Techniques d'enquête, juin 1998
Vol. 24, n° 1, pp. 3-10
Statistique Canada

Échantillonnage et estimation à partir de bases de sondage listes multiples

PHILLIP S. KOTT, JOHN F. AMRHEIN et SUSAN D. HICKS[1]

RÉSUMÉ

Un grand nombre des enquêtes économiques et agricoles visent des objectifs multiples. Il serait donc pratique de pouvoir stratifier la population-cible de ces enquêtes de différentes manières – et ainsi répondre à un certain nombre d'objectifs – puis de combiner les échantillons pour le dénombrement. Nous examinons dans ce document quatre méthodes d'échantillonnage distinctes qui prélèvent des échantillons similaires, toutes stratifications confondues, ce qui permet de réduire la taille globale de l'échantillon. L'efficacité de ces stratégies d'échantillonnage est évaluée à la lumière des données extraites d'une enquête sur l'agriculture. Nous indiquons ensuite comment un estimateur par calage (c.-à-d. pondéré de nouveau) peut accroître l'efficacité statistique, en reproduisant dans l'estimation ce que l'on sait de la taille de la strate initiale. La méthode itérative du quotient, qui a été proposée dans certains ouvrages, n'est en fait qu'une méthode de calage.

MOTS CLÉS: Étalonnage; échantillonnage par collocation; nombre aléatoires permanents; échantillonnage de Poisson; échantillonnage systématique avec probabilité proportionnelle à la taille.

1. INTRODUCTION

Un grand nombre des enquêtes fondées sur une base de sondage liste, qui sont menées par le National Agricultural Statistics Service (NASS), sont intégrées, en ce que des données portant sur un éventail de sujets hétérogènes – depuis les superficies en culture aux stocks de céréales – sont recueillies par le biais d'une enquête unique plutôt que par l'exécution de plusieurs enquêtes indépendantes. Bankier (1986), Skinner (1991) et Skinner, Holmes et Holt (1994) ont démontré qu'il est possible d'accroître l'efficacité d'une vieille méthode qui consiste à combiner des échantillons aléatoires simples stratifiés prélevés séparément (où chaque échantillon provient d'une base de sondage liste dont le plan de stratification diffère); une telle stratégie d'estimation combinée donnerait des variances inférieures à celles qui seraient obtenues des enquêtes indépendantes compilées séparément.

Encore plus intéressant dans bien des cas serait un plan d'échantillonnage qui aurait tendance à sélectionner les mêmes unités de chaque base de sondage ce qui, par le fait même, réduirait les coûts et le fardeau de réponse associés à une enquête intégrée. Nous examinons ici plusieurs plans d'échantillonnage de ce type, dont trois sont fondés sur l'utilisation de nombres aléatoires permanents. Le quatrième plan utilise une variante de la méthode d'échantillonnage systématique avec probabilité proportionnelle à la taille. Le but visé, avec chacun des plans présentés, est d'atteindre, voire de dépasser (du moins en moyenne) une série particulière d'objectifs relatifs à la taille des échantillons.

Nous démontrons également comment un estimateur par calage (repondéré) peut améliorer l'efficacité relative, en reproduisant dans l'estimation ce que l'on sait sur la taille de la strate initiale. Dans la dernière section, nous montrons que l'utilisation d'une technique de calage peut servir à autre chose qu'à refléter uniquement la taille de la strate initiale.

Une autre stratégie pour réduire le fardeau de réponse consiste à utiliser des instruments séparés pour répondre à différents objectifs et à sélectionner des échantillons distincts pour chaque instrument. Cette méthode augmente le nombre total d'unités sélectionnées, mais elle réduit le fardeau de réponse pour chacune de ces unités. Le NASS utilise cette approche pour son étude sur la gestion des ressources agricoles (Agricultural Resources Management Study, voir Kott et Fetter 1997), mais ce n'est *pas* de cette approche dont il sera ici question.

2. ÉCHANTILLONNAGE INDÉPENDANT ET ESTIMATION SANS BIAIS

Supposons que nous avons F bases de sondage indépendantes, par exemple une base de sondage pour les stocks de sorgho, une autre pour l'avoine et une base générale pour l'ensemble des stocks de céréales. Chaque base de sondage est stratifiée indépendamment et des échantillons aléatoires simples sans remise sont prélevés dans chaque strate, de chaque base de sondage. Supposons maintenant que la base de sondage f (disons celle de l'avoine) renferme H_f strates, que la strate h (grosses exploitations productrices d'avoine) dans la base f contient N_{fh} unités de population, dont n_{fh} unités sont sélectionnées. L'union de l'ensemble des F bases de sondage doit couvrir l'ensemble de la population (liste), mais il n'est

[1] Phillip S. Kott, Research Division; John F. Amrhein, Survey Sampling Branch et Susan D. Hicks, Estimates Division, National Agricultural Statistics Service, USDA.

pas nécessaire qu'au moins une base particulière soit complète. Il peut y aussi avoir chevauchement entre les bases de sondage.

