{E}\left[\left(\frac{N}{n}\sum_{i=1}^{n} x_{K_i}\right)^2\right] - Y^2$$

$$= \frac{N^2}{n^2}\mathrm{E}\left[\left(\sum_{i=1}^{n} x_{K_i}\right)^2\right] - Y^2. \qquad (2.4)$$

Maintenant,

$$\mathrm{E}\left[\left(\sum_{i=1}^{n} x_{K_i}\right)^2\right] = \sum_{i=1}^{n} \mathrm{E}\left(x_{K_i}^2\right) + 2\mathrm{E}\left(\sum\sum_{i<i'} \left(x_{K_i} x_{K_{i'}}\right)\right). \qquad (2.5)$$

On peut écrire

$$\mathrm{E}\left(x_{K_i}^2\right) = \sum_{k=1}^{M}\left[x_k^2 \Pr(K_i = k)\right] = \sum_{k=1}^{M} \frac{y_k^2}{s_k^2}\frac{s_k}{N} = \frac{1}{N}\sum_{k=1}^{M} \frac{y_k^2}{s_k}. \qquad (2.6)$$

Comme nous l'avons indiqué à la section 2.1, nous pouvons sélectionner un échantillon d'arcs qui mène ensuite à la sélection des éléments de la population. Chaque arc (jk) est associé à une valeur $x_k = y_k/s_k$ de l'élément de la population k à sa destination. Nous pouvons donc récrire la double sommation en (2.5) comme étant le cumul des paires non ordonnées admissibles d'arcs, R^*.

$$2\mathrm{E}\left(\sum_{i'}\sum_{i<i'} \left(x_{K_i} x_{K_{i'}}\right)\right) =$$

$$2\binom{n}{2}\sum\sum_{[jk,j'k']\in R^*} \left[(x_k x_{k'})\Pr(K_i = k, K_{i'} = k')\right]. \qquad (2.7)$$

Compte tenu de l'indépendance de la randomisation et du choix des unités de la base de sondage:

$$\Pr(\text{sélection } [jk, j'k'] \text{ dans } R^*) = \Pr(\text{sélection } \{j, j'\} \text{ dans } F)$$

$$\Pr(\text{sélection } [jk, j'k'] \text{ dans } R^* \mid \text{sélection } \{j, j'\} \text{ dans } F) = \frac{1}{\binom{N}{2}} s_{jk} s_{j'k'}.$$

Si l'on substitue ceci dans (2.7), on obtient alors

$$n(n-1) \sum \sum_{[jk,j'k'] \in R^*} \left[(x_k x_{k'}) \frac{1}{\binom{N}{2}} s_{jk} s_{j'k'} \right] =$$

$$\frac{2n(n-1)}{N(N-1)} \sum \sum_{[jk,j'k'] \in R^*} \left[\left(\frac{y_k s_{jk}}{s_k} \frac{y_{k'} s_{j'k'}}{s_{k'}} \right) \right]. \quad (2.8)$$

Si l'on introduit maintenant (2.6) et (2.8) dans (2.5), on obtient,

$$\mathrm{E}\left[\left(\sum_{i=1}^{n} x_{K_i} \right)^2 \right] = \frac{n}{N} \left[\sum_{k=1}^{M} \frac{y_k^2}{s_k} + \right.$$

$$\left. \frac{2(n-1)}{(N-1)} \sum \sum_{[jk,j'k'] \in R^*} \left(\frac{y_k s_{jk}}{s_k} \frac{y_{k'} s_{j'k'}}{s_{k'}} \right) \right]. \quad (2.9)$$

Enfin, si l'on introduit(2.9) dans (2.4), nous obtenons le résultat en (2.3).

L'équation (2.3) est une généralisation de la formule élaborée par Bandyopadhyay et Adhikari (1993) pour calculer la variance de l'estimation du total d'une population, avec une structure plusieurs à plusieurs. On peut voir que l'équation (2.3) correspond à leur formule, lorsque la base de sondage est limitée à une structure plusieurs à un.

Corollaire 2-1: Voici une autre façon d'exprimer la formule de la variance dans le **Théorème 2-2:**

$$\mathrm{V}(\hat{Y}) = \frac{N}{n} \left[\sum_{k=1}^{M} \frac{y_k^2}{s_k} + \frac{(n-1)}{(N-1)} \left(\left(\sum_{jk} \frac{y_k}{s_k} s_{jk} \right)^2 - \right. \right.$$

$$\left. \left. \sum_{jk} \left(\frac{y_k}{s_k} s_{jk} \right)^2 - 2 \sum \sum_{[jk,jk'] \in R'} \frac{y_k s_{jk}}{s_k} \frac{y_{k'} s_{jk'}}{s_{k'}} \right) \right] - Y^2.$$

Preuve:
Si nous écrivons

$$\left(\sum_{jk} \frac{y_k s_{jk}}{s_k} \right)^2 = \sum_{jk} \left(\frac{y_k s_{jk}}{s_k} \right)^2 +$$

$$2 \sum_{[jk,j'k'] \in R^*} \sum \frac{y_k s_{jk}}{s_k} \frac{y_{k'} s_{j'k'}}{s_{k'}} + 2 \sum_{[jk,jk'] \in R'} \sum \frac{y_k s_{jk}}{s_k} \frac{y_{k'} s_{jk'}}{s_{k'}}.$$

Il s'ensuit que:

$$\sum_{[jk,j'k'] \in R^*} \sum \frac{y_k s_{jk}}{s_k} \frac{y_{k'} s_{j'k'}}{s_{k'}} = \frac{1}{2} \left(\sum_{jk} \frac{y_k s_{jk}}{s_k} \right)^2 -$$

$$\frac{1}{2} \sum_{jk} \left(\frac{y_k s_{jk}}{s_k} \right)^2 - \sum_{[jk,jk'] \in R'} \sum \frac{y_k s_{jk}}{s_k} \frac{y_{k'} s_{jk'}}{s_{k'}}.$$

Nous obtenons le résultat en remplaçant l'expression qui précède dans (2.3).

Cette formule est plus simple au plan computationnel. À noter que l'équation (2.3) exige que l'on fasse la somme du terme

$$\left(\frac{y_k s_{jk}}{s_k} \frac{y_{k'} s_{j'k'}}{s_{k'}} \right)$$

pour toutes les paires non ordonnées admissibles d'arcs (R^*), alors qu'avec cette autre formule il suffit de faire la totalisation des paires d'arcs inadmissibles (R'). Dans la plupart des scénarios pratiques, le nombre de paires d'arcs admissibles sera bien supérieur au nombre de paires inadmissibles.

2.2 Moyenne d'une population

2.2.1 Estimateur de la moyenne d'une population

L'estimateur de la moyenne d'une population, qui est présenté ici, est un prolongement de l'estimateur présenté par Hansen et coll. (1953a), ici appliqué à la structure plusieurs à plusieurs.

En rapport avec les n prélèvements de F, définissons les variables aléatoires s_{K_i} et $z_{K_i} = 1/s_{K_i}$, de sorte que s_{K_i} a la valeur s_k si $K_i = k$ pour $i = 1, ..., n$ et $k = 1, ..., M$. L'estimateur de la moyenne d'une population,

$$\bar{Y} = \frac{1}{M} \sum_{k=1}^{M} y_k,$$

devient, avec un plan d'ÉASSR et une structure de plusieurs à plusieurs:

$$\hat{\bar{Y}} = \frac{\sum_{i=1}^{n} x_{K_i}}{\sum_{i=1}^{n} z_{K_i}}. \tag{2.10}$$

2.2.2 Erreur quadratique moyenne (EQM) de la moyenne d'une population

L'estimateur de la moyenne d'une population est biaisé, parce qu'il s'agit d'un estimateur par quotient. Cependant, on sait très bien que ce biais devient négligeable avec de gros échantillons et que le biais est d'ordre $1/n$ (Cochran 1977, p. 160).

Notre approximation de l'erreur quadratique moyenne exige une sommation de R^{**} – l'ensemble de toutes les paires d'arcs *ordonnées admissibles*. Par conséquent, si $[jk, j'k'] \in R^{*}$, alors $[jk, j'k'] \in R^{**}$ et $[j'k', jk] \in R^{**}$.

Pour calculer la valeur approximative de l'erreur quadratique moyenne de l'estimateur (2.10), nous utilisons

$$\mathrm{EQM}(\hat{\bar{Y}}) \approx \frac{M^2}{nN\left(\sum_{k=1}^{M} \frac{1}{s_k}\right)^2}$$

$$\left[\left(\sum_{k=1}^{M} \frac{y_k^2}{s_k} + \frac{2(n-1)}{(N-1)} \sum \sum_{[jk,j'k'] \in R^{*}} \frac{y_k s_{jk}}{s_k} \frac{y_{k'} s_{j'k'}}{s_{k'}}\right)\right.$$

$$- 2\bar{Y}\left(\sum_{k=1}^{M} \frac{y_k}{s_k} + \frac{(n-1)}{(N-1)} \sum \sum_{[jk,j'k'] \in R^{**}} \frac{y_k s_{jk}}{s_k} \frac{s_{j'k'}}{s_{k'}}\right)$$

$$\left. + \bar{Y}^2\left(\sum_{k=1}^{M} \frac{1}{s_k} + \frac{2(n-1)}{(N-1)} \sum \sum_{[jk,j'k'] \in R^{*}} \frac{s_{jk}}{s_k} \frac{s_{j'k'}}{s_{k'}}\right)\right]. \tag{2.11}$$

Pour justifier cette approximation, supposons que

$$\bar{x} = \frac{\sum_{i=1}^{n} x_{K_i}}{n}, \quad \bar{z} = \frac{\sum_{i=1}^{n} z_{K_i}}{n} \quad \text{et} \quad \bar{Z} = \frac{\sum_{k=1}^{M} \frac{1}{s_k}}{M}.$$

