Techniques d'enquête, décembre 1978
Vol. 4, nº 2
Statistique Canada, nº 12-001-X au catalogue

Certaines méthodes de mise à jour des bases de sondage[1] pour les enquêtes et leurs effets sur l'estimation

J.D. Drew, G.H. Choudhry et G.B. Gray[2]

Résumé

Les bases conçues pour des enquêtes permanentes servent parfois à effectuer des enquêtes spéciales qui nécessitent un échantillon distinct de celui de l'enquête permanente. Cet article présente une méthode sans biais de mise à jour d'une base de sondage, qui prolonge celle de Keyfitz (1951) en l'appliquant au cas où une partie de la base a été réservée à des enquêtes autres que l'enquête permanente. Une autre méthode, simple mais biaisée, est aussi exposée.

Les auteurs élargissent ensuite la portée de la technique de Platek et Singh (1975) sur la conception d'un échantillon permanent à partir d'une base aréolaire nécessitant des mises à jour, en incorporant à cette technique le renouvellement des unités d'échantillonnage de premier degré, l'établissement d'une base réservée aux enquêtes spéciales et des procédures visant à éviter de tirer deux fois la même unité finale.

Pour évaluer les méthodes proposées, les auteurs appliquent la méthode de Monte Carlo à des données du recensement, en simulant le plan de sondage de l'Enquête sur la population active.

Mots-clés : Mise à jour de l'échantillon; renouvellement de l'échantillon; base aréolaire.

1 Introduction

Les enquêtes par sondage intègrent souvent des plans qui reposent sur des probabilités inégales de sélection d'unités dans les strates. Étant donné que de nombreuses caractéristiques sont fortement corrélées avec la taille relative des unités, les estimations fondées sur de tels plans sont en général plus efficaces que les estimations fondées sur des plans où la taille des unités n'est pas prise en compte. Dans les enquêtes permanentes, la taille des unités d'échantillonnage peut changer au fil du temps en raison des naissances et des décès dans les unités finales d'échantillonnage (p. ex. la construction ou la démolition de logements dans le cas des enquêtes auprès des ménages). Un taux de croissance uniforme parmi les unités d'échantillonnage entraîne une diminution de la corrélation entre les caractéristiques mesurées à partir de l'enquête

1. Adapté d'un document présenté lors de la réunion annuelle de l'American Statistical Association, qui a eu lieu à San Diego en Californie aux États-Unis du 14 au 17 août 1978.
2. J.D. Drew, G.H. Choudhry, et G.B. Gray, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada.

et les mesures de la taille, ce qui se traduit par des estimations moins efficaces qu'au cours de la période initiale.

Dans le cas des plans de sondage fondés sur des bases aréolaires, une solution au problème des tailles relatives provenant des données réside dans leur vérification périodique par des dénombrements sur le terrain prévus à des intervalles réguliers, suivis d'une révision des probabilités de sélection, et enfin, d'un changement nécessaire dans l'échantillon pour refléter les nouvelles probabilités. Keyfitz [4] a présenté une méthode selon laquelle les probabilités de sélection révisées pouvaient être intégrées tout en maximisant la probabilité de conserver l'unité échantillonnée à l'origine dans une strate. Plus récemment, Kish et Scott [5] ont adapté la procédure de Keyfitz à d'autres cas, en particulier des cas où les unités sont transférées d'une strate à une autre. Le principal inconvénient des méthodes mentionnées ci-dessus est qu'elles ne peuvent être appliquées qu'aux plans d'échantillonnage dans lesquels une unité est sélectionnée par strate. Cela laisse entendre qu'il est impossible d'obtenir des estimations de variance sans biais.

Rao, Hartley et Cochran [7] ont conçu une procédure d'échantillonnage appelée méthode des groupes aléatoires dans laquelle les estimations sans biais et leurs variances peuvent être obtenues en sélectionnant une unité par groupe aléatoire. Comme l'ont suggéré Platek et Singh [6], la procédure de mise à jour de Keyfitz peut être appliquée à chaque groupe aléatoire.

À la section (2), nous présentons une extension sans biais de la procédure de mise à jour de l'échantillon de Keyfitz [4] au cas où une unité de premier degré (UPD) est sélectionnée par strate avec une probabilité inégale, mais où une partie de l'UPD, à l'exclusion de la strate sélectionnée, est réservée exclusivement à une utilisation spéciale pour l'enquête. Les unités sont réservées en appliquant un certain mécanisme de probabilité connu, et au moment de l'actualisation de l'échantillon, l'enquête permanente est limitée à la partie non réservée de la base. La méthode intègre les « probabilités de travail » en suivant une approche semblable à celle utilisée par Fellegi [1] dans sa méthode de sélection où la probabilité est proportionnelle à la taille sans remise.

À la section (3), nous étendons l'étude de la stratégie d'actualisation à un échantillon rotatif dans lequel la méthode des groupes aléatoires

est appliquée. Après avoir sélectionné une unité avec probabilité proportionnelle à la taille (PPT) dans chaque groupe aléatoire pour l'enquête permanente, une portion déterminée des unités restantes dans chaque groupe est réservée au moyen de l'échantillonnage aléatoire simple sans remise (EASSR) pour les enquêtes spéciales. Pour le mode de renouvellement propre à l'étude, il est démontré que lorsque les unités sont réservées de la manière qui vient d'être mentionnée, les probabilités de sélection pour l'enquête permanente demeurent intactes avant l'actualisation. La procédure d'actualisation impartiale de la section (2) est adaptée au mode de renouvellement. Comme solution de rechange, une procédure d'actualisation biaisée, qui se rapproche des probabilités de travail par les probabilités de sélection révisées, est envisagée.

À la section (4), les unités réservées de chaque groupe aléatoire au sein d'une strate sont fusionnées pour former une base de sondage spéciale. La méthode systématique avec PPS randomisée de Hartley et Rao [3] est utilisée pour sélectionner des échantillons de la base de sondage spéciale, et une procédure d'estimation pour les enquêtes spéciales est décrite.

À la section (5), nous présentons les résultats d'une étude de Monte Carlo fondée sur la méthode des groupes aléatoires. Cette méthode est utilisée dans le cadre de l'Enquête sur la population active canadienne dans les régions autoreprésentatives.