Un estimateur sans biais pour la population totale $T = \sum_{i \in P} y_i$ est l'estimateur de multiplicité simple proposé par Skinner (1991):

$$t_M = \sum_{i \in P} y_i n_{(i)} / E[n_{(i)}], \qquad (1)$$

où P représente l'ensemble de la population et $n_{(i)}$ est le nombre de fois que l'unité i est sélectionnée de quelque base de sondage que ce soit pour faire partie de l'échantillon. On remarquera que $n_{(i)} = 0$ pour les unités de population qui ne font pas partie de l'échantillon. Dans la grande majorité des applications, $n_{(i)}$ sera égal à 1 pour la plupart des unités échantillonnées, mais il est également possible que $n_{(i)} > 1$ avec ce plan d'échantillonnage.

Le nombre prévu de fois que l'unité i sera sélectionnée pour former l'échantillon correspond à $E[n_{(i)}] = \sum^F p_{if}$, où p_{if} est la probabilité de sélection de l'unité i dans l'échantillon aléatoire simple stratifié tiré de la base de sondage F; en d'autres mots, $p_{if} = n_{fh}/N_{fh}$, où l'unité i se trouve dans la strate h, de la base de sondage f.

Il existe également un estimateur de Horvitz-Thompson pour T en vertu du plan, soit $t_{\mathrm{HT}} = \sum_{i \in S} y_i/\pi_i$, où S représente l'échantillon et où $\pi_i = 1 - (1 - p_{i1})(1 - p_{i2}) \cdots (1 - p_{iF})$. Voir Bankier (1986) pour plus d'information sur cette approche.

3. STRATÉGIES D'ÉCHANTILLONNAGE UTILISANT DES NOMBRES ALÉATOIRES PERMANENTS

Le plan d'échantillonnage décrit précédemment est indépendant, toutes bases confondues. Pour bon nombre d'enquêtes, toutefois, il serait utile que le plan d'échantillonnage ne soit *pas* indépendant, car toutes les unités dans l'échantillon combiné ont le même instrument d'enquête et parce que bon nombre d'unités se retrouvent dans un certain nombre de bases de sondage. Par conséquent, étant donné les objectifs relatifs à la taille de l'échantillon prélevé de la base ou la strate, un plan d'échantillonnage qui aurait tendance à sélectionner la même unité dans chaque base de sondage devrait se traduire par un nombre moins élevé de contacts (et, partant, par des coûts d'enquête moindres) qu'un échantillonnage indépendant toutes bases confondues.

Supposons, à cette fin, que chaque unité doive atteindre ou dépasser l'objectif p_{if} dans chaque base de sondage. Cette valeur cible est constante pour toutes les unités qui se trouvent dans la strate h de la base de sondage f. Nous réservons pour la dernière section notre évaluation de la politique qui consiste à mettre l'accent sur des valeurs-cibles pour p_{if} – ou encore pour n_{if}. Nous nous contenterons ici de dire que bon nombre d'organismes statistiques, dont le NASS, ont une telle politique.

Selon un plan d'échantillonnage potentiel, un *nombre aléatoire permanent* (NAP) – obtenu de la distribution uniforme sur l'intervalle [0, 1) – est attribué à chaque unité de la population. L'unité i est sélectionnée pour faire partie de l'échantillon de la base de sondage f, lorsque son NAP est inférieur à p_{if}.

Nous obtenons ainsi un échantillon de Poisson, où la probabilité que l'unité i soit sélectionnée pour faire partie de l'échantillon correspond à $\pi_i = \max_f \{p_{if}\}$, laquelle, manifestement, est au moins aussi grande que chaque p_{if} individuelle pour une unité donnée. Selon un tel plan d'échantillonnage, l'estimateur sans biais de Horvitz-Thompson pour T correspond à $t_p = \sum_{i \in S} y_i / \max_f \{p_{if}\}$.

Dans l'échantillonnage de Poisson, la taille de l'échantillon est aléatoire. Une façon de réduire la variance de la taille de l'échantillon est d'utiliser une variante de ce plan d'échantillonnage. Dans l'échantillonnage par collocation fondé sur les NAP, un NAP unique est attribué à chaque unité de population, lequel est choisi parmi les éléments de la série $\{e/N, (1 + e)/N, (2 + e)/N, \ldots, (N - 1 + e)/N\}$, où e est une variable aléatoire uniforme prélevée dans l'intervalle [0, 1). Pour ce faire, on peut d'abord tirer un NAP temporaire pour chaque unité, puis une valeur pour e. L'unité dont le NAP temporaire est le plus faible se voit attribuer un NAP par collocation dont la valeur est e/N, celle dont le NAP temporaire est le deuxième plus faible obtient la valeur $(1 + e)/N$, et ainsi de suite jusqu'à ce que la valeur $(N - 1 + e)/N$ soit assignée à l'unité dont le NAP temporaire est le plus élevé. L'estimateur t_p demeure sans biais dans l'échantillonnage par collocation.

En raison de la nature aléatoire de la taille des échantillons obtenus par échantillonnage de Poisson et échantillonnage par collocation, il se peut que les objectifs relatifs à la taille de l'échantillon tiré de la base de sondage ou de la strate ne soient pas atteints pour un échantillon particulier que l'on tire. Un troisième plan d'échantillonnage fondé sur les NAP est mis en oeuvre avec comme objectifs les n_{fh} valeurs cibles, ce qui supprime cette possibilité. Dans ce dernier plan, les unités dans la strate h de la base de sondage f, affichant les n_{fh} plus faibles NAP, sont sélectionnées pour former l'échantillon (cette méthode est très similaire à l'échantillonnage séquentiel de Poisson utilisé par Ohlsson 1995). Avec ce plan fondé sur les *NAP avec échantillon fixe,* les probabilités de sélection des unités échantillonnées doivent être calculées pour l'estimateur de Horvitz-Thompson – une tâche difficile qui doit parfois être faite de façon approximative, par simulation.

4. ÉCHANTILLONNAGE SYSTÉMATIQUE AVEC PROBABILITÉ PROPORTIONNELLE À LA TAILLE

Un autre plan d'échantillonnage, qui présente les mêmes probabilités de sélection que le plan d'échantillonnage de Poisson (et l'échantillonnage par collocation) décrit dans la section qui précède, se déroule comme suit:

0) Au besoin, créer une «strate» additionnelle pour chaque base de sondage, qui réunit les unités qui ne figurent dans aucune strate du plan d'échantillonnage.

1) Diviser la population en cellules s'excluant mutuellement, par classement croisé des strates de diverses bases de sondage. Une paire d'unités dans une cellule particulière se retrouvera ainsi dans la même strate de chaque base de sondage (p. ex. la grande strate des stocks d'avoine, la strate moyenne des stocks de céréales et la strate sans sorgho).

2) Classer par ordre aléatoire les unités dans chaque cellule, puis classer les cellules dans un ordre quelconque. On obtient ainsi une liste de toutes les unités de population.

3) À partir de cette liste, prélever un échantillon systématique avec probabilité proportionnelle à la taille (PPT), en utilisant la valeur π_i décrite dans la discussion sur l'échantillonnage de Poisson comme mesure de la «taille» (le mot «taille» apparaît ici entre guillemets, car π_i n'est pas, à proprement parler, une mesure de la taille). Cette approche assure que la probabilité de sélection d'une unité est égale à π_i.

Le plan d'échantillonnage systématique PPT décrit précédemment donnera toujours un échantillon dont la taille se rapproche de $\sum_{i\in P}\pi_i$. En fait, si $\sum_{i\in P}\pi_i$ est un nombre entier, alors la taille de l'échantillon sera exactement égale à cette somme. Dans les autres cas, la taille de l'échantillon sera égale à celui des deux nombres entiers qui se rapproche le plus de $\sum_{i\in P}\pi_i$. De même, le nombre prévu d'unités échantillonnées dans une cellule, C, sera égal à $\sum_{i\in C}\pi_i$, alors que la taille réelle de l'échantillon sera, soit $\sum_{i\in C}\pi_i$, soit un des deux nombres entiers qui s'en rapprochent le plus.

Examinons maintenant une strate particulière h dans la base de sondage f, pour laquelle la taille visée de l'échantillon est égale à n_{fh}. Pour une unité i dans cette strate, $\pi_i \geq n_{fh}/N_{fh}$ en vertu du plan. Supposons que $P(fh)$ désigne la série des unités de population dans la strate fh. Le nombre prévu d'unités échantillonnées dans fh est égal à $\sum_{i\in P(fh)}\pi_i \geq n_{fh}$. Rien ne *garantit* que la taille réelle de l'échantillon dans la strate sera supérieure ou égale à n_{fh}. Cependant, compte tenu de l'inégalité précitée et des limites inférieures de la taille de l'échantillon des cellules dans fh, la taille de l'échantillon dans la strate fh ne sera jamais de beaucoup inférieure à n_{fh}.

Les avantages de ce plan d'échantillonnage, par rapport à celui de Poisson ou à l'échantillonnage par collocation, tiennent au fait que ce plan produit un échantillon de taille plus stable et qu'il accroît la probabilité d'atteindre les exigences relatives à la base de sondage ou à la strate. Ces objectifs relatifs à la base de sondage ou à la strate sont en revanche toujours atteints avec l'échantillonnage fondé sur les NAP avec échantillon fixe, mais ceci a un prix: dans l'ensemble, la taille globale de l'échantillon est moins stable et les probabilités de sélection peuvent être très difficiles à déterminer.

5. ÉVALUATION DES AUTRES TECHNIQUES D'ÉCHANTILLONNAGE

Afin d'évaluer ces techniques d'échantillonnage empiriquement, nous avons choisi trois États dans lesquels est menée l'enquête sur l'utilisation de produits chimiques dans la production maraîchère du NASS, et nous avons répété 100 fois les trois techniques fondées sur les NAP, la méthode d'échantillonnage systématique PPT et l'échantillonnage indépendant, toutes bases de sondage confondues. Pour chaque répétition, les mêmes NAP ont été utilisés pour les trois techniques basées sur les NAP. Une base de sondage distincte a été créée pour chaque denrée à l'étude à l'intérieur d'un État (le nombre de bases de sondage varie de 2, au Minnesota, à 23 en Californie). Les unités de population ont été réparties comme suit entre l'une des quatre strates, dans chacune des bases de sondage: deux strates probabilistes, une strate à tirage complet et une strate nulle. Les limites des states ont été déterminées au moyen d'une méthode modifiée de Lavallée et Hidiroglou (1988), et les unités ont été réparties entre les strates selon la règle $\text{cum}^3\sqrt{f(x)}$ (Sweet et Sigman 1995). Cette stratification a été choisie, pour reproduire ce que serait un plan d'échantillonnage unidimensionnel acceptable ou assez répandu.