Comme $\hat{\bar{Y}}$ est un ratio de deux estimations, l'approximation bien connue de l'erreur quadratique moyenne (Cochran 1977, p. 32-33) peut être utilisée:

$$\mathrm{EQM}(\hat{\bar{Y}}) = \mathrm{E}\left(\frac{\bar{x} - \bar{Y}\bar{z}}{\bar{z}}\right)^2 \approx \mathrm{E}\left(\frac{\bar{x} - \bar{Y}\bar{z}}{\bar{Z}}\right)^2 =$$

$$\frac{1}{\bar{Z}^2}\left[\mathrm{E}(\bar{x}^2) - 2\bar{Y}\mathrm{E}(\bar{z}\bar{x}) + \bar{Y}^2\mathrm{E}(\bar{z}^2)\right] =$$

$$\frac{M^2}{n^2\left(\sum_{j=1}^{M} z_j\right)^2}\left[\mathrm{E}\left(\sum_{i=1}^{n} x_{K_i}\right)^2 - \right.$$

$$\left. 2\bar{Y}\mathrm{E}\left(\sum_{i=1}^{n} x_{K_i} \sum_{i=1}^{n} z_{K_i}\right) + \bar{Y}^2\mathrm{E}\left(\sum_{i=1}^{n} z_{K_i}\right)^2\right]. \tag{2.12}$$

La première espérance dans (2.12) est tout simplement (2.9). Ensuite, l'utilisation de l'équation (2.1) dans le terme central de (2.12) donne

$$\mathrm{E}\left(\sum_{i=1}^{n} x_{K_i} \sum_{i=1}^{n} z_{K_i}\right) = \mathrm{E}\left(\sum_{i=1}^{n} x_{K_i} \frac{1}{s_{K_i}}\right) +$$

$$\mathrm{E}\left(\sum_{\substack{i=1 \\ i \neq i'}}^{n} \sum_{i'=1}^{n} x_{K_i} \frac{1}{s_{K_{i'}}}\right) = \frac{n}{N} \sum_{k=1}^{M} \frac{y_k}{s_k} + \mathrm{E}\left(\sum_{\substack{i=1 \\ i \neq i}}^{n} \sum_{i'=1}^{n} x_{K_i} \frac{1}{s_{K_{i'}}}\right).$$

Si l'on utilise (2.7) et (2.9), on obtient

$$\mathrm{E}\left(\sum_{\substack{i=1 \\ i \neq i}}^{n} \sum_{i'=1}^{n} x_{K_i} \frac{1}{s_{K_{i'}}}\right) = n(n-1)\mathrm{E}\left(x_{K_i} \frac{1}{s_{K_{i'}}}\right) =$$

$$n(n-1) \sum_{[jk,j'k'] \in R^{**}} \sum \left[\left(\frac{y_k}{s_k} \frac{1}{s_{k'}}\right) \Pr\left(x_{K_i} = \frac{y_k}{s_k}, \frac{1}{s_{K_{i'}}} = \frac{1}{s_{k'}}\right)\right] =$$

$$n(n-1) \sum_{[jk,j'k'] \in R^{**}} \sum \frac{y_k}{s_k} \frac{1}{s_{k'}} \left(\frac{1}{N(N-1)} s_{jk} s_{j'k'}\right) =$$

$$\frac{n(n-1)}{N(N-1)} \sum_{[jk,j'k'] \in R^{**}} \sum \frac{y_k s_{jk}}{s_k} \frac{s_{j'k'}}{s_{k'}}.$$

À noter que la double somme englobe toutes les paires *ordonnées* et admissibles d'arcs. Par conséquent,

$$\mathrm{E}\left(\sum_{i=1}^{n} x_{K_i}\frac{1}{s_{K_i}}\right) + \mathrm{E}\left(\sum_{\substack{i=1\\ i\neq i'}}^{n}\sum_{i'=1}^{n} x_{K_i}\frac{1}{s_{K_{i'}}}\right) =$$

$$\frac{n}{N}\sum_{k=1}^{M}\frac{y_k}{s_k} + \frac{n(n-1)}{N(N-1)}\sum_{[jk,j'k']\in R^{**}}\sum\frac{y_k s_{jk}}{s_k}\frac{s_{j'k'}}{s_{k'}} =$$

$$\frac{n}{N}\left(\sum_{k=1}^{M}\frac{y_k}{s_k} + \frac{(n-1)}{(N-1)}\sum_{[jk,j'k']\in R^{**}}\sum\frac{y_k s_{jk}}{s_k}\frac{s_{j'k'}}{s_{k'}}\right).$$

Enfin, comme (2.1),

$$\mathrm{E}\left(\sum_{i=1}^{n} z_{K_i}\right)^2 = \mathrm{E}\left(\sum_{i=1}^{n}\frac{1}{s_{K_i}}\right)^2 =$$

$$\frac{n}{N}\left(\sum_{k=1}^{M}\frac{1}{s_k} + \frac{2(n-1)}{(N-1)}\sum\sum_{[jk,j'k']\in R^{*}}\frac{S_{jk}}{s_k}\frac{s_{j'k'}}{s_{k'}}\right).$$

Lorsqu'on introduit ces espérances dans l'équation (2.12), on obtient alors (2.11).

3. ESTIMATEURS POUR DES BASES DE STRUCTURE PLUSIEURS À PLUSIEURS, PAR ÉCHANTILLONNAGE ALÉATOIRE STRATIFIÉ

3.1 Introduction

Nous présentons dans cette section des estimateurs pour les chiffres, la moyenne et le total d'une population, avec une structure plusieurs à plusieurs selon un plan d'échantillonnage aléatoire stratifié. Il convient toutefois de décrire d'abord la méthode d'échantillonnage en vertu de laquelle ces estimations sont appropriées. La figure 3.1 présente un exemple qui sera utilisé tout au long de cette section.

3.2 Méthode d'échantillonnage

Les scénarios décrits pour l'ÉASSR sont également utilisés pour l'échantillonnage aléatoire stratifié. Il y a toutefois quelques problèmes additionnels qui peuvent survenir dans ce dernier cas.

Revenons à l'étude sur les caractéristiques des immeubles, qui a justifié la conduite de la présente recherche. Supposons que la taille du bâtiment est la valeur de l'élément de la population à la figure 3.1 et que la consommation d'électricité associée à l'adresse de voirie est la variable de stratification. Comme il existe un rapport de plusieurs à plusieurs entre la base de sondage (adresses de voirie) et la population-cible (immeubles commerciaux), les problèmes suivants sont venus s'ajouter à ceux mentionnés à la section 2.1:

1. Erreur de stratification: Par exemple, l'unité 2 de la base de sondage (adresses de voirie) dans la strate 1 semblait correspondre à un grand immeuble, en raison de la forte consommation d'électricité qui y était associée; cette unité a donc été placée dans la première strate. Les données recueillies ont toutefois révélé que l'adresse correspondait en fait à deux petits immeubles (éléments de la population 2 et 3). Dans un autre exemple, les unités 5 et 6 de la base de sondage, dans la strate 2, semblaient être deux petits immeubles et ont donc été placées dans la deuxième strate. Or l'élément de la population 7 qui y correspond est un immeuble unique ayant deux adresses.
2. Chevauchement: Par exemple, les unités 3 et 4 de la base de sondage, dans la strate 1, de même que les unités 1 et 2 dans la strate 2, ont toutes une adresse différente et figurent donc dans la base comme deux petits et deux grands immeubles. Or les données recueillies révèlent que les quatre adresses ne correspondent en fait qu'à un seul bâtiment (par ex. un mail linéaire). Dans ce dernier cas, non seulement y a-t-il erreur de stratification, mais également les unités de la base de sondage associées à un même immeuble ne sont pas toutes incluses dans la même strate. En d'autres mots, un même élément de la population (un immeuble) chevauche plusieurs strates.

Dans la section qui suit, nous élaborons des estimateurs pour le total et les chiffres de population et démontrons que ces estimateurs sont sans biais, malgré l'erreur de stratification et le chevauchement. Cependant, comme c'est habituellement le cas, l'erreur de stratification augmente la variance des estimations. En outre, dans la mesure où le chevauchement produit une erreur de stratification, celui-ci augmente également la variance des estimations.

3.3 Totaux et chiffres de population

3.3.1 Estimateur du total d'une population

L'estimateur présenté ici s'appuie sur une méthode de pondération qui consiste à prolonger l'estimateur proposé par Hansen et coll. (1953a, p. 62-64), pour l'appliquer à l'échantillonnage aléatoire stratifié avec base de sondage plusieurs à plusieurs.

Supposons que F a été divisé en L strates exhaustives s'excluant mutuellement $F_1, ..., F_L$, respectivement de taille $N_1, ..., N_L$. Les unités dans F_h seront représentées par hj où $j = 1, ..., N_h$ et $h = 1, ..., L$. Supposons également qu'un échantillon aléatoire stratifié (sans remise) de taille $n = n_1 + ... + n_L$ a été prélevé, où n_h est la taille de l'échantillon prélevé de F_h. Supposons que $hJ_1, ..., hJ_{n_h}$ représentent les variables aléatoires de sorte que $hJ_i = hj$ lorsque le i-ième prélèvement de F_h donne lieu à la sélection de hj. Supposons enfin que $hK_1, ..., hK_{n_h}$ représentent des variables aléatoires de sorte que $hK_i = k$ si le i-ième prélèvement de F_h est suivi de la sélection de k de T. Si hjk représente l'arc qui part de l'unité hj dans F_h et se