2 Actualisation de l'échantillon lorsqu'une partie de la base de sondage est réservée : (cas sans renouvellement)

Prenons une strate qui comporte N unités d'échantillonnage de premier degré. Une mesure de taille X_i est associée à la $i^{ème}$ unité dans la strate; i = 1, 2, ..., N. Une unité de la strate est sélectionnée pour une enquête permanente avec ppt où p_i, la probabilité de sélectionner l'unité i pour l'enquête permanente, est donnée par :

$$p_i = X_i / \sum_{i=1}^{N} X_i ; \qquad i=1,2,...,N.$$

Nous supposons qu'il n'y a pas de renouvellement des UPD pour l'enquête permanente. À la suite de la sélection initiale d'une unité pour l'enquête permanente, une partie des UPD restantes est réservée à des enquêtes spéciales, selon un mécanisme de probabilité inconnu. Au moment de l'actualisation de l'échantillon, l'enquête permanente est limitée à la partie non réservée de la base de sondage. Supposons que s désigne un ensemble de n unités réservées pour les enquêtes spéciales et que S correspond à un tel ensemble (notons que S est une fonction alors que s est une réalisation), alors Pr(s) désigne la probabilité de réserver les ensembles d'unités dans n'importe quel ordre. Supposons que C désigne l'enquête permanente. Nous avons :

$$\begin{aligned} \Pr(s) &= \sum_{j \notin S} \Pr(j \text{ sélectionné pour C}) \; . \; \Pr(s|j \text{ sélectionné pour C}) \\ &= \sum_{j \notin S} P_j . \Pr(s|j \text{ sélectionné pour C}). \end{aligned} \tag{2.1}$$

La seule restriction imposée aux méthodes de réservation des unités est que le calcul de Pr(s) devrait être pratique.

Au moment de l'actualisation, des mesures de taille X_i' révisées sont obtenues pour chaque unité i = 1, 2, ..., N. Nous exigeons que les nouvelles probabilités de sélection pour l'enquête permanente C soient :

$$p_i' = \frac{X_i'}{\sum_{i=1}^{N} X_i'} \qquad i=1,2,\ldots,N. \tag{2.2}$$

Il convient de mentionner que les probabilités de sélection révisées pour l'enquête permanente sont limitées par la non-sélection des unités réservées. Nous définissons donc les « probabilités de travail » $p_i(2)$, i = 1, 2, ..., N, de telle sorte que la probabilité globale de sélectionner l'unité i lorsque sa moyenne est calculée sur tous les ensembles réservés possibles de n sur (N - 1) unités à l'exclusion de l'unité i devrait être égale à p_i', c'est-à-dire :

$$\sum_{s}{}' \Pr(s) \left(\frac{p_i(2)}{1 - \sum_{j \in s} p_j(2)}\right) = p_i' \qquad i=1,2,\ldots,N \;, \tag{2.3}$$

où $\sum'_s$ désigne la somme globale possible des n-tuples non ordonnés de (N - 1) unités, à l'exclusion de l'unité i, et Pr(s) est défini par l'expression (2.1). Par conséquent, à partir de l'équation (2.3), nous avons :

$$p_i(2) = \frac{p'_i}{\sum'_s \frac{Pr(s)}{1 - \sum_{j \in s} p_j(2)}} \qquad i=1,2,\ldots,N. \tag{2.4}$$

La solution pour $p_i(2)$ peut être obtenue de façon itérative en utilisant p_i comme valeurs initiales. Il convient de mentionner qu'à mesure que N et n augmentent, des difficultés combinatoires surviennent rapidement, puisque $N \binom{N-1}{n}$ sommations sont impliquées pour chaque itération. La probabilité conditionnelle après l'actualisation de sélectionner l'unité i, compte tenu de l'ensemble des unités réservées, est la suivante :

$$\mathbb{I}'_{i|s} = \frac{p_i(2)}{1 - \sum_{j \in s} p_j(2)} . \tag{2.5}$$

La probabilité *a posteriori* que l'enquête permanente contienne la $i^{ème}$ unité comme unité sélectionnée étant donné que l'ensemble de s unités était réservé est :

$$\mathbb{I}_{i|s} = \frac{Pr(\text{i sélectionné pour C et les ensembles d'unités réservées})}{Pr(s)}$$

$$= \frac{pi \cdot (Pr(s \mid \text{i sélectionné pour C}))}{Pr(s)} . \tag{2.6}$$

Nous effectuons maintenant une actualisation de Keyfitz en fonction des unités (N - n) disponibles en comparant $\mathbb{I}'_{i|s}$ avec $\mathbb{I}_{i|s}$ pour $i \notin s$. Afin de réviser les probabilités conditionnelles $\mathbb{I}_{i|s}$ à $\mathbb{I}_{i|s}$, nous entreprenons la procédure d'actualisation de Keyfitz.

Définissons des ensembles d'unités l et D à augmentation et à diminution conditionnelles, de façon à ce que

$$i \in l \quad \text{si} \quad \mathbb{I}'_{i|s} \geq \mathbb{I}_{i|s}$$

et

$$i \in D \quad \text{autrement.}$$

Si $i \in l$, conserver l'unité. Si $i \in D$, conserver l'unité avec la probabilité $\Pi'_{i|s} / \Pi_{i|s}$, et si elle est rejetée, comme ce serait le cas avec la probabilité $\left(1 - \Pi'_{i|s}/\Pi_{i|s}\right)$, sélectionner une unité de l'ensemble l selon la probabilité qui suit :

$$\frac{\Pi'_{i|s} - \Pi_{i|s}}{\sum_{i \in l} (\Pi'_{i|s} - \Pi_{i|s})} \quad \text{pour } i \in l.$$

Alors $P_{i|s}$, la probabilité conditionnelle de sélectionner l'unité i selon la procédure de Keyfitz, compte tenu de l'ensemble s d'unités réservées, est la suivante :

$$P_{i|s} = \Pi_{i|s} \left(\frac{\Pi'_{i|s}}{\Pi_{i|s}}\right) = \Pi'_{i|s} \quad \text{pour } i \in D$$

$$P_{i|s} = \Pi_{i|s} + \sum_{j \in D} \Pi_{j|s} \left(1 - \frac{\Pi'_{j|s}}{\Pi_{j|s}}\right) \left(\frac{\Pi'_{i|s} - \Pi_{i|s}}{\sum_{i \in l} (\Pi'_{i|s} - \Pi_{i|s})}\right)$$

$$= \Pi_{i|s} + \Pi'_{i|s} - \Pi_{i|s} = \Pi'_{i|s}. \quad \text{pour } i \in l.$$

Par conséquent, lors de l'actualisation, la $i^{ème}$ unité est sélectionnée avec la probabilité conditionnelle $\Pi'_{i|s}$. En faisant la moyenne sur toutes les réserves possibles de n sur (N - l) unités, à l'exclusion de l'unité i, nous obtenons la probabilité moyenne globale, P_i pour l'unité i à sélectionner après l'actualisation, comme suit :

$$P_i = \sum_s' Pr(s)(\Pi'_{i|s})$$

$$= p'_i \quad \text{par (2.3 et 2.5)} \quad i = 1, \ldots, N.$$

Par conséquent, le modèle d'actualisation est sans biais. Étant donné qu'une seule unité est sélectionnée par strate pour l'enquête permanente,

la variance est fonction des probabilités de sélection des unités et n'est pas touchée par la réservation des unités.

3 Actualisation de l'échantillon lorsqu'une partie de la base est réservée : (cas avec renouvellement)

Les résultats de la section précédente sont appliqués à la stratégie de Platek et Singh [6] pour un échantillon continu aréolaire qui doit être actualisé. La portée de cette stratégie s'applique à une situation où l'enquête permanente comprend le renouvellement des UPD. Ici, seuls les plans d'échantillonnage autopondéré sont pris en compte pour l'enquête permanente, de sorte que lorsqu'une partie de la base a été réservée, il faut que le mécanisme de réserve n'influe pas sur les probabilités de sélection des unités pour l'enquête permanente lors du renouvellement de l'échantillon.

Par souci de simplicité, nous avons considéré comme modèle une méthode de groupes aléatoires à deux degrés avec sélection PPT des UPD (grappes), la sélection systématique des unités finales d'échantillonnage (logements) et le renouvellement de l'échantillon dans et entre les UPD; cette méthode est utilisée dans le cadre de l'Enquête sur la population active du Canada dans les grandes villes. Les résultats peuvent être généralisés aux plans comportant plus de deux degrés de sélection.

Comme nous l'avons mentionné précédemment, nous avons N unités dans une strate (groupe aléatoire) et une mesure de la taille X_i associée à chaque unité i = 1, 2, ..., N. Nous souhaitons procéder à l'échantillonnage dans la strate à un taux de 1/R. Nous définissons ensuite une fraction de sondage inverse (FSI) de la grappe sous forme de nombres entiers :

$$\left.\begin{array}{l} R_i \geq 1 \qquad i=1, 2, \ldots, N \\ \text{de sorte que} \quad \sum_{i=1}^{N} |R_i - R.\, p_i| \text{ est réduit au minimum} \\ \text{et} \quad \sum_i R_i = R \end{array}\right\} \qquad (3.1)$$

Il convient de mentionner que les fractions de sondage inverse sous forme de nombres entiers sont plus pratiques que les autres nombres pour la mise en œuvre sur le terrain et pour le renouvellement de l'échantillon.

Définissons R échantillons uniques ordonnés dans chaque groupe aléatoire comme suit :

$$j|R_i \qquad j=R_i, R_i-1, \ldots 2, 1; \qquad i=1, 2, \ldots, N$$

consistant en une grappe échantillonnée i à sous-échantillonner systématiquement au taux $1/R_i$ pour j occasions successives avant le renouvellement des UPD. Autrement dit, nous avons l'ensemble suivant de R échantillons ordonnés :

$$R_1|R_1,\ (R_1-1)|R_1,\ \ldots,\ R_N|R_N,\ \ldots,\ 1|R_N.$$

Au départ, l'un des échantillons ci-dessus est sélectionné en générant un nombre aléatoire r, $1 \le r \le R$. Supposons que l'échantillon sélectionné est $j|R_i$, où $\sum_{k=0}^{i-1} R_k < r \le \sum_{k=1}^{i} R_k$ pour certains $i \in \{1, 2, \ldots, N\}$, et $j = \sum_{k=1}^{i} R_k - r + 1$; R_o est défini comme étant nul. Ensuite, un autre nombre aléatoire r_i, $1 \le r_i \le R_i$ est généré et les échantillons systématiques déterminés par l'origine choisie au hasard r_i, $(r_i + 1) \bmod R_i, \ldots, (r_i + j - 1) \bmod R_i$ sont respectivement associés aux échantillons $j|R_i, (j-1)|R_i, \ldots, 1|R_i$[1]. Après chaque intervalle de temps constant prédéterminé, le renouvellement se fait dans l'échantillon suivant de la liste. Au moment du renouvellement vers la prochaine grappe, c'est-à-dire la grappe $i^* = (i + 1) \bmod N$, avec l'échantillon $R_{i^*} \mid R_{i^*}$, un nombre aléatoire r_{i^*}, $1 \le r_{i^*} \le R_{i^*}$ est généré, et les échantillons systématiques déterminés par l'origine choisie au hasard r_{i^*}, $(r_{i^*} + 1) \bmod R_{i^*}, \ldots (r_{i^*} + R_{i^*} - 1) \bmod R_{i^*}$ sont respectivement associés aux échantillons $R_{i^*} \mid R_{i^*}$, $(R_{i^*} - 1) \mid R_{i^*} \ldots, 1 \mid R_{i^*}$, et ainsi de suite. En pratique, les nombres aléatoires r_i, $i = 1, 2, \ldots, N$ sont tous générés au moment de l'introduction initiale de l'échantillon, et le calendrier de renouvellement est créé en fonction des échantillons systématiques réels ou des origines.

[1] $R_i \bmod R_i$ est considéré comme égal à R_i au lieu de 0. Cette convention sera adoptée tout au long du présent document.

Selon ce plan de renouvellement, la probabilité de sélectionner la grappe i à tout moment est donnée par :

$$Pr(i \varepsilon C) = R_i/R \doteq p_i \quad .$$

Étant donné que la grappe i est sélectionnée, la probabilité que chaque origine soit dans l'échantillon à tout moment est donnée par $1/R_i$, de sorte que la probabilité globale de sélectionner chaque origine est $(1 / R_i)(R_i / R)$ ou $1 / R$. Par conséquent, puisqu'il s'agit d'un plan d'échantillonnage autopondéré, si y_{ik} est le total de la caractéristique pour l'origine k dans la grappe i, alors $R\, y_{ik}$ est un estimateur sans biais du total du groupe y.

Examinons maintenant ce qu'il advient des probabilités de sélection lorsque des réserves sont faites à partir de la base, en adoptant la règle selon laquelle si l'unité qui aurait été renouvelée est réservée, le renouvellement aura lieu dans la prochaine unité non réservée. Par souci de simplicité, nous considérons le cas d'une unité réservée. Étant donné que la probabilité de sélectionner une grappe à tout moment est donnée par R_i/R, nous pouvons supposer sans perte de généralité qu'au moment $t = 0$ la grappe i se trouve dans l'enquête permanente, et qu'au moment $t \in (0, 1)$, une grappe, disons $k = i$, est réservée avec la probabilité $p^*_{k|i}$. Ensuite, à $t = 1$, l'occasion du prochain renouvellement de l'échantillon, la probabilité que la grappe i fasse partie de l'échantillon pour l'enquête permanente C, c'est-à-dire $Pr(i \in C | t = 1)$, est donnée par :

$$\begin{aligned} Pr(i \in C \mid t = 1) = {} & Pr(i \in C \mid t = 0).\ Pr\ (\text{la grappe i ne sera pas renouvelée à } t = 1) \\ & + Pr(i - 1 \in C \mid t = 0) \bullet Pr\ (\text{la grappe } i - 1 \text{ ne sera pas renouvelée à } t = 1) \bullet Pr(\text{la grappe i n'est pas réservée}) \\ & + Pr(i - 2 \mid C\ t = 0) \bullet Pr\ (\text{la grappe } i - 2 \text{ ne sera pas renouvelée à } t = 1) \bullet Pr(\text{la grappe } i - 1 \text{ n'est pas réservée}) \end{aligned}$$

$$= \frac{R_i}{R}\left(1 - \frac{1}{R_i}\right) + \frac{R_{i-1}}{R}\,\frac{1}{R_{i-1}}\left(1 - p^*_{i|i-1}\right) + \frac{R_{i-2}}{R}\,\frac{1}{R_{i-2}}\,p^*_{i-1|i-2}$$

$$= \frac{R_i - 1}{R} + \frac{1}{R}\left(1 - p^*_{i|i-1}\right) + \frac{1}{R}\,p^*_{i-1|i-2} \quad . \tag{3.2}$$

Maintenant, l'équation (3.2) est égale à R_i/R si et seulement si $p^*_{i|i-1} = p^*_{i-1|i-2}$ pour tous les i. Cette condition s'applique de manière non exclusive si une grappe est réservée avec une probabilité égale, excluant l'unité sélectionnée pour l'enquête permanente. La probabilité *a posteriori* que l'unité i soit dans l'enquête permanente C étant donné que l'unité j a été réservée est donnée par :

$$\mathbb{I}_{i|j} = \frac{\Pr(i \in C,\ j \text{ réservé})}{\Pr(j \text{ réservé})} = \frac{p_i\ \frac{1}{N-1}}{\sum_{i \neq j} p_i\ \frac{1}{N-1}} = \frac{p_i}{1 - p_j} \quad . \tag{3.3}$$

Par conséquent, l'expression de $\mathbb{I}_{i|j}$ est simplifiée si une unité est réservée avec une probabilité égale.

En général, il est possible de démontrer que lorsque n sur N - 1 grappes sont réservées avec une probabilité égale excluant la sélection de l'enquête permanente, les probabilités de sélection pour l'enquête permanente sont préservées, et l'expression de la probabilité *a posteriori* $\mathbb{I}_{i|s}$ est simplifiée comme suit :

$$\mathbb{I}_{i|s} = \frac{p_i}{1 - \sum_{j \in s} p_j} \quad . \tag{3.4}$$

Cependant, pour la même raison que nous avons choisi un mode d'échantillonnage avec PPT pour l'enquête permanente, un tel plan serait, dans la plupart des cas, avantageux pour l'enquête spéciale. Par conséquent, au lieu de sélectionner une ou plusieurs unités expressément pour une enquête spéciale particulière avec une probabilité égale

excluant la sélection pour l'enquête permanente, notre stratégie consistera plutôt à réserver une partie de la base de sondage, disons un tiers, en suivant le mécanisme ci-dessus pour la réservation des UPD, puis à sélectionner des unités pour l'enquête spéciale à partir de la partie réservée à la suite d'un modèle avec PPT.

Si les réserves sont faites de la manière mentionnée ci-dessus, il n'y aura pas de biais de sélection pour l'enquête permanente avant l'actualisation. Dans le reste de la présente section, nous montrons la façon dont la méthode générale décrite à la section (2) peut être adaptée au plan de renouvellement particulier pour atteindre les probabilités souhaitées après l'actualisation, tout en évitant les chevauchements de logements entre les échantillons avant et après l'actualisation.

Selon cette méthode de réservation des UPD, les équations (2.1) et (2.3) se réduisent respectivement à :

$$Pr(s) = \left(1 - \sum_{i \in s} p_i\right) \frac{1}{\binom{N-1}{n}} \tag{3.5}$$

et

$$\sum_{s}{}' \left(1 - \sum_{i \in s} p_i\right) \frac{1}{\binom{N-1}{n}} \left(\frac{p_i(2)}{1 - \sum_{i \in s} p_i(2)}\right) = p_i' \qquad i=1, 2, \ldots, N \quad , \tag{3.6}$$

où p_i' est défini dans l'équation (2.2).

En appliquant la procédure d'actualisation de l'échantillon de Keyfitz à l'aide des probabilités conditionnelles décrites à la section (2), une grappe $i \in s$ pourrait être sélectionnée pour l'enquête permanente, la probabilité conditionnelle $\Pi'_{i|s}$ étant donnée par :

$$\Pi'_{i|s} = \frac{p_i(2)}{1 - \sum_{j \in s} p_j(2)}$$

de sorte que lorsqu'on fait la moyenne de toutes les réserves possibles, la probabilité de sélectionner la grappe i devient p_i'. Cependant, après

avoir conservé une grappe de cette façon au moment de l'actualisation, il serait souhaitable de rester dans la grappe seulement assez longtemps pour que l'échantillonnage puisse être limité aux logements inutilisés. Cela indique une mise en correspondance (voir l'annexe A) des échantillons possibles avant l'actualisation avec les échantillons possibles après l'actualisation, de sorte qu'après le plan de renouvellement, il n'y ait pas de chevauchement des logements, et que les probabilités requises après l'actualisation sont atteintes.

La FSI de la grappe basée sur de nouvelles tailles sera définie comme elle l'a été précédemment, R_i' remplaçant R_i et p_i' remplaçant p_i, N dans l'expression (3.1).

Étant donné que nous utiliserons une mise en correspondance un à un à partir des échantillons possibles avant l'actualisation avec les échantillons possibles après l'actualisation pour effectuer l'actualisation de l'échantillon de type de Keyfitz comme décrit à l'annexe A, et qu'il pourrait n'y avoir que $\left(R - \sum_{j \in s} R_j\right)$ échantillons possibles avant l'actualisation, nous définissons la FSI de la grappe après l'actualisation comme étant les nombres entiers $R_{i|s}(2) \geq 1$ pour $i \notin s$

$$\left.\begin{array}{ll} \text{de sorte que} & \sum_{i \notin s} \left(R_{i|s}(2) - (R - \sum_{j \in s} R_j) . \Pi'_{i|s}\right) \\ & \\ \text{est réduit au minimum et que} & \sum_{i \notin s} R_{i|s}(2) = R - \sum_{j \in s} R_j . \end{array}\right\} \quad (3.7)$$

Ainsi, la grappe $i \notin s$ sera sélectionnée avec la probabilité conditionnelle $\frac{R_{i|s}(2)}{R - \sum_{j \in s} R_j}$ au lieu de $\Pi'_{i|s}$. Fait à noter, cette procédure de calcul peut uniquement donner lieu à des erreurs lors de l'arrondissement des tailles en nombres entiers. Dans l'expression (3.6), pour calculer les probabilités de travail $p_i(2)$, p_i' a été considéré comme suit :

$\frac{X_i}{\sum_i X_i}$ au lieu de R_i'/R de sorte que l'effet attribuable à l'arrondissement en nombres entiers ne soit pas introduit deux fois.

Comme nous échantillonnerons au taux $R_{i|s}(2)$ plutôt qu'au taux R_i' dans la grappe i sélectionnée, nous appliquerons un poids compensatoire égal au ratio $\frac{R_{i|s}(2)}{R_i'}$ à l'étape de l'estimation. Comme auparavant, si y_{ij} est le total de la caractéristique pour l'échantillon sélectionné k dans la grappe i, alors $R\left(\frac{R_{i|s}(2)}{R_i'}\right) y_{ik}$ est un estimateur pour le total de la strate, dont le seul biais est attribuable à l'arrondissement en nombres entiers.

En raison de la complexité du calcul des « probabilités de travail » et des limites pratiques de cette méthode, une solution de rechange simple, bien que biaisée, est présentée ici. Il a été a observé empiriquement que, lorsque $n / (N - 1) \leq 1/3$

$$p_i(2) \doteq p_i' \qquad i=1, 2, \ldots, N$$

de sorte que nous définissions maintenant la probabilité conditionnelle de sélectionner l'unité i pour l'enquête permanente C, étant donné que l'ensemble s d'unités était réservé, comme suit :

$$\pi^*_{i|s} = \frac{p_i'}{1 - \sum_{j \in s} p_j'}$$

et nous définissons le $R'_{i|s} \geq 1$ de la FSI pour $i \notin s$ en remplaçant $R_{i|s}(2)$ par $R'_{i|s}$ et $\mathbb{I}'_{i|s}$ par $\mathbb{I}^*_{i|s}$ dans l'équation (3.7).

Ensuite $R\left(\frac{R'_{i|s}}{R_i'}\right) y_{ik}$ est l'estimateur pour le total de la strate, et la mise en correspondance des échantillons avant l'actualisation avec les échantillons après l'actualisation est identique au cas précédent.

Il convient de mentionner que si le nombre d'échantillons après l'actualisation pouvait être choisi en tant que $R - \sum_{i \in s} R_i'$ au lieu de $R - \sum_{i \in s} R_i$, alors les poids $\frac{R_{i|s}(2)}{R_i'}$ en général seraient près de 1, et l'écart par rapport à un plan autopondéré serait réduit au minimum. Cependant,

la procédure de mise en correspondance pour le cas où le nombre d'échantillons avant et après l'actualisation n'est pas égal devient très compliquée. De plus, dans le cadre de cette mise en correspondance, la probabilité de conserver la grappe actuellement sélectionnée ne sera pas maximisée comme dans la méthode de Keyfitz.

4 Stratégie d'utilisation de la base de sondage spéciale

Dans une strate, les unités réservées (grappes) de chaque groupe aléatoire sont fusionnées pour former une base de sondage spéciale. Avant de présenter la méthodologie pour la base de sondage spéciale, il convient de souligner que, s'il n'était pas nécessaire de fournir une capacité d'actualisation de la base de sondage et de l'échantillon, les enquêtes autres que l'enquête permanente pourraient également reposer sur la base de sondage, éviter le chevauchement avec l'enquête permanente en espaçant simplement leurs sélections à un certain intervalle de celles de l'enquête permanente. Toutefois, au moment de l'actualisation, que ce soit au moyen de la méthode de Keyfitz ou d'une sélection indépendante, la sélection pour l'enquête permanente pourrait changer, ce qui entraînerait un conflit avec les échantillons sélectionnés pour les enquêtes spéciales. Par ailleurs, si l'enquête spéciale est limitée à la même grappe dans laquelle se trouve la sélection pour l'enquête permanente, cela peut, sur le plan opérationnel, lier l'enquête permanente et l'enquête spéciale, au point de nuire aux deux enquêtes. Par exemple, l'enquête spéciale serait liée aux délais d'exécution de l'enquête permanente pour l'introduction d'unités d'échantillonnage, tandis que l'utilisation sporadique de la base de sondage pour l'enquête spéciale aurait un effet perturbateur sur les opérations de maintien de l'échantillon pour l'enquête permanente.

Étant donné que la taille de l'échantillon peut varier pour différentes enquêtes spéciales, un plan de sondage systématique randomisé avec PPT [3] est proposé, car cette méthode est souple en ce qui concerne le nombre d'unités sélectionnées [2]. Les enquêtes spéciales successives s'appuieraient, dans la mesure du possible, sur les UPD communes pour réduire au minimum les coûts de listage; cependant, lorsque la base de sondage est mise à jour, une sélection entièrement indépendante serait

effectuée dans la base de sondage spéciale, ce qui éviterait le chevauchement pour ce qui est des logements grâce au mécanisme d'ajustement de l'ordre décrit à l'annexe (A).

Supposons que pour chaque groupe aléatoire g, nous sélectionnons les grappes avec échantillonnage aléatoire simple à partir des grappes (Ng - 1) disponibles, à l'exclusion de la sélection pour l'enquête permanente, où g = 1, 2, ..., G. Ainsi, à l'intérieur d'une sous-unité $n = \sum_{g=1}^{G} n_g$ parmi $N = \sum_{g=1}^{G} N_g$, des grappes sont réservées pour la base de sondage spéciale.

Étant donné que l'enquête permanente est plus susceptible de se faire dans de grandes grappes, la probabilité globale qu'une grappe soit réservée à la base de sondage spéciale diminue à mesure que la taille de la grappe augmente. Un plan d'échantillonnage sans biais qui tient compte de cela est probablement moins efficace qu'un plan biaisé qui suppose que la probabilité que la grappe i se trouve dans la base de sondage spéciale est égale à n/N pour tous les i. Selon cette dernière hypothèse, pour un taux d'échantillonnage global de $1/R_o$ à partir de la sous-unité, $1/W_o$ correspond au taux d'échantillonnage équivalent de la base de sondage spéciale. Alors

$$\frac{n}{N}\,(1/W_o) = 1/R_o$$

ou

$$W_o = \frac{n}{N}\,R_o.$$

Définissons $W_o' = \left[\frac{n}{N}\,R_o\right]$.

Un poids compensatoire, ω, pour compenser l'effet de l'arrondissement sera appliqué à l'estimation, où :

$$\omega = \frac{W_o'}{W_o} = \frac{W_o'}{\frac{n}{N}\,R_o}\ .$$

Ensuite, les fractions de sondage inverse pour les grappes dans la base de sondage spéciale sont définies comme étant les nombres entiers $W_i \geq 1$ pour $i \in s$ de sorte que :

$$\sum_{i \in s} W_i = W_o'$$

et

$$\sum_{i \in s} \left(W_i - W_o' \left(\frac{X_i}{\sum_{i \in s} X_i}\right)\right)$$

est réduit au minimum, ce qui subdivise la base de sondage spéciale en W_o' échantillons systématiques. La sélection de M de ces échantillons pour une enquête spéciale correspond à un taux d'échantillonnage de M/R_o pour l'ensemble de la base de sondage.

Supposons que y_m = la réponse de l'échantillon sélectionné.

Ensuite, $y = \sum_{m=1}^{M} y_m$ = la réponse totale de l'échantillon.

Deux estimateurs pour le chiffre de population sont pris en compte :

$$\hat{y}_1 = \omega R_o y/M$$

$$= \left(\frac{N}{n}\right) W_o' y/M, \qquad (4.1)$$

et

$$\hat{y}_2 = \frac{\left(\frac{X}{N}\right)}{\left(\frac{X_s}{n}\right)} \omega R_o y/M$$

$$= \left(\frac{X}{X_s}\right) W_o' y/M, \qquad (4.2)$$

où

$$X = \sum_{i=1}^{N} X_i, \quad X_s = \sum_{i \in s} X_i.$$

L’ajustement du ratio $\frac{\left(\frac{X}{N}\right)}{\left(\frac{X_s}{n}\right)}$ dans $\hat{y}_2$ compense les écarts dans la taille de la base de sondage spéciale par rapport à un sous-échantillon n/N de la base de sondage, introduit en raison de la variabilité de l’échantillonnage ainsi que du biais attribuable à l’hypothèse d’un échantillonnage aléatoire simple pour réserver des unités dans l’ensemble de la sous-unité.

Nous avons observé, dans les études de Monte Carlo, que $\hat{y}_2$ donne des résultats plus cohérents que $\hat{y}_1$; par conséquent l’estimateur pris en compte pour la base de sondage spéciale de la section (5) est $\hat{y}_2$.

5 Étude de Monte Carlo

a) Description

L’Enquête sur la population active canadienne suit un plan d’échantillonnage stratifié à plusieurs degrés [6]. Dans les régions autoreprésentatives composées de grandes villes et de régions métropolitaines, représentant plus des deux tiers du pays, un plan d’échantillonnage stratifié à deux degrés est utilisé. Les strates sont composées de sous-unités dont les populations varient de 6 000 à 25 000, tandis que les UPD (grappes) sont composées de côtés d’îlots, et les unités finales d’échantillonnage sont composées de logements.

Afin d’évaluer les gains de fiabilité des données découlant de la mise à jour des procédures et la pertinence de la procédure suggérée pour les enquêtes spéciales, une étude de Monte Carlo a été réalisée pour sept sous-unités de la population active (strates) ayant des taux de croissance variables entre les recensements de 1966 et de 1971.

Pour les secteurs de dénombrement (SD) du recensement comprenant ces sous-unités, les données du recensement de 1971 ont été obtenues au niveau individuel pour le tiers de l’échantillon des ménages qui ont reçu un questionnaire détaillé du recensement. Aux fins de l’étude, les établissements comme les hôpitaux et les foyers pour personnes âgées ont été exclus. Pour la plupart, les SD de 1971 ont été choisis pour représenter les grappes de l’Enquête sur la population active (EPA). Cependant, afin que la répartition des tailles des grappes au sein des sous-unités corresponde de près à la répartition connue des grappes par province et type de zone pour le plan de l’EPA, certains des SD plus

importants ont été subdivisés pour former deux grappes ou plus. Les nouvelles mesures de la taille ont été obtenues à partir des chiffres des ménages se rapportant au tiers de l'échantillon, tandis que les anciennes mesures de la taille correspondantes ont été obtenues en prenant le tiers des chiffres des logements pour les SD de 1966 et en utilisant les tableaux de conversion des SD de 1971 à 1966.

Dans la présente étude, nous avons considéré l'estimation des six caractéristiques suivantes :

i) Population,

ii) Nombre de ménages,

iii) Nombre de personnes employées,

iv) Nombre de personnes sans emploi,

v) Nombre de personnes inactives,

vi) Revenu total.

Cinq méthodes différentes ont été simulées 1 000 fois indépendamment au sein de chaque sous-unité. Une méthode est définie comme un modèle de sélection associé à une procédure d'estimation. Les méthodes sont décrites ci-dessous.

Méthode 1 - Méthode des groupes aléatoires fondée sur de nouvelles mesures de la taille avec base de sondage complète disponible pour l'enquête permanente.

Méthode 2 - Après la sélection comme dans la méthode 1, une portion d'un tiers de chaque groupe aléatoire a été réservée avec une probabilité égale en ne prenant pas en compte la grappe sélectionnée pour l'enquête permanente, et les grappes réservées de chaque groupe aléatoire ont été fusionnées pour former la base de sondage spéciale. Dans la base de sondage spéciale, le plan et la procédure d'estimation décrits à la section 4 ont été suivis.

Méthode 3 - Cette méthode est identique à la méthode 1, mais elle est fondée sur d'anciennes mesures de la taille.

Méthode 4 - Après la sélection faite selon la méthode 3, un tiers de chaque groupe aléatoire a été réservé, et l'échantillon a été mis à jour à l'aide du modèle de « probabilité de travail » décrit à la section 3.

Méthode 5 – Cette méthode est identique à la méthode 4, sauf que l'échantillon a été mis à jour au moyen du modèle de « probabilité révisée » décrit à la section 3.

Supposons Y_h = le total de la caractéristique pour la sous-unité h est fondé sur le Recensement de 1971; (h = 1, 2, ..., 7),

et $y_{hr}^{(m)}$ = l'estimation de Y_h à partir de la r^{ime} réplication à l'aide de la méthode m; (r = 1, 2, ..., 1 000; m = 1, 2, ..., 5).

Ensuite, la valeur moyenne de 1 000 estimations pour la méthode m, la sous-unité h, est donnée par :

$$\bar{y}_h^{(m)} = \frac{1}{1\,000} \sum_{r=1}^{1\,000} y_{hr}^{(m)} .$$

En combinant les sept sous-unités, le chiffre de population Y est donné par :

$$Y = \sum_{h=1}^{7} Y_h ,$$

et en combinant les estimations pour toutes les sous-unités, nous avons :

$$y_r^{(m)} = \sum_{h=1}^{7} y_{hr}^{(m)}$$

et

$$\bar{y}^{(m)} = \sum_{h=1}^{7} \bar{y}_h^{(m)}$$

$$= \frac{1}{1\,000} \sum_{r=1}^{1\,000} y_r^{(m)}$$

Définissons l'écart de la méthode m, $D^{(m)}$, comme l'écart de la moyenne de 1 000 estimations, à l'aide de la méthode m, par rapport à la population totale y, c'est-à-dire :

$$D^{(m)} = \bar{y}^{(m)} - y,$$

et le pourcentage de l'écart relatif (ÉR) :

$$\text{ÉR}^{(m)} = 100(\bar{y}^{(m)} - y)/y.$$

L'estimation de l'écart-type (É.-T.) de $y_{hr}^{(m)}$ est :

$$\hat{\text{É.-T.}}(y_{hr}^{(m)}) = \left[\frac{1}{1\,000} \sum_{r=1}^{1\,000} (y_{hr}^{(m)} - \bar{y}_h^{(m)})^2\right]^{1/2}$$

Par conséquent, l'estimation de l'écart-type de $y_r^{(m)}$ est :

$$\hat{\text{É.-T.}}(y_r^{(m)}) = \left(\sum_{h=1}^{7} [\hat{\text{É.-T.}}(y_{hr}^{(m)})]^2\right)^{1/2},$$

et l'estimation de l'écart-type de $\bar{y}^{(m)}$ est :

$$\hat{\text{É.-T.}}(\bar{y}^{(m)}) = \hat{\text{É.-T.}}(y_r^{(m)})/(1\,000)^{1/2}.$$

Le pourcentage estimé du coefficient de variation (C.V.) est alors :

$$\hat{\text{C.V.}}(\bar{y}^{(m)}) = 100\,\hat{\text{É.-T.}}(\bar{y}^{(m)})/\bar{y}^{(m)}$$

Dans la sous-unité h, définissons l'efficacité de la méthode m par rapport à la méthode 1 comme suit :

$$\text{EFF}_h\ (m \text{ vs } 1) = 100\ (\text{EQM})_h^{(1)}/(\text{EQM})_h^{(m)}$$

où

$$(\text{EQM})_h^{(m)} = \frac{1}{1\,000} \sum_{r=1}^{1\,000} (y_{hr}^{(m)} - y_h)^2.$$

Enfin, définissons l'efficacité de la méthode m par rapport à la méthode 1 comme suit :

$$EFF(m \text{ vs } 1) = 100\ (EQM)^{(1)}/(EQM)^{(m)}$$

où

$$EQM^{(m)} = [\hat{E}.-T.(y_r^{(m)})]^2 + (\sum_{h=1}^{7} D_h^{(m)})^2.$$

b) Analyse des résultats

Bien que l'objectif principal de l'étude ait été d'évaluer les deux systèmes d'actualisation (c.-à-d. les méthodes 4 et 5) et le rendement de la base de sondage spéciale proposée, il a également été possible d'étudier les gains découlant de l'actualisation de l'échantillon lorsque la base complète est disponible. Examinons donc brièvement ces gains.

Nous pouvons observer dans les tableaux (5.1) et (5.2) qu'à l'exception de la caractéristique des personnes au chômage, qui n'est pas très fortement corrélée avec les mesures de la taille, l'efficacité a tendance à diminuer (donc les gains ont tendance à augmenter) avec la corrélation décroissante entre les anciennes et les nouvelles mesures de la taille. Alors que nous pourrions nous attendre à ce que, dans la pratique, plus le taux de croissance est élevé, plus cette corrélation est faible, les sous-unités 83,112 et 95,135 ne confirment pas ces attentes. Même dans les secteurs où la croissance globale est relativement modérée, des gains substantiels dans les estimations d'enquêtes simples peuvent découler de l'actualisation, comme le démontre la sous-unité 51 201. Cependant, en raison de l'efficacité des techniques couramment utilisées dans les procédures d'estimation pour les enquêtes à grande échelle, comme la stratification *a posteriori* selon les catégories d'âge et de sexe, les gains de précision pour les estimations finales de l'enquête seront probablement plus faibles. Il serait intéressant d'approfondir cet aspect.

Tableau 5.1
Corrélations[1] et pourcentage de croissance[2]

	Sous-unité						
	33 102	**83 112**	**95 135**	**51 201**	**80 114**	**53 120**	**51 110**
Corrélation	0,87	0,79	0,78	0,65	0,63	0,51	0,48
pourcentage de croissance	5,83	54,00	17,41	11,06	18,37	39,16	39,02

[1] corrélation entre les anciennes mesures de la taille et les nouvelles.
[2] pourcentage de croissance pour la période entre les recensements de 1966 et de 1971.

Tableau 5.2
Efficacité de la méthode 3 par rapport à la méthode 1

Caractéristique	Sous-unité						
	33 102	**83 112**	**95 135**	**51 201**	**86 114**	**53 120**	**51 110**
Population	87,8	27,4	25,3	30,0	48,1	23,8	8,6
Ménages	33,6	6,6	4,3	5,1	3,0	4,0	1,8
Personnes occupées	78,3	37,3	58,6	39,0	29,9	24,6	13,5
Personnes au chômage	82,1	85,4	86,4	99,3	78,3	79,3	88,3
Personnes non actives	87,2	57,7	43,1	50,7	89,4	55,4	31,7
Revenu	93,3	42,1	46,2	35,4	26,5	26,5	10,8

Le rendement des méthodes d'actualisation (4 et 5) et de la base de sondage spéciale par rapport à la méthode 1 peut être observé à partir d'une analyse des tableaux 5.3 et 5.4.

Du point de vue de l'efficacité (tableau 5.3) lorsqu'un tiers de la base de sondage a été réservé, il y a peu de différence entre l'actualisation au moyen des méthodes 4 et 5. Les gains d'efficacité selon les deux méthodes sont les plus faibles pour les caractéristiques des personnes au chômage et des personnes non actives (91 % à 93 %). Cette faible perte d'efficacité pour la méthode 4 est probablement attribuable à l'arrondissement des tailles en nombres entiers et à l'écart du plan autopondéré, car autrement, comme il est indiqué à la section (1), la variance selon les méthodes 1 et 4 devrait être identique. Il semble plausible d'attribuer la perte d'efficacité de la méthode 5 aux mêmes causes.

Tableau 5.3
Efficacité globale

Caractéristique	Méthode			
	1	**2**	**4**	**5**
Population	100	103,9	98,6	98,1
Ménages	100	107,8	102,0	100,7
Personnes occupées	100	101,1	101,5	100,4
Personnes au chômage	100	95,1	91,1	92,4
Personnes non actives	100	96,7	91,8	93,2
Revenu	100	103,2	101,4	99,9

Pour les autres caractéristiques, l'efficacité est de l'ordre de 98 % à 102 %. L'efficacité de la base de sondage spéciale diminue pour se fixer à 95 % pour les personnes au chômage et à 96,7 % pour les personnes non actives, mais pour d'autres caractéristiques, elle se situe entre 101 % et 108 %. Les gains d'efficacité ne semblent pas être influencés de façon appréciable par la procédure consistant à réserver une partie de la base de sondage, puis à extraire l'échantillon de la partie réservée plutôt que de l'extraire de la base de sondage entière. Ce phénomène semble être attribuable à la fois au plan dans la base de sondage spéciale et à l'estimateur de ratio proposé (4.2).

Tableau 5.4
Pourcentage d'écarts relatifs/pourcentage estimé du coefficient de variation

Caractéristique	Valeur de la population	Méthode			
		1	2	4	5
Population	49 389	0,17 0,1485	-0,12 0,1458	0,00 0,1497	0,11 0,1500
Ménages	14 264	0,07 0,0512	0,01 0,0493	0,01 0,0507	0,02 0,0510
Personnes non occupées	19 951	0,30 0,1731	-0,45 0,1719	-0,05 0,1721	0,08 0,1730
Personnes au chômage	1 615	0,35 0,7391	-0,22 0,7578	0,70 0,7739	0,22 0,7687
Personnes non actives	12 288	-0,10 0,2414	0,30 0,2454	0,52 0,2515	0,53 0,2495
Revenu (en milliers de dollars)	250 547	0,08 0,0972	-0,02 0,0957	-0,06 0,0965	-0,03 0,0972

Selon le tableau (5.4), le pourcentage d'écarts relatifs est faible dans tous les cas. La comparaison du pourcentage d'écarts relatifs pour les méthodes théoriquement impartiales (1 et 4) avec celles des autres méthodes donne à penser que le biais des méthodes 2 et 5 n'est pas important. Il convient de mentionner que, bien que des statistiques t significatives à un niveau de 95 % aient été obtenues pour la caractéristique employée selon la méthode 2 et non dans la population active pour les méthodes 4 et 5, ces biais semblent néanmoins sans signification pratique, ceux-ci étant inférieurs à 1 % de la valeur de la population. De plus, il convient de souligner que même si nous n'avons pas présenté d'écarts pour des sous-unités individuelles, ceux-ci ont été calculés, et nous avons observé qu'aucune méthode n'a sous-estimé ou surestimé une caractéristique pour toutes les sous-unités.

En conclusion, nous estimons que les tableaux 5.3 et 5.4 démontrent la pertinence globale de la stratégie que nous avons présentée, tant du point de vue de l'enquête permanente que des enquêtes spéciales. Nous supposons que dans des circonstances similaires à celles de l'étude, les deux systèmes de mise à jour fonctionneront de la même manière, de sorte que la méthode 5 devrait être privilégiée pour des raisons de simplicité des calculs.

Bibliographie

[1] Drew, J.O., "Sample Update - A Mapping Procedure for Cases where the number of Pre- and Post-Update Sample Are Not Equal", Note technique à l'interne de Statistique Canada, Household Surveys Development Division (1978).

[2] Fellegi, I.P., "Sampling with Varying Probabilities Without Replacement: Rotating and Non-Rotating Samples", *Journal of the American Statistical Association*, Vol. 58 (1963), pp. 183-201.

[3] Gray, G.B., "On Increasing the Sample Size (number of psu's)", Note technique à l'interne de Statistique Canada, Household Surveys Development Division (1973).

[4] Hartley, H.O. et Rao, J.N.K., "Sampling with Unequal Probabilities and Without Replacement", *Annals of Mathematical Statistics* (1962), Vol. 33, pp. 350-374.

[5] Keyfitz, N., "Sampling with Probabilities Proportional to Size: Adjustment for Changes in the Probabilities", *Journal of the American Statistical Association*, Vol. 46 (1951), pp. 105-109.

[6] Kish, L. et Scott, A., "Retaining Units after Changing Strata and Probabilities", *Journal of the American Statistical Association*, Vol. 66 (1971), pp. 461-470.

[7] Platek, R. et Singh, M.P., "A Strategy for Updating Continuous Surveys", *Techniques d'enquête* (Statistical Services, Statistique Canada), Vol. 1, No. 1 (juin 1975), pp. 16-26.

[8] Rao, J.N.K., Hartley, