No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2005 : Défis méthodologiques reliés aux besoins futurs d'information

2005

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2005 de Statistique Canada
Défis méthodologiques reliés aux besoins futurs d'information

L'ESTIMATION EN BASE SIMPLE DANS UNE ENQUÊTE À BASES MULTIPLES

Fulvia Mecatti[1]

RÉSUMÉ

Les enquêtes à bases de sondage multiples ont été proposées, à l'origine, dans un dessein d'*optimalité* afin de générer des économies de coûts, particulièrement dans les cas où une liste complète est disponible, mais coûteuse à échantillonner. Dans la pratique moderne de l'échantillonnage, il arrive fréquemment qu'une liste d'unités complète et à jour, devant servir de base d'échantillonnage, n'est pas disponible. On trouve plutôt un ensemble de deux ou plusieurs listes partielles, habituellement chevauchantes qui, une fois combinées, offrent une couverture adéquate de la population ciblée. Ainsi, le regroupement de listes partielles peut constituer une base multiple. Les travaux antérieurs portant sur la théorie de l'estimation à bases multiples se concentrent sur les bases doubles et ne se préoccupent que rarement du problème pratique important de l'estimation de la variance. En utilisant une démarche fondée sur la notion de *multiplicité*, nous proposons un estimateur à base simple et à pondération fixe pour les enquêtes à bases multiples. Le nouvel estimateur s'applique naturellement à n'importe quelle base et n'exige aucune information à propos de l'appartenance des unités à un domaine. De plus, il est analytiquement simple, de sorte que sa variance est donnée exactement et est facilement estimée. On présente également une étude de simulation comparant le nouvel estimateur avec les principaux concurrents à base simple.

MOTS CLÉS : confidentialité, population difficile à échantillonner, multiplicité; simulation, estimation de la variance.

1. INTRODUCTION

Dans la pratique moderne de l'échantillonnage, il est fréquent de se heurter à un cas où la liste d'unités unique, complète et à jour qui doit servir de base d'échantillonnage n'est pas disponible et ne peut être créée sans effectuer des tris coûteux et difficilement réalisables. Cette situation se présente, par exemple, lorsque nous devons travailler sur une population *peu nombreuse* comme les personnes ayant une maladie rare, sur une population *insaisissable ou cachée* comme celle des sans-abri, des immigrants illégaux ou des consommateurs de drogue, bref sur une population *particulière* ou *difficile à échantillonner* (Sudman et Kalton, 1986, Lepkowski, 1991). À la place, on peut disposer d'un ensemble de $Q \geq 2$ listes. En général, les listes sont partielles et elles se chevauchent, bien que leur union puisse offrir une couverture adéquate de la population ciblée *U*. Voilà un contexte favorable aux enquêtes à bases multiples (BM). Si les travaux antérieurs sur les BM remontent au début des années soixante, Lohr et Rao précisent néanmoins, dans un récent article, qu'« à mesure que les États-Unis, le Canada et d'autres pays se diversifient, des bases de sondage différentes sauront mieux saisir les sous-groupes de la population; nous nous attendons à ce que les plans de sondage modulaires faisant appel à des bases multiples soient généralisés dans l'avenir » (Lohr et Rao, 2005) [traduction libre].

Les travaux antérieurs se concentrent principalement sur le cas de $Q = 2$ listes d'unités disponibles, c'est-à-dire l'enquête à base double (BD), où deux bases *A* et *B* sont données avec $A \cup B = U$ et les dimensions N_A et N_B sont généralement connues, par exemple $N_A + N_B \geq N$, où *N* désigne la taille de la population, généralement inconnue. Les deux bases chevauchantes sont virtuellement divisées en trois *domaines* disjoints : $a = A - B = A \cap B^C$, $b = B - A = B \cap A^C$ (où ^{C} indique la complémentation) et le soi-disant *domaine chevauchant* $ab = A \cap B$. Ainsi, la taille de la population est égale à la somme des tailles des domaines $N_a + N_b + N_{ab} = N$ et le total

[1]Département de la statistique, Université de Milan-Bicocca, Via Bicocca degli Arcimboldi, 8 – Ed. U7, 20126 Milano, Italie. (fulvia.mecatti@unimib.it)

$Y = \sum_{i \in A \cup B} y_i$ de la variable de sondage, supposé le paramètre à estimer, est égal à la somme des totaux des domaines $Y_a + Y_b + Y_{ab} = Y$ où, par exemple, $Y_a = \sum_{i \in a} y_i$. Deux échantillons aléatoires s_A et s_B sont sélectionnés indépendamment des deux bases selon un plan de sondage donné, qui peut être différent pour chaque base. Les données d'échantillon des deux bases sont utilisées pour produire des estimations des totaux de domaine, qui sont combinés pour fournir une estimation de la population totale *Y*.

C'est Hartley (1974) qui a proposé le premier estimateur BD. Celui-ci s'est concentré sur le cas où la base complète *A* est disponible, mais coûteuse à échantillonner, et sur celui où une seconde base *B* est disponible, mais partielle. Afin de générer des économies de coûts tout en obtenant une efficacité égale ou supérieure, les données coûteuses de la base complète *A* sont combinées à l'information moins coûteuse de la base *B* selon une *approche d'optimalité*. Plus particulièrement, l'estimateur pour la population totale *Y* est produit en combinant les estimateurs des totaux de domaine avec des *poids optimaux* $\hat{Y}_H = \sum_{i \in s_a^{(A)}} y_i + \sum_{i \in s_b^{(B)}} y_i + \vartheta \sum_{i \in s_{ab}^{(A)}} y_i + (1-\vartheta) \sum_{i \in s_{ab}^{(B)}} y_i$, c'est-à-dire avec ϑ minimisant la variance de l'estimateur, où par exemple $s_{ab}^{(A)}$ indique le sous-échantillon s_A des unités incluses dans le domaine chevauchant *ab*. L'estimateur optimal de Hartley a été successivement amélioré. Plus particulièrement, Fuller et Burmeister (1972) ont introduit un estimateur optimal pour la BD qui a été interprété comme un estimateur du maximum de vraisemblance (Skinner, 1991) et a été démontré comme asymptotiquement efficace (Lohr et Rao, 2000). Des problèmes pratiques et méthodologiques peuvent surgir lorsque la démarche d'optimalité est adoptée et les estimateurs optimaux sont appliqués : *i)* on constate que les poids optimaux dépendent de variances et de covariances inconnues de telle sorte qu'ils doivent être estimés à partir de données d'échantillon, ce qui pourrait s'avérer compliqué et affecter l'optimalité en soi; *ii)* les poids estimés dépendent des valeurs des variables de l'enquête et diffèrent donc pour chaque variable de l'enquête, ce qui est peu pratique dans le contexte d'une enquête à usages multiples; *iii)* la généralisation à une configuration en BM ne se fait pas de manière simple et directe et peut même s'avérer impossible (Skinner, 1991); *iv)* l'enjeu pratique important de l'estimation de variance est à peine effleuré; *v)* il faut savoir à quel domaine appartient chaque unité échantillonnée, c'est-à-dire qu'il faut bien classer les unités échantillonnés de chacune des bases par domaine afin d'appliquer les estimateurs optimaux. Une telle hypothèse est forte, comme le mentionnent par exemple Lohr et Rao (2005), puisque les estimateurs optimaux sont sensibles à une mauvaise classification par domaine des unités échantillonnées.

Le point *ii)* a été abordé par Skinner et Rao (1996), parallèlement à l'application de plans de sondage complexes, en introduisant un estimateur de pseudo-maximum de vraisemblance (PMV) pour bases doubles; dans Lohr et Rao (2005), le point iii) est traité à la fois pour les estimateurs optimaux et PMV; dans le présent article, nous nous concentrons sur les points *iv)* et *v)*. La section 2 rappelle plus particulièrement la démarche pour les bases simples. Dans la section 3, un estimateur de multiplicité pour BM est proposé. Un certain nombre de résultats de simulation sont présentés dans la section 4.