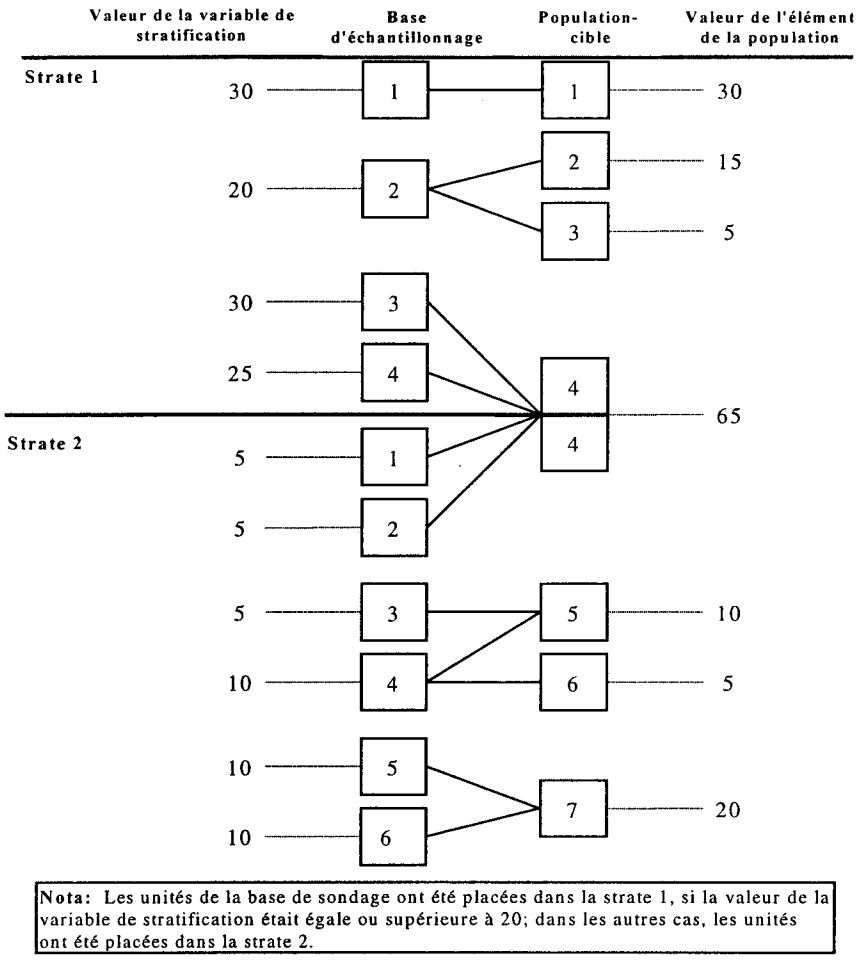

Figure 3.1. Exemple de la correspondance entre la base de sondage et la population-cible dans l'échantillonnage aléatoire stratifié

Tableau 3.1
Probabilités d'arc pour la figure 3.1

Arc hjk	111	122	123	134	144	214	224	235	245	246	257	267
s_{hjk}	1	35796	35796	1	1	1	1	1	35796	35796	1	1

et se termine à k dans T, alors la probabilité marginale de l'arc aléatoire (hJ_i, hK_i) est calculée par

$$\Pr\{(hJ_i, hK_i) = (hjk)\} = \frac{1}{N_h} s_{hjk},$$

où $s_{hjk} = \Pr(hK_i = k \mid hJ_i = hj)$ est une probabilité d'arc. À noter que s_{hjk} est la probabilité conditionnelle de sélectionner l'élément de population k dans T lorsque l'unité hj de la base de sondage a été choisie. Le tableau 3.1 indique les probabilités d'arc pour la figure 3.1, lorsqu'on présume de probabilités de randomisation égales.

Supposons que W_k représente l'ensemble des unités hj dans F dont l'arc se termine à k dans T. Par exemple, $W_4 = \{(1, 3), (1, 4), (2, 1), (2, 2)\}$. Supposons également que le facteur de pondération pour l'élément de la population est $s_k = \sum_{hj \in W_k} S_{hjk}$.

Le tableau 3.2 présente les facteurs de pondération (s_k) pour tous les éléments de la population à la figure 3.1. Les observations faites à la section 2.3.1, au sujet des facteurs de pondération des arcs (s_{hjk}) et des facteurs de pondération des éléments de population (s_k), s'appliquent ici également.

Tableau 3.2
Facteurs de pondération des éléments de la population (s_k) pour la figure 3.1

k	1	2	3	4	5	6	7
s_k	1	35796	35796	1+1+1+1=4	1+1/2=3/2	35796	1+1=2

Pour chaque $h = 1, ..., L$ et $i = 1, ..., n_h$, supposons que x_{hK_i} est une variable aléatoire de telle sorte que $x_{hK_i} = y_k/s_k$ si k dans T est sélectionné après la sélection de quelque unité hj dans F_h.

L'estimateur du total d'une population, selon un plan d'échantillonnage aléatoire stratifié avec une base de sondage de structure plusieurs à plusieurs, est:

$$\hat{Y}_{st} = \sum_{h=1}^{L} \hat{Y}_h, \text{ où } \hat{Y}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} x_{hK_i}. \qquad (3.1)$$

3.3.2 Variance de l'estimateur pour le total d'une population

Avant de définir la variance de l'estimateur (3.1), quelques termes additionnels doivent être définis. Supposons que q_{hk} représente le «facteur de pondération de l'élément de la strate». Ce facteur additionnel est nécessaire en raison du risque de chevauchement. Supposons maintenant que U_{hk} représente l'ensemble des unités dans F_h dont les arcs se terminent à l'élément de population k, par exemple $U_{24} = \{(2, 1), (2, 2)\}$. Définissons maintenant $q_{hk} = \sum_{hj \in U_{hk}} s_{hjk}$. Pour illustrer ceci, rappelons-nous, à la figure 3.1, que l'élément de la population 4 est représenté par deux unités de la base dans la strate 2, de sorte que $q_{24} = \sum_{2j \in U_{24}} s_{2j4} = 2$.

Le facteur de pondération q_{hk} joue le rôle de s_k lorsque la sélection se limite à F_h. En fait, $q_{hk} = s_k$ lorsqu'il n'y a pas de chevauchement. La probabilité de sélectionner quelque unité de F_h à l'étape i sur n_h est égale à $1/N_h$. Cependant, la probabilité de sélectionner un élément de la population k représenté par une unité dans F_h est égale à $\Pr(hK_i = k) = q_{hk}/N_h$, pour toutes les $i = 1, ..., n_h$.

Pour faire la preuve, nous introduisons le terme «total réparti entre les strates», représenté par Y_h^*. En fait, les valeurs des éléments de la population qui sont représentés par les unités dans les strates multiples sont réparties entre ces strates. Supposons que V_h représente l'ensemble des éléments de la population associés aux unités dans F_h. Dans notre exemple, $V_1 = \{1, 2, 3, 4\}$ et $V_2 = \{4, 5, 6, 7\}$. Supposons que

$$Y_h^* = \sum_{k \in V_h} y_k q_{hk} / s_k,$$

où y_k est la valeur de l'élément de population k, $k = 1, 2, ..., M$. Lorsqu'il y a chevauchement, l'utilisation des facteurs de pondération q_{hk} et s_k répartit la mesure y_k entre les strates dans lesquelles l'élément de population k est représenté. L'utilisation de ces facteurs de pondération sert en fait à répartir la valeur de l'élément de population entre les strates, en fonction du nombre de fois que cet élément est représenté dans une strate par rapport au nombre total de fois qu'il est représenté dans la base de sondage. À la figure 3.1, par exemple, Y_1^* et Y_2^* se calculent comme suit:

$$Y_1^* = \frac{30(1)}{1} + \frac{15(1/2)}{1/2} + \frac{5(1/2)}{1/2} + \frac{65(2)}{4} = 82{,}5$$

$$Y_2^* = \frac{65(2)}{4} + \frac{10(3/2)}{3/2} + \frac{5(1/2)}{1/2} + \frac{20(2)}{2} = 67{,}5.$$

À noter que $\sum_{h=1}^{L} Y_h^* = Y$, qu'il y ait ou non chevauchement.

Théorème 3-1: L'estimateur du total d'une population (3.1) est sans biais.

Preuve:

À partir de (3.1),

$$\mathrm{E}\left(\hat{Y}_{st}\right) = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} \mathrm{E}\left(x_{hK_i}\right). \qquad (3.2)$$

Pour chaque $i = 1, ..., n_h$,

$$\mathrm{E}\left(x_{hK_i}\right) = \sum_{k \in V_h} \frac{y_k}{s_k} \Pr(hK_i = k) =$$

$$\sum_{k \in V_h} \frac{y_k}{s_k} \frac{q_{hk}}{N_h} = \frac{1}{N_h} \sum_{k \in V_h} \frac{y_k q_{hk}}{s_k} = \frac{1}{N_h} Y_h^*. \qquad (3.3)$$

Si l'on remplace (3.3) dans l'équation (3.2), on obtient $E(\hat{Y}_{st}) = Y$.

Dans le résultat principal qui suit, nous avons besoin de la notation suivante. Supposons que R_h^* et R_h' représentent respectivement les ensembles de paires non ordonnées admissibles et inadmissibles qui partent de F_h. Les définitions de ces termes sont identiques aux concepts correspondants pour l'ÉASSR, mais elles se limitent maintenant aux strates.

Théorème 3-2: La variance de (3.1) est:

$$\mathrm{V}\left(\hat{Y}_{st}\right) = \sum_{h=1}^{L} S_h^2, \qquad (3.4)$$

où,

$$S_h^2 = \frac{N_h}{n_h}\left[\sum_{k \in V_h} q_{hk}\left(\frac{y_k}{s_k}\right)^2 + \frac{2(n_h - 1)}{(N_h - 1)} \times \right.$$

$$\left. \sum_{[hjk, hj'k'] \in R_h^*} \left(\frac{y_k s_{hjk}}{s_k} \frac{y_{k'} s_{hj'k'}}{s_{k'}}\right)\right] - \left(\sum_{k \in V_h} \frac{y_k q_{hk}}{s_k}\right)^2. \qquad (3.5)$$

Preuve:

Écrivons d'abord

$$\mathrm{V}\left(\hat{Y}_{st}\right) = \mathrm{E}\left(\hat{Y}_{st}\right)^2 - Y^2 = \mathrm{E}\left(\sum_{h=1}^{L} \hat{Y}_h^2\right) -$$

$$\sum_{h=1}^{L} \left(Y_h^*\right)^2 + 2\left(\mathrm{E}\left(\sum_{h<h'} \hat{Y}_h \hat{Y}_{h'}\right) - \sum_{h<h'} Y_h^* Y_{h'}^*\right). \qquad (3.6)$$

Les deux derniers termes s'annulent parce que $\hat{Y}_h$ et $\hat{Y}_{h'}$ sont indépendants. Ceci est logique, puisque la répartition crée une nouvelle population stratifiée sans chevauchement et que les échantillons choisis dans les différentes strates sont indépendants. Par conséquent, avec

$$S_h^2 = \mathrm{E}\left(\hat{Y}_h\right)^2 - \left(Y_h^*\right)^2, \quad \mathrm{V}\left(\hat{Y}_{st}\right) = \mathrm{E}\left(\sum_{h=1}^{L} \hat{Y}_h^2\right) -$$

$$\sum_{h=1}^{L} \left(Y_h^*\right)^2 = \sum_{h=1}^{L} \left(\mathrm{E}\left(\hat{Y}_h^2\right) - \left(Y_h^*\right)^2\right) = \sum_{h=1}^{L} S_h^2.$$

Maintenant,

$$\mathrm{E}\left(\hat{Y}_h\right)^2 = \frac{N_h^2}{n_h^2}\mathrm{E}\left(\sum_{i=1}^{n_h} x_{hK_i}\right) =$$

$$\frac{N_h^2}{n_h^2}\left(\sum_{i=1}^{n_h} \mathrm{E}\left(x_{hK_i}\right)^2 + 2\mathrm{E}\left(\sum_{i<i'} x_{hK_i} x_{hK_{i'}}\right)\right). \tag{3.7}$$

Pour chaque $i = 1, ..., n_h$,

$$\mathrm{E}\left(x_{hK_i}\right)^2 =$$

$$\sum_{k \in V_h}\left[\left(\frac{y_k}{s_k}\right)^2 \Pr(hK_i = k)\right] = \sum_{k \in V_h}\left[\left(\frac{y_k}{s_k}\right)^2 \frac{q_{hk}}{N_h}\right]. \tag{3.8}$$

En utilisant les équations (2.7) et (2.8),

$$2\mathrm{E}\left(\sum_{i>i'} x_{hK_i} x_{hK_{i'}}\right) = 2\binom{n_h}{2}\mathrm{E}\left(x_{hK_i} x_{hK_{i'}}\right) =$$

$$n_h(n_h - 1)\sum_{[hjk,hj'k'] \in R_h^*} \frac{y_k}{s_k}\frac{y_{k'}}{s_{k'}}\Pr(hK_i = k, hK_{i'} = k') =$$

$$n_h(n_h - 1)\sum \sum_{[hjk,hj'k'] \in R_h^*}\left[\left(\frac{y_k}{s_k}\frac{y_{k'}}{s_{k'}}\right)\binom{N_h}{2}^{-1} s_{jk} s_{j'k'}\right] =$$

$$\frac{2n_h(n_h - 1)}{N_h(N_h - 1)}\sum \sum_{[hjk,hj'k'] \in R_h^*}\left[\frac{y_k s_{hjk}}{s_k}\frac{y_{k'} s_{hj'k'}}{s_{k'}}\right]. \tag{3.9}$$

L'équation (3.5) découle maintenant de (3.8), (3.9) et de la définition de Y_h^*.

En utilisant la méthode du corollaire 2-1, l'équation (3.5) peut être simplifiée aux fins de calcul, comme suit:

$$S_h^2 = \frac{N_h}{n_h}\left[\sum_{k \in V_h} q_{hk}\left(\frac{y_k}{s_k}\right)^2 + \frac{(n_h - 1)}{(N_h - 1)}\left[\left(\sum_{hjk \in A_h}\frac{y_k s_{hjk}}{s_k}\right)^2 -\right.\right.$$

$$\left.\left.\sum_{hjk \in A_h}\left(\frac{y_k s_{hjk}}{s_k}\right)^2 - 2\sum_{[hjk,hjk'] \in R_h'}\left(\frac{y_k s_{hjk}}{s_k}\frac{y_{k'} s_{hjk'}}{s_{k'}}\right)\right]\right] -$$

$$\left(\sum_{k \in V_h}\frac{y_k q_{hk}}{s_k}\right)^2,$$

où A_h représente l'ensemble des arcs qui partent des unités dans F_h.

3.4 Moyennes de la population

3.4.1 Estimateur de la moyenne d'une population

L'estimateur élaboré ici pour la moyenne d'une population selon l'échantillonnage aléatoire stratifié constitue un prolongement de l'estimateur proposé par Hansen et coll. 1953a (p. 62-64), ici appliqué à un échantillon aléatoire stratifié prélevé d'une base de plusieurs à plusieurs.

L'estimateur de la moyenne d'une population, selon un plan d'échantillonnage aléatoire stratifié avec base de sondage de plusieurs à plusieurs, est représenté par:

$$\hat{\bar{Y}}_{st} = \sum_{h=1}^{L}\frac{N_h}{N}\hat{\bar{Y}}_h, \text{ où } \hat{\bar{Y}}_h = \frac{\sum_{i=1}^{n_h} x_{hK_i}}{\sum_{i=1}^{n_h}\frac{1}{s_{hK_i}}}. \tag{3.10}$$

Comme dans le cas de l'ÉASSR, l'estimateur de la moyenne d'une population est biaisé, parce qu'il s'agit d'un estimateur par quotient.

4. CONCLUSIONS

Nous avons présenté dans ce document des estimateurs du total, des chiffres et de la moyenne d'une population qui conviennent à une base de sondage de plusieurs à plusieurs, selon des plans d'échantillonnage aléatoire simple et stratifié.

La méthode de pondération décrite ici a été appliquée à une étude sur les immeubles commerciaux pour laquelle un échantillon aléatoire stratifié avait été utilisé. Pour cette étude, pour laquelle la base de sondage était formée des adresses de voirie, les intervieweurs ont noté toute adresse additionnelle se rapportant à l'immeuble sélectionné. Il a ensuite été déterminé si ces adresses additionnelles

figuraient dans la base de sondage et si elles étaient ou non reliées à d'autres éléments de la population (immeubles commerciaux). Dans le cas de scénarios plus complexes, les intervieweurs ont parfois eu recours à des esquisses des immeubles et à l'étiquetage de toutes les adresses pertinentes, ce qui nous a permis de déterminer la structure de tous les sous-graphes CM dans notre échantillon et de déterminer les facteurs de pondération appropriés (s_k).

Nous avons également défini les formules pour calculer la variance de certains des estimateurs présentés dans le document. Il convient de préciser que ces formules sont des paramètres de la population qui ne peuvent être converties facilement en estimations de l'échantillon correspondant. En fait, les auteurs ne connaissent aucune méthode optimale pour estimer les variances décrites dans cet article. Il existe cependant de nombreuses méthodes exigeant beaucoup de calculs (méthode BRR, méthode «boostrap», *etc.*) pour estimer la variance dans le cadre de sondages complexes (Wolter 1985). Il y a lieu également de préciser que chacune de ces méthodes d'estimation de la variance vise un objectif commun, à savoir les formules de la variance que nous avons élaborées ici.

L'utilité de ces formules réside toutefois dans leur application à l'étude des effets des imperfections de la base de sondage, et des caractéristiques de la population, sur la précision des estimations. Une telle étude, qui pourrait faire l'objet d'une autre recherche, devrait mener à la formulation de recommandations et de lignes directrices à l'intention des chercheurs, sur la façon de gérer une base de sondage de structure plusieurs à plusieurs. En d'autres mots, le chercheur devrait, à partir de la base de sondage et des caractéristiques de la population, être en mesure de prendre des décisions stratégiques sur les options qui s'offrent: recenser la population par interview pour supprimer les imperfections de correspondance ou utiliser les estimateurs décrits dans le présent article.

Un autre domaine de recherche futur serait de comparer la précision de nos estimateurs à celle d'autres estimateurs, par exemple l'estimateur de Horvitz-Thompson. Comme nous l'avons indiqué dans l'introduction, l'estimateur de Horvitz-Thompson peut être appliqué à une méthode d'échantillonnage basée sur une structure de plusieurs à plusieurs. L'avantage de l'estimateur de Horvitz-Thompson est qu'il permet, moyennant des probabilités d'inclusion de premier et de second ordres bien définies, d'obtenir à la fois une estimation des caractéristiques de la population et une estimation sans biais de sa variance. En outre, les probabilités d'inclusion du premier ordre peuvent être calculées d'une manière similaire à celle de Musser (1993) à partir uniquement d'information provenant des sous-graphes CM. Ces probabilités sont toutefois très difficiles à calculer dans une base de sondage complexe de structure plusieurs à plusieurs comme la nôtre. Il est en revanche relativement facile de calculer les facteurs de pondération nécessaires pour nos estimateurs.

REMERCIEMENTS

Les auteurs aimeraient remercier le personnel de la Cinergy Corporation pour leur avoir donné l'occasion de mener l'enquête sur les caractéristiques des immeubles, qui est à l'origine de la présente recherche. Nous voulons également remercier les trois examinateurs pour leurs excellentes suggestions qui ont contribué à améliorer sensiblement ce document.

BIBLIOGRAPHIE

BANDYOPADHYAY, S., et ADHIKARI, A.K. (1993). Échantillonnage dans des bases imparfaites contenant un nombre inconnu d'enregistrements répétés. *Techniques d'enquête*, 19, 205-209.

BIRNBAUM, Z.W., et SIRKEN, M.G. (1965). Design of sample surveys to estimate the prevalence of rare diseases: Three unbiased estimates. *Vital and Health Statistics*, PHS Publication 1000, Ser. 2, *Data Evaluation and Methods Research*, no. 11. Hyattsville, MD: National Center for Health Statistics, Public Health Service, U.S. Department of Health and Human Services.

CASADY, R.J., et SIRKEN, M.G. (1980). A multiplicity estimator for multiple frame sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 601-605.

COCHRAN, W.G. (1977). *Sampling Techniques*, (3ième éd.). New York: Wiley & Sons.

HANSEN, M.H., HURWITZ, W.N., et MADOW, W.G. (1953a). *Sample Survey Methods and Theory 1, Methods and Applications*. New York: Wiley & Sons.

HANSEN, M.H., HURWITZ, W.N., et MADOW, W.G. (1953b). *Sample Survey Methods and Theory 2, Theory*. New York: Wiley & Sons.

HORVITZ, D.G., et THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

KISH, L. (1965). *Survey Sampling*. New York: Wiley & Sons.

LAVALLÉE, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 27-35.

LESSLER, J.T., et KALSBEEK, W.D. (1992). *Nonsampling Error in Surveys*. New York: Wiley & Sons.

MUSSER, O. (1993). Unbiased estimation in the presence of frame duplication. *Proceedings of the International Conference on Establishment Surveys*, 889-892.

SIRKEN, M.G. (1972a). Household surveys with multiplicity. *Journal of the American Statistical Association*, 65, 257-266.

SIRKEN, M.G. (1972b). Stratified sample surveys with multiplicity. *Journal of the American Statistical Association*, 65, 224-227.

U.S. DEPARTMENT OF ENERGY, Energy Information Administration (1992). *Commercial Buildings Energy Consumption Survey*.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

WRIGHT, T., et TSAO, H.J. (1983). A frame on frames: An annotated bibliography, (Éd., Tommy Wright). *Statistical Methods and the Improvement of Data Quality*, Orlando, Florida: Academic Press, 25-72.