Un échantillon correspondant au tiers de la population a été prélevé de chacune des strates probabilistes. Le tableau 1 présente une comparaison de la taille globale de l'échantillon obtenu par chaque technique d'échantillonnage. Comme il fallait s'y attendre, le plan fondé sur la base de sondage indépendante a produit les échantillons les plus gros. Pour leur part, les trois techniques fondées sur les NAP ont donné des échantillons de taille similaire, la méthode de Poisson donnant les écarts-types les plus élevés lors de chacun des trois essais (États). Il semble que la méthode PPT soit la plus stable.

Le tableau 2 présente le pourcentage des échantillons prélevés des strates par la méthode de Poisson et la méthode PPT et dont la taille a été inférieure aux objectifs visés. Le fait qu'il n'y ait pas eu davantage de problèmes avec les tailles réelles par la méthode de Poisson s'explique notamment par la présence de ce que nous appelons les «visiteurs». Un visiteur est une unité qui n'a pas été choisie

dans la base de sondage d'un produit particulier, mais qui se retrouve néanmoins dans l'échantillon du fait qu'elle a été sélectionnée dans une autre base de sondage. La présence de ces visiteurs a tendance à produire des échantillons dont la taille est en moyenne supérieure à la taille visée.

Tableau 1
Tailles réelles moyennes de l'échantillon (plus de 100 répétitions)

État	Base de sondage indépendante	Méthode avec échantillon fixe	Méthode NAP de Poisson	Méthode NAP par collocation	Méthode systématique PPT
CA	496	388	375	374	373
	(8,8)	(9,6)	(11,1)	(5,6)	(0,14)
MI	658	513	504	501	502
	(9,3)	(9,2)	(13,6)	(6,0)	(0,48)
NJ	563	359	343	344	343
	(8,1)	(8,6)	(13,8)	(4,6)	(0,17)

La taille de la population s'établit comme suit: CA-775; MI-1 041; NJ-785.
Les écarts-types sont indiqués entre parenthèses.

Tableau 2
Pourcentage des strates probabilistes pour lesquelles la taille réelle de l'échantillon est inférieur à l'objectif visé (100 répétitions)

État	Méthode NAP de Poisson	Méthode NAP par collocation	Méthode systématique PPT
CA	11 %	11 %	6,3 %
MI	12 %	12 %	6,3 %
NJ	11 %	8 %	1,4 %

La figure 1 (voir p.7) présente les distributions cumulatives des différences entre la taille réelle et la taille visée de l'échantillon; ces différences sont exprimées en pourcentages de la taille visée pour la strate échantillonnée; il s'agit, en d'autres mots, de la distribution cumulative de la taille (réelle-visée)/visée, au niveau de la strate probabiliste. Au Michigan, par exemple, il y a eu formation de 13 bases de sondage pour les denrées, chacune étant composée de deux strates probabilistes. L'échantillonnage à partir de ces bases de sondage a été répété 100 fois, de sorte que la fonction de distribution cumulative (FDC) pour chaque technique a utilisé 2 600 points. Les résultats obtenus avec les deux méthodes de Poisson sont représentés par une seule ligne, car ils coïncident. La méthode de Poisson ne produit pas un suréchantillonnage aussi marqué que ceux obtenus par les méthodes basées sur un échantillon fixe ou sur la base de sondage indépendante, mais elle comporte en revanche un risque de sous-échantillonnage, comme on le remarque au tableau 2. À l'inverse, les techniques basées sur un échantillon fixe (avec bases de sondage dépendante et indépendante) ne donnent pas lieu à un sous-échantillonnage, mais elles occasionnent davantage de suréchantillonnage que les méthodes de Poisson et PPT. Par ailleurs, la méthode PPT s'accompagne d'un certain sous-échantillonnage, qui n'est toutefois pas comparable à celui associé à la méthode de Poisson. Le plan d'échantillonnage PPT est également celui qui affiche la pente la plus prononcée de toutes les FDC, ce qui signifie qu'il y a moins de suréchantillonnage.

Avec la méthode de Poisson et l'échantillonnage par collocation, la probabilité de sélection de l'unité i est $\pi_i = \max_f(p_{fh})$ où h représente la strate dans laquelle se trouve l'unité i pour la base de sondage f. La même probabilité de sélection est utilisée pour la technique PPT. Par contre, les probabilités de sélection avec la méthode basée sur les NAP avec échantillon fixe sont difficiles à déterminer et elles doivent parfois être simulées.