2. L'ESTIMATION À BASE SIMPLE

Comme solution de remplacement de la démarche d'optimalité, une *méthode fondée sur la notion de base simple* (BS) peut être utilisée. Dans un estimateur à base simple, les données des deux bases sont combinées en utilisant des poids fixés dépendamment des probabilités d'inclusion en fonction du plan induit par les deux plans des bases transposés à l'échantillon *total*, c'est-à-dire l'union des deux échantillons des bases (Bamkier, 1986; Kalton et Anderson; 1986, Skinner, 1991) : $\hat{Y} = \sum_{i \in s_A} w_i y_i + \sum_{i \in s_B} w_i y_i$ où $w_i = (\pi_{Ai} + \pi_{Bi})^{-1}$ avec $\pi_{Ai} = 0$ si $i \in b$ et $\pi_{Bi} = 0$ si $i \in a$. Pour l'échantillonnage aléatoire simple (ÉAS) de chaque base, l'estimateur à BS pour la BD est donné par : $f_A^{-1} \sum_{i \in s_a^{(A)}} y_i + (f_A + f_B)^{-1} \left(\sum_{i \in s_{ab}^{(A)}} y_i + \sum_{i \in s_{ab}^{(B)}} y_i \right) + f_B^{-1} \sum_{i \in s_b^{(B)}} y_i$ où $f_A = n_A / N_A$ et $f_B = n_B / N_B$ c'est-à-dire les fractions d'échantillonnage des bases. Étant donné que les poids fixes diffèrent habituellement des poids optimaux, l'estimateur BS est en général (asymptotiquement) moins efficient que les estimateurs optimaux (Lohr et Rao, 2000). De plus, il faut toujours bien classer les unités échantillonnées par domaine. D'un autre côté, l'estimateur BS ne requiert pas l'identification des unités en double puisque les unités échantillonnées du domaine chevauchant sont pondérées par le même coefficient fixe; de plus, il s'étend naturellement au cadre de la BM. À cette fin, la notation à BM ingénieuse de Lohr et Rao (2005) est appliquée ici de manière extensive. Soit un

ensemble de $Q \geq 2$ bases chevauchantes $A_1 \cdots A_q \cdots A_Q$, en supposant que $\bigcup_q A_q = U$. Définissons les ensembles d'indices K comme sous-ensembles de l'étendue de l'indice de la base $q = 1 \cdots Q$. Pour chaque ensemble d'indices $K \subseteq \{1 \cdots q \cdots Q\}$, un domaine est défini sous la forme $D_K = \left(\bigcap_{q \in K} A_q\right) \cap \left(\bigcap_{q \notin K} A_q^C\right)$ avec $2^Q - 1$ domaines différents. Par exemple, avec $Q = 3$, il y a 7 domaines D_K désignés par les ensembles d'indices $K = \{\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}\}$. Définissons comme *indicateur d'appartenance à un domaine* la variable aléatoire $\delta_i^{(D_K)}$ qui prend la valeur 1 si $i \in D_K$ et 0 dans le cas contraire; le total de la population, qui doit être estimé, est alors exprimé sous la forme d'une somme sur l'ensemble de $2^Q - 1$ domaines par le biais de l'appartenance de l'unité à un domaine, c'est-à-dire $Y = \sum_K \sum_{i \in \bigcup_q A_q} \delta_i^{(D_K)} y_i$. Prenons s_q comme échantillon choisi dans la base A_q selon un plan donné, indépendamment pour $q = 1 \cdots Q$. L'estimateur BS (simple) dans le cas général de BM est alors donné par $\hat{Y}_{SF} = \sum_K \sum_{q \in K} \sum_{i \in s_q} w_i \delta_i^{(D_K)} y_i$ avec des poids fixes w_i garantissant que le plan n'est pas biaisé. Sous un ÉAS de chaque base, il est donné par :

$$\hat{Y}_{SF} = \sum_K \sum_{q \in K} \left(\sum_{q \in K} f_q \right)^{-1} \sum_{i \in s_q} \delta_i^{(D_K)} y_i \qquad (1)$$

Comme il est expliqué ci-dessus, les poids fixes sont en général non optimaux. Ainsi, afin d'améliorer la performance, on a proposé de corriger l'estimateur simple de BS au moyen de la méthode itérative du quotient en utilisant les tailles de base connues N_q. Pour la BD, Skinner (1991) a dérivé la forme restrictive (fermée) de l'estimateur de la méthode itérative du quotient de la BS lorsque le nombre *r* d'itérations augmente à l'infini. Dans le cadre de la BM, l'estimateur de la méthode itérative du quotient en BS $\hat{Y}_{SFrak}$ devient plus complexe sur le plan computationnel : l'estimateur (1) doit être corrigé à toutes les itérations $r \bmod(Q)$ pour ce qui concerne la dimension N_q des bases en cause dans chaque domaine D_K, c'est-à-dire $\forall K \ni q$ itérativement jusqu'à la convergence. Il convient de remarquer, comme les estimateurs à BS ou itérés donnent lieu à l'indicateur d'appartenance au domaine $\delta_i^{(D_K)}$, tel qu'il apparaît par exemple dans l'équation (1), leur application requiert une classification par domaine connue et appropriée des unités échantillonnées. Outre le risque de mauvaise classification, cela sous-entend une collecte d'information supplémentaire, étant donné qu'il faut aussi demander aux unités échantillonnées, parallèlement à la variable de l'enquête, *à quelles* bases elles appartiennent (en plus de celle à partir de laquelle elles ont été échantillonnées). Cette hypothèse peut être retirée en adoptant une *démarche fondée sur la notion de multiplicité* permettant de corriger l'estimateur à BS.

3. UN ESTIMATEUR DE MULTIPLICITÉ

La notion de multiplicité a été introduite pour la première fois dans le cadre de l'échantillonnage par réseau (Sirken, 2004, Casady et Sirken, 1980). Elle fait aussi partie des outils de la méthode généralisée de partage des poids (Lavallée, 2002) ainsi que de la théorie de l'estimation par l'échantillonnage de centre (Mecatti, 2004), dont le cadre équivaut à celui de la BM sous certaines conditions. En utilisant une démarche fondée sur la notion de multiplicité, nous proposons un estimateur à BM dont le plan n'est pas biaisé. Contrairement aux estimateurs optimaux et à BS, l'estimateur de multiplicité ne dépend pas de l'appartenance à un domaine des unités échantillonnées pour déterminer *à combien* de bases elles appartiennent plutôt *qu'à quelles* bases. Comme nous l'avons déjà signalé, cet aspect représente un avantage pratique, du fait qu'il atténue le risque de mauvaise classification et réduit la quantité d'information demandée aux unités échantillonnées qui pourraient être, d'une manière ou d'une autre, sensibles à l'appartenance à une base, par exemple lors de l'échantillonnage d'ex-détenus, de toxicomanes, de patients ou d'immigrants illégaux.

Dans Lohr et Rao (2005), la multiplicité du domaine D_K est définie comme la cardinalité de l'ensemble d'indices *K*. Étant donné que les domaines sont mutuellement exclusifs, c'est-à-dire que chaque unité *i* appartient à un domaine et à un seul, la multiplicité est également une caractéristique de chaque unité $m_i = \sum_q \delta_i^{(A_q)}$ où $\delta_i^{(A_q)}$ désigne

l'indicateur d'appartenance à une base, c'est-à-dire la variable aléatoire qui prend 1 comme valeur si $i \in A_q$ et 0 dans le cas contraire. La multiplicité d'unité m_i est égale au nombre de bases au sein desquelles chaque unité *i* est incluse. Ainsi, on peut l'observer en demandant simplement aux unités *à combien de bases elles appartiennent*. En utilisant la démarche fondée sur la notion de multiplicité, le total de la population à estimer est exprimé comme une somme par rapport aux bases plutôt qu'une somme par rapport aux domaines :

$$Y = \sum_K \sum_{i \in \bigcup_q A_q} \delta_i^{(D_K)} y_i \, m_i^{-1} \sum_{q=1}^{Q} \delta_i^{(A_q)} = \sum_{q=1}^{Q} \sum_{i \in \bigcup_q A_q} \delta_i^{(A_q)} y_i m_i^{-1} \sum_K \delta_i^{(D_K)}$$
$$= \sum_{q=1}^{Q} \sum_{i \in A_q} y_i m_i^{-1} \qquad (2)$$

Il convient de noter que dans l'équation (2), l'indicateur d'appartenance au domaine est exclu (parce qu'il est $\sum_K \delta_i^{(D_K)} = 1, \ \forall i \in \bigcup_q A_q$). L'équation (2) représente un avantage pratique puisque les domaines sont des objets virtuels tandis que l'échantillonnage est effectué au sein des bases *Q*. Par conséquent, un estimateur de multiplicité à BS est donné par :

$$\hat{Y}_{SFmulti} = \sum_{q=1}^{Q} \sum_{i \in s_q} w_i y_i m_i^{-1} \qquad (3)$$

avec des poids fixes w_i pour s'assurer, par exemple, que le calcul soit non biaisé sous les plans des bases sélectionnés, c'est-à-dire l'inverse de la probabilité d'inclusion. Pour un ÉAS, nous avons $w_i = f_q^{-1}, \forall i \in s_q$. Vu sa structure de type Horvitz-Thompson, la variance exacte de l'estimateur (3) est donnée en forme analytique et est donc facilement estimée. Pour l'ÉAS de chaque base, l'estimateur de variance est donné par

$$Var\left(\hat{Y}_{SFmulti}\right) = \sum_{q=1}^{Q} \frac{N_q - n_q}{n_q\left(N_q - 1\right)} \left[N_q \sum_{i \in A_q} y_i^2 m_i^{-2} - \left(\sum_{i \in A_q} y_i m_i^{-1} \right)^2 \right]. \qquad (4)$$

Un estimateur de variance non biaisé courant pour l'ÉAS est donc

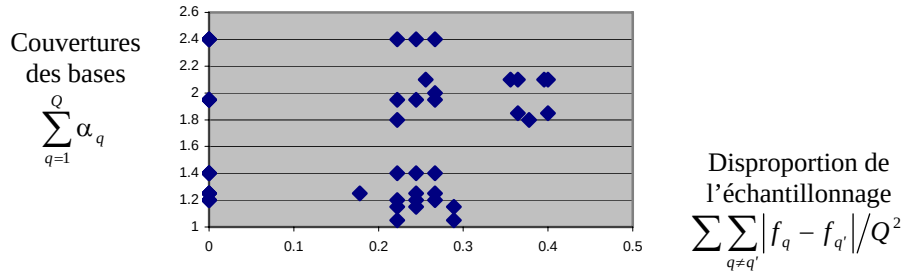
$$\hat{v}\left(\hat{Y}_{SFmulti}\right) = \sum_{q=1}^{Q} \frac{N_q\left(N_q - n_q\right)}{n_q^2\left(N_q - 1\right)} \left[N_q \sum_{i \in s_q} y_i^2 m_i^{-2} - f_q^{-1} \left(\sum_{i \in s_q} y_i m_m^{-1} \right)^2 \right]. \qquad (5)$$

L'estimateur (3) est d'abord comparé avec d'autres estimateurs concurrents directs à base simple, soit l'estimateur à BS simple et celui de la méthode itérative, selon un ÉAS de chaque base. Pour commencer, une constatation s'impose : alors que dans l'estimateur à BS simple donné par (1) les valeurs des unités sont pondérées par $\left(\sum_{q \in K} f_q\right)^{-1} \delta_i^{(D_K)} = w_i(K)$, c'est-à-dire un coefficient *moyen* calculé à partir des bases de chaque domaine, dans l'estimateur de multiplicité elles sont pondérées par $\left(f_q m_i\right)^{-1} = w_i(q)$, c'est-à-dire un coefficient de base *spécifique*. Ainsi, elles coïncident lors d'un échantillonnage proportionnel de chaque base, c'est-à-dire pour des fractions d'échantillon constantes $f_q = f \ \ (q = 1 \cdots Q)$, alors qu'elles diffèrent lors d'un échantillonnage non proportionnel.

4. UNE ÉTUDE EN SIMULATION

Une étude en simulation a été menée afin de comparer l'estimateur de multiplicité proposé avec les concurrents à BS directs et d'analyser les propriétés inférentielles pour des échantillons de taille finie.

Figure 1 : *Populations simulées*



La simulation porte sur des populations artificielles de taille $N = 120$ divisées aléatoirement en $Q = 3$ bases chevauchantes, la variable d'enquête est fournie à l'aide d'une distribution discrète uniforme allant de 1 à 100, et le total de la population Y est supposé le paramètre à estimer. Différents scénarios sont produits en variant la couverture de la base $\alpha_q = N_q / N$ sous la contrainte $\sum_q \alpha_q > 1$ de façon à contrôler le chevauchement entre les bases, et en variant les fractions d'échantillonnage $f_q = N_q / N$. Dans la figure 1, les populations simulées sont représentées par des points sur le plan formé par les deux principaux paramètres de simulation : la somme des couvertures des bases sur l'axe vertical et une mesure de dispersion parmi les fractions d'échantillonnage sur l'axe horizontal. Lorsque la dispersion des fractions d'échantillonnage augmente, la disproportion de l'échantillonnage augmente elle aussi. Ainsi, en simulant le long de l'axe vertical, on étudie le cas de l'échantillonnage proportionnel, alors que l'échantillonnage disproportionné est simulé sur l'ensemble du plan. Dans cette étude, 45 populations/points sont considérés. Pour chaque population, une distribution de Monte Carlo des trois estimateurs à BS $\hat{Y}_{SF}$,$\hat{Y}_{SFrak}$ et $\hat{Y}_{SFmulti}$ est produite selon un ÉAS de chaque base, puis on en calcule la moyenne empirique et l'erreur quadratique moyenne empirique (EMSE). L'objectif principal de l'étude en simulation consiste à évaluer l'estimateur de multiplicité proposé en fonction de sa performance relative de manière à ce que le ratio de performance empirique $EEff = Var\left(\hat{Y}_{SFmulti}\right) / EMSE\left(\hat{Y}\right)$ soit fourni. Le ratio *EEff* peut donner un résultat inférieur, égal ou

Tableau 1 : *Échantillonnage proportionnel. Performance de $\hat{Y}_{SFmulti} \equiv \hat{Y}_{SF}$ relative à $\hat{Y}_{SFrak}$*

Couvertures de bases égales

α_q	*EEff*
élevée 0,8	[0,97, 1,03]
faible 0,4	[0,97, 1,03]

Tableau 2 : *Échantillonnage proportionnel. Performance de $\hat{Y}_{SFmulti} \equiv \hat{Y}_{SF}$ relative à $\hat{Y}_{SFrak}$*

Couvertures de bases différentes

α_q	***EEff***
0,2 0,4 0,8	[0,99, 1,14]
0,2 0,15 0,9	[1,09, 1,12]
0,8 0,9 0,25	[1,02, 1,08]

supérieur à 1 signifiant que $\hat{Y}_{SFmulti}$ est plus, aussi ou moins performant que l'estimateur compare. L'erreur de Monte Carlo a été contrôlée en acceptant exclusivement les simulations donnant un biais de Monte Carlo relatif inférieur à 1 % pour les estimateurs que l'on sait non biaisés. Il a donc fallu maintenir le nombre de cycles de simulation dans une fourchette de 32 000 à 60 000, dans les pires des cas.

Les tableaux 1 et 2 présentent certains résultats indicatifs de simulation dans le cas d'un échantillonnage proportionnel, lorsque la multiplicité et les estimateurs simples à BS coïncident. Dans ce cas, les simulations considèrent l'effet de la correction du ratio de la méthode itérative du quotient sur l'impartialité et la performance. En tant qu'indication générale, la correction du ratio de la méthode itérative du quotient n'affecte pas l'impartialité,

puisque $\hat{Y}_{SFrak}$ donne toujours des résultats non biaisés ou presque. Nous nous concentrons alors sur la performance. Les résultats du tableau 1 montrent que la correction du ratio de la méthode itérative de correction n'améliore pas l'estimateur à BS simple dans le cas de couvertures de bases égales. Autrement dit,, lorsque les tailles de bases sont toutes égales $N_q = N/Q$, elles ne fournissent pas d'information supplémentaires et la méthode itérative est dénuée d'intérêt. À mesure que la différence entre les couvertures des bases augmente, on constate certains gains de performance, allant de 2 % à 14 %. De tous les cas étudiés selon l'échantillonnage proportionnel, la correction du ratio de la méthode itérative du quotient améliore l'estimateur à BS simple dans 60 % des cas tout en fournissant la même performance pour les 40 % restants. Environ 62 cycles d'itérations sont nécessaires, en moyenne, pour atteindre la convergence. Les tableaux 3 à 5 présentent les principaux résultats de simulation dans le cas de l'échantillonnage non proportionnel. Il s'agit d'abord d'établir les effets de la correction de multiplicité en comparant l'estimateur de multiplicité avec l'estimateur à BS simple. Le tableau 3 porte sur le cas de couvertures de bases égales : $\hat{Y}_{SFmulti}$ est moins efficient que l'estimateur à BS simple dans seulement un cas à couverture de bases élevée, alors que des gains de performance entre 12 % et 22 % sont enregistrés dans tous les autres cas. Dans l'ensemble, la correction de multiplicité améliore l'estimateur à BS simple, surtout en contexte où la couverture de la base de sondage s'amenuise. Quant au tableau 4, il montre que, pour différentes couvertures de bases, les gains majeurs de performance de l'estimateur de multiplicité pour l'estimateur à BS simple sont enregistrés entre 12 % et 73 % à mesure que la disproportion de l'échantillonnage augmente. Dans l'ensemble, la correction de multiplicité améliore l'estimateur à BS simple, surtout en contexte de disproportion d'échantillonnage accrue.

Tableau 3 : *Échantillonnage non proportionnel*
Perfromance de $\hat{Y}_{SFmulti}$ relative à $\hat{Y}_{SF}$
Couvertures de bases égales

α_q	f_q	*EEff*	Gains de performance
0,4	0,1 0,3 0,6	0,78	22 %
	0,75 0,75 0,15	0,84	16 %
0,6	0,09 0,05 0,5	0,79	21 %
	0,2 0,2 0,75	0,88	12 %
0,7	0,6 0,05 0,25	0,86	14 %
0,8	0,2 0,2 0,75	1,14	-14 %

Tableau 4 : *Échantillonnage non proportionnel*
Performance de $\hat{Y}_{SFmulti}$ relative à $\hat{Y}_{SF}$
Couvertures de bases différentes

α_q	f_q	*EEff*	Gains de performance
0,2 0,15 0,9	0,9 0,01 0,81	0,27	73 %
	0,75 0,75 0,15	0,81	19 %
0,9 0,7 0,5	0,05 0,25 0,95	0,88	12 %
	0,5 0,08 0,9	0,85	15 %

Enfin, l'estimateur $\hat{Y}_{SFmulti}$ a été comparé avec l'estimateur de la méthode itérative du quotient $\hat{Y}_{SFrak}$ en tant que solutions de remplacement pour correction de l'estimateur à BS simple. Puisque l'estimateur de multiplicité nécessite moins d'information que l'estimateur de la méthode itérative du quotient pour l'appartenance des unités échantillonnées à un domaine et les tailles des bases, on s'attend à une certaine perte de performance. Ce qui nous intéresse est de savoir *combien* et *quand*. À cet égard, les résultats de simulation sont très dispersés. Le tableau 5 tente de les résumer et d'en tirer une indication générale. Tous les résultats de simulation pour l'échantillonnage non proportionnel ont été classés d'après le ratio de performance empirique en trois catégories : 1) perte de performance *nulle* ou *négligeable* (moins de 10 %), 2) perte de performance *légère* (entre 10 % et 20 %) et 3) perte de performance *lourde* (plus de 20 %). La deuxième colonne du tableau 5 montre que des pertes de performance nulles ont été enregistrées dans 15 % des cas avec des pertes de 0 % en moyenne (troisième colonne); dans 23 % des cas, la perte est de 18 % en moyenne, et une perte de performance lourde (59 % en moyenne) a été enregistrée dans 62 % des cas. Les trois dernières colonnes du tableau 5 présentent le minimum, le maximum et le 75^{e} quantile des pertes de performance observées, de façon à ce que les valeurs négatives indiquent les gains de performance de la correction de multiplicité par rapport à la correction du ratio de la méthode d'itération des quotients. En tant qu'indication générale, $\hat{Y}_{SFmulti}$ donne des résultats aussi – ou légèrement moins – efficaces que $\hat{Y}_{SFrak}$, tout particulièrement dans les cas de couvertures de bases décroissantes.

Tableau 5 : *Échantillonnage non proportionnel*
Performance de $\hat{Y}_{SFmulti}$ *par rapport à* $\hat{Y}_{SFrak}$

Perte de performance		Cas	Moyenne	min.	max.	75e
nulle	< 10 %	15 %	0 %	-4 %	7 %	-1 %
légère	10 % à 20 %	23 %	18 %	15 %	20 %	20 %
lourde	> 20 %	62 %	59 %	21 %	112 %	75 %

REMERCIEMENTS

L'auteur tient à remercier Jon N.K. Rao pour ses observations utiles et ses suggestions ingénieuses.

RÉFÉRENCES

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

Casady, R. J. et Sirken, M. G. (1980), "A multiplicity Estimator for Multiple Frame Sampling", *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 601-605.

Fuller, W. A. et Burmeister, L. F. (1972), "Estimates for samples selected from two overlapping frames", *Proceedings of the Social Statistical Section, American Statistical Association*, pp. 245-249.

Hartley, H. O. (1974), "Multiple Frame Methodology and Selected Applications", *Sankya*, C, 36, pp. 99-118.

Kalton, G. et Andersen, D. W. (1986), "Sampling rare populations", *Journal of Royal Statistical Society*, A, 149, pp. 65-82.

Lavallée, P. (2002), *Le Sondage Indirect ou la Méthode Généralisée du Partage des Poids*, Editions de l'Université de Bruxelles, Editions Ellipses.

Lepkowski, J. M. (1991), "Sampling the Difficult-to-Sample", *Proceedings of the Symposium of the American Institute of Nutrition*, pp. 416-423.

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

Lohr, S. L. et Rao, J. N. K. (2005), "Estimation in Multiple Frame Surveys", *Journal of the American Statistical Association*, to appear.

Mecatti, F. (2004), " Échantillonnage de centres : stratégie d'enquête auprès des populations difficiles à échantillonner", *Recueil du Symposium 2004 de Statistique Canada*, Statistique Canada.

Sirken, M. G. (2004), "Enquêtes auprès de populations rares ou difficiles à rejoindre au moyen d'échantillonnage par réseau : revue historique ", *Recueil du Symposium 2004 de Statistique Canada*, Statistique Canada.

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

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

Sudman, S. et Kalton, G. (1986), “New Developments in the Sampling of Special Populations”, *Ann. Rev. Sociol.*, 12, pp. 401-429.