Pareille simulation a été faite avec les données de la Californie. La technique avec échantillon fixe a été répétée 10 000 fois. Comme toutes les strates probabilistes ont été échantillonnées à un taux de 1/3, les probabilités simulées (c.-à-d. les fréquences relatives) peuvent être comparées à 1/3. Les probabilités de sélection simulées moyennes pour les 10 000 essais sont illustrées à la figure 2 par une fonction du nombre de bases de sondage dans lesquelles l'unité se trouve dans une strate probabiliste. Il existe 19 denrées d'intérêt dans cet État mais, dans exactement 16 ou 19 bases de sondage, il n'y avait aucune unité dans les strates probabilistes. Or la probabilité de sélection d'une unité a tendance à augmenter parallèlement au nombre de strates probabilistes qui la renferment. Cette probabilité de sélection est de 1/3 seulement lorsque l'unité est exactement dans une de ces strates.

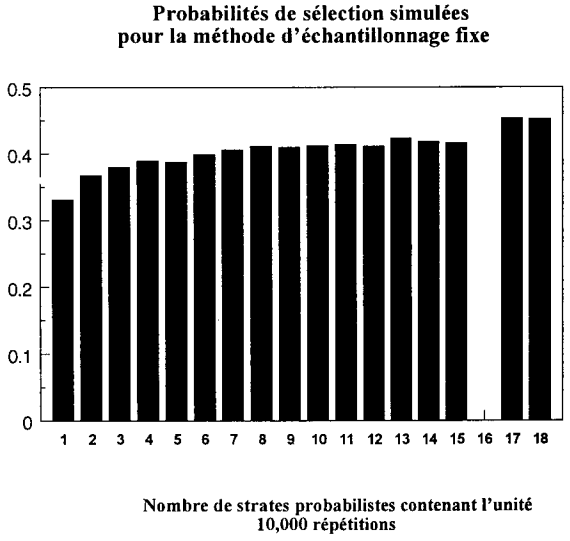

Figure 2. Probabilités de sélection simulées pour la méthode d'échantillonnage fixe-Californie

6. CALAGE

Le problème qui se pose, autant avec t_M qu'avec t_P (ou t_{HT}), est qu'ils ne sont souvent pas de très bons estimateurs de T en termes de précision (variance). Une des propriétés de l'échantillonnage aléatoire simple stratifié à base de sondage unique est que l'estimateur classique avec facteur d'extension estime parfaitement la taille de la strate (c.-à-d. avec une variance nulle). Or avec notre plan fondé sur des bases de sondage multiples, ni t_M ni t_P ne pourra, dans la plupart des applications, estimer N_{fh} à la perfection.

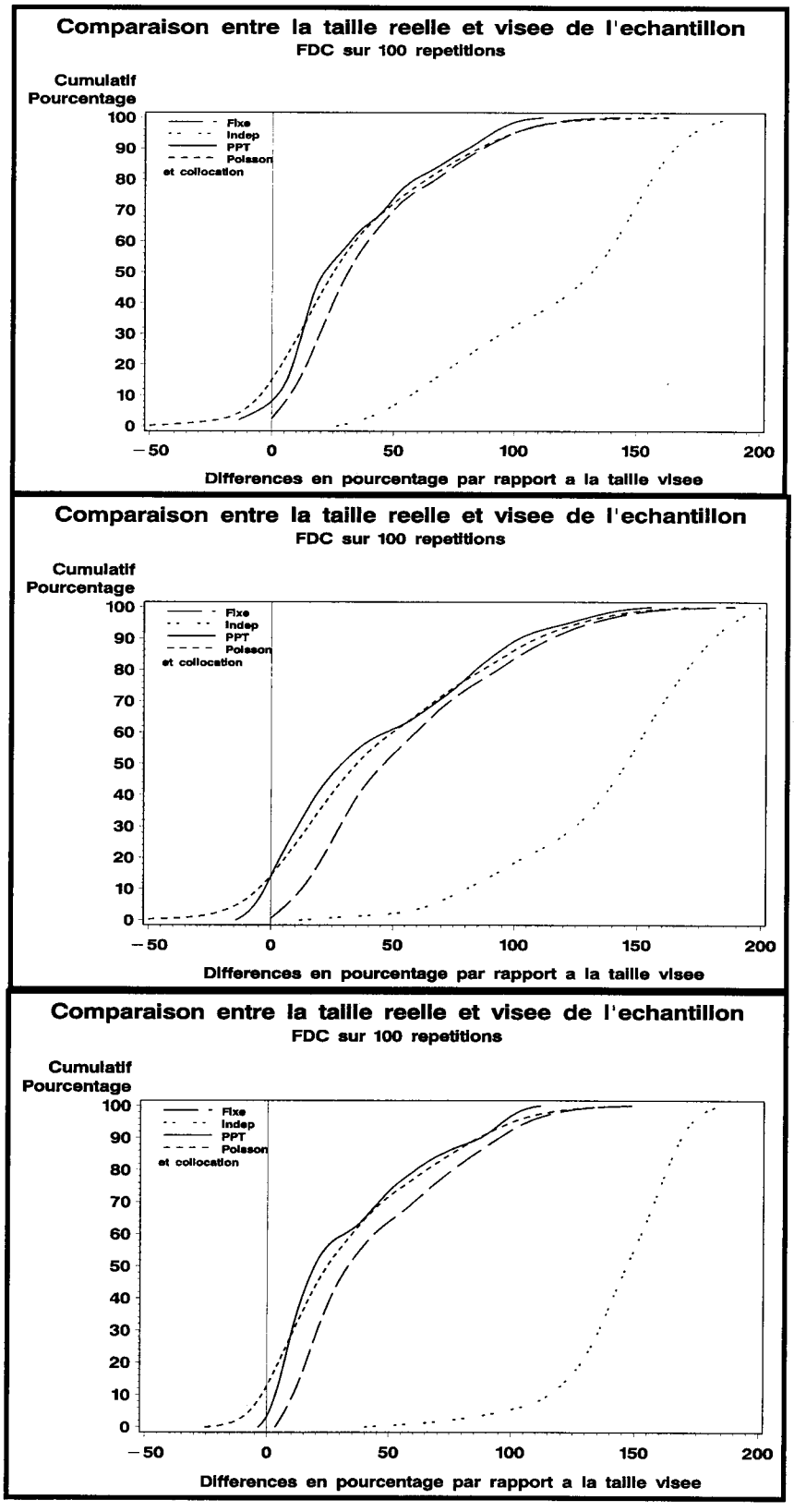

Figure 1. Comparaison entre la taille réelle et visée de l'échantillon pour les strates échantillonées. Ligne supérieure - MI; ligne médiane - CA; ligne inférieure - NJ

Supposons que $w_i^0 = n_{(i)}/E[n_{(i)}]$ est le *poids d'échantillonnage initial* de l'unité i dans t_M. De même, $w_i^0 = 1/\max_f\{p_{if}\}$ dans t_P et $1/\pi_i$ de façon plus générale, pour un estimateur de Horvitz-Thompson. Bankier (1986) propose d'utiliser la méthode itérative du quotient pour créer une série de poids corrigés, de sorte que

$$\sum_{i \in S_{fh}} w_i^C = N_{fh} \tag{2}$$

pour chaque strate h dans chaque base de sondage f, où S_{fh} est la partie de l'échantillon qui se trouve dans la strate h de la base de sondage f, quelles que soient la ou les bases de sondage d'où les unités ont été sélectionnées.

Deville et Särndal (1992) parlent d'une *équation de calage* (2). Ils soulignent qu'il existe diverses façons de calculer les poids de calage, w_i^C, de manière à ce que l'équation (2) se vérifie et que la valeur de w_i^C/w_i^0 se rapproche, dans une certaine mesure, de 1 pour toutes les unités i. L'une de ces méthodes est la méthode itérative du quotient proposée par Bankier (1986). Une autre méthode, décrite en détail par Deville et Särndal (1992), est basée sur les moindres carrés. Cependant, que l'on utilise l'une ou l'autre méthode, nous obtenons l'estimateur

$$t_C = \sum_{i \in S} w_i^C y_i,$$

où S représente l'échantillon complet, estimateur qui sera pratiquement sans biais puisque w_i^C/w_i^0 se rapproche de 1 pour toutes les unités i.

L'estimateur t_C est également sans biais sous le modèle:

$$y_i = \beta_0 + \sum_{f=1}^{F} \sum_{h=2}^{H_f} d_{ifh}\,\beta_{fh} + \epsilon_i, \tag{3}$$

où la variable fictive, d_{ifh}, est égale à 1, lorsque l'unité i se trouve dans la strate h de la base de sondage f (échantillonnée ou non) et est égale à zéro dans les autres cas, alors que ϵ_i est une variable aléatoire avec une valeur moyenne de zéro. β_0 et β_{fh} sont des constantes inconnues (β_0 représente la valeur moyenne de y pour une unité dans la première strate de chaque base de sondage; c'est pourquoi la deuxième somme exclut $h = 1$). Les mêmes valeurs de d_{ifh}, s'appliquent à chaque question d'enquête (y) d'intérêt, tandis que la valeur de β change pour chaque question. Pour bon nombre de questions, la valeur de β_{fh} sera égale à zéro lorsque la base de sondage f (disons les stocks de céréales) n'a aucun rapport avec la question (les superficies ensemencées avec de l'avoine, par exemple).

Isaki et Fuller (1982) parlent de la valeur probable de l'erreur quadratique moyenne de t_C comme étant l'«erreur quadratique moyenne prévue» de l'estimateur. Cette valeur est très utilisée durant la phase de planification d'un sondage.

Si le modèle dans l'équation (3) se vérifie et que les ϵ_i sont sans corrélation, alors l'erreur quadratique moyenne prévue de t_C est

$$\begin{aligned}
E_\epsilon[\mathrm{EQM}_D(t_C)] &= E_\epsilon\{E_D[\textstyle\sum_s w_i^C y_i - \sum_P y_i)^2]\} \\
&= E_D\{E_\epsilon[(\textstyle\sum_s w_i^C y_i - \sum_P y_i)^2]\} \\
&= E_D\{E_\epsilon[(\textstyle\sum_s w_i^C \epsilon_i - \sum_P \epsilon_i)^2]\} \\
&= E_D\{\textstyle\sum_s [(w_i^C)^2 - 2w_i^C]E_\epsilon(\epsilon_i^2)\} + \sum_P E_\epsilon(\epsilon_i^2) \\
&\approx E_D\{\textstyle\sum_s [(1/\pi_i)^2 - 2/\pi_i]E_\epsilon(\epsilon_i^2)\} + \sum_P E_\epsilon(\epsilon_i^2) \\
&= \textstyle\sum_P (1/\pi_i - 1)E_\epsilon(\epsilon_i^2),
\end{aligned} \tag{4}$$

puisque $w_i^C \approx 1/\pi_i$. Il est intéressant de souligner que, quelle que soit la méthode d'échantillonnage utilisée (méthode de Poisson, échantillonnage par collocation ou échantillonnage systématique PPT), on obtient des estimateurs dont l'erreur quadratique moyenne prévue est asymptotiquement à peu près égale. Ce résultat surprenant tient en partie à la nature même de l'estimateur par calage, mais il résulte également du fait que, lorsque nous prenons la valeur probable de la variance approximative du modèle, à la dernière ligne de l'équation (4), nous faisons la moyenne de tous les échantillons possibles et supprimons ainsi la principale source de variation entre les trois plans d'échantillonnage.

Supposons maintenant que nous avions utilisé l'échantillonnage aléatoire simple stratifié et que nous avions sélectionné l'unité i dont la probabilité est $p_{if} \le \pi_i$, où f désigne la base de sondage qui se rapporte à y. On constate sans difficulté que la variance prévue de l'estimateur simple avec facteur d'extension aurait alors été $\sum_P (1/p_{if} - 1)E_\epsilon(\epsilon_i^2)$, ce qui donne une valeur pour le moins aussi élevée que le côté droit de l'équation (4). Il est donc avantageux – du moins pour les grands échantillons – d'intégrer les échantillons de diverses bases de sondage, comme nous l'avons fait ici. Nous ignorons toutefois quelle doit être, en pratique, la taille de l'échantillon pour que les résultats asymptotiques soient pertinents. Nous savons cependant que la taille de l'échantillon doit, tout au moins, être bien des fois supérieure au nombre de paramètres du modèle dans l'équation (3).

Il convient d'apporter quelques précisions sur l'estimation de l'erreur quadratique moyenne de t_C. L'estimateur de l'erreur quadratique moyenne, préconisé par Deville et Särndal (1992) – un estimateur qui allie à la fois un bon plan et des propriétés basées sur le modèle – ne peut être appliqué, à moins que la probabilité de sélection conjointe (π_{ij}) pour chaque paire d'unités d'échantillonnage (i et j) soit connue. Or, de tous les plans d'échantillonnage dont nous avons discuté, ces probabilités ne se calculent facilement

que pour la variante de Poisson de la méthode fondée sur les NAP (où $\pi_{ij} = \pi_i \pi_j$).

Comme nous l'avons remarqué avec l'équation (4), l'erreur quadratique moyenne prévue de l'estimateur par calage est la même, que l'on utilise la méthode d'échantillonnage de Poisson fondée sur les NAP, la méthode NAP par collocation ou l'échantillonnage systématique PPT. Ces résultats portent à croire que l'estimateur de l'erreur quadratique moyenne par la méthode de Poisson pourrait être acceptable avec chacun des trois plans d'échantillonnage. Il existe, à l'appui de cette allégation, un autre argument plus solide axé sur le modèle, mais il ne sera pas invoqué ici.

7. DISCUSSION

Dans la section qui précède, nous avons indiqué que, si les poids de calage étaient conçus de manière à vérifier l'équation (2), alors l'estimateur obtenu serait sans biais dans le modèle de l'équation (3). Pour bon nombre d'applications, toutefois, il se pourrait qu'un autre modèle de calage convienne mieux que celui présenté dans l'équation (3). À titre d'exemple, si une variable de contrôle continue était utilisée pour la stratification d'une base de sondage particulière, il serait alors plus plausible d'utiliser cette variable directement dans le modèle, plutôt qu'indirectement par le biais des identificateurs de la base de sondage ou de la strate.

La méthode itérative du quotient est une forme de calage dans un modèle particulier. Il est donc tout indiqué d'utiliser le modèle le plus raisonnable qui existe. Par comparaison avec la méthode itérative du quotient, la méthode des moindres carrés a l'avantage de pouvoir être facilement appliquée aux variables de contrôle continues. Singh et Mohl (1996) présentent un examen détaillé d'autres algorithmes de calage, incluant une extension de la méthode itérative du quotient aux variables continues. Brewer (1994) propose une variante intéressante de la méthode des moindres carrés, qui ne figure pas dans Singh et Mohl (1996).

Bon nombre des enquêtes économiques et agricoles utilisent des plans d'échantillonnage avec renouvellement, ce qui s'est avéré un moyen efficace d'assurer un équilibre entre les coûts et le fardeau de réponse. Bien que nos résultats empiriques privilégient l'échantillonnage systématique PPT pour les objectifs relatifs à la taille de l'échantillon, les trois plans d'échantillonnage fondé sur les NAP se prêtent beaucoup mieux au renouvellement de l'échantillon. (Voir par exemple Ohlsson (1995) à ce sujet.) Qui plus est, les méthodes fondées sur les NAP permettent d'intégrer différentes bases de sondage, à différentes périodes de l'année (avec l'échantillonnage systématique PPT, il est difficile de réaffecter l'échantillon à la base de sondage initiale). Ceci est une caractéristique particulièrement utile pour les enquêtes agricoles, du fait que les saisons de croissance varient d'une culture à l'autre.

En résumé, le plan d'échantillonnage fondé sur les NAP avec échantillon fixe est excellent pour ce qui est d'atteindre les objectifs relatifs à la taille, mais il est difficile à utiliser en pratique parce que les probabilités de sélection sont habituellement inconnues et qu'elles doivent être simulées. Pour sa part, le plan d'échantillonnage systématique PPT est lui aussi très bon pour atteindre la taille visée, mais il est difficile à intégrer dans un plan avec renouvellement de l'échantillon. De plus, l'estimation de l'erreur quadratique moyenne requiert la formulation d'hypothèses relatives au modèle. Notre exemple empirique montre que l'échantillonnage par collocation donne des résultats qui ne sont que légèrement supérieurs à ceux obtenus par la méthode de Poisson, pour ce qui est des objectifs relatifs à la taille de l'échantillon. Il convient toutefois de préciser que des résultats différents pourraient être obtenus avec d'autres configurations des bases de sondage, des strates et des taux d'échantillonnage. L'échantillonnage par collocation se prête également aux plans avec renouvellement, comme l'échantillonnage de Poisson. Cependant, comme l'échantillonnage PPT, le premier requiert la formulation d'un modèle pour estimer l'erreur quadratique moyenne.

Enfin, l'établissement de cibles pour p_{if} ou n_{if} constitue une méthode populaire, mais indirecte, de contrôler la variance de l'estimateur t_C associé à chaque base de sondage. Ce sont ces objectifs qui nous ont amené à prendre la décision ponctuelle de poser π_i égal à $\max_f \{p_{if}\}$. Une stratégie plus directe serait de définir des objectifs de variance prévus (asymptotiques) pour l'estimateur de chaque base de sondage, en utilisant l'équation (4) et les valeurs prévues pour $E_\epsilon(\epsilon_i^2)$. On pourrait alors choisir, disons la série de π_i qui réduit au minimum la taille prévue de l'échantillon, tout en satisfaisant à ces objectifs de variance. Une approche similaire est adoptée par Amrhein, Fleming et Bailey (1997), qui utilisent l'algorithme de Chromy d'une manière analogue à Sigman et Monsour (1995). La méthode d'échantillonnage de Poisson fondée sur les NAP, la méthode par collocation basée sur les NAP et l'échantillonnage systématique PPT demeurent trois solutions viables pour la sélection de l'échantillon, lorsque les π_i optimales ont été déterminées.

BIBLIOGRAPHIE

AMRHEIN, J.F., FLEMING, C.M., et BAILEY, J.T. (1997). Determining the probabilities of selection in a multivariate probability proportional to size sample design. Dans *Recueil Symposium 97: Nouvelles orientations pour les enquêtes et les recensements*, Statistique Canada. À paraître.

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

BREWER, K.R.W. (1994). Survey sampling inference: some past perspectives and present prospects. *Pakistan Journal of Statistics*, 10(1)A, 213-233.

DEVILLE, J-C., et SÄRNDAL, C.-E. (1992). Calibration estimator in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

ISAKI, C.T., et FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

KOTT, P.S., et FETTER, M.J. (1997). A multi-phase sample design to co-ordinate surveys and limit response burden. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. À paraître.

LAVALLÉE, P., et HIDIROGLOU, M. (1988). Sur la stratification de population asymétriques. *Techniques d'enquête*, 14, 35-45.

OHLSSON, E. (1995). Coordination of samples using permanent random numbers. Dans *Business Survey Methods* (Éds. B.G. Cox, D.A. Binder, N. Chinnappa, A. Christianson, M.J. Colledge, et P.S. Kott). New York: Wiley, 153-169.

SINGH, A.C., et MOHL, C.A. (1996). Comprendre les estimateurs de calage dans les enquêtes par échantillonnage. *Techniques d'enquête*, 22, 107-116.

SIGMAN, R.S., et MONSOUR, N. J. (1995). Selecting samples from list frames of businesses. Dans *Business Survey Methods* (Éds. B.G. Cox, D.A. Binder, N. Chinnappa, A. Christianson, M.J. Colledge, et P.S. Kott). New York: Wiley, 133-152.

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

SKINNER, C.J., HOLMES, D.J., et HOLT, D. (1994). Multiple frame sampling for multivariate stratification. *Revue Internationale de Statististique*, 62, 3, 333-347.

SWEET, E., et SIGMAN, R.S. (1995). *User Guide for the Generalized* SAS *Univariate Stratification Program*, Economical Statistical Methods and Programming Division, Bureau of the Census, U.S. Department of Commerce, n° rapport ESM-9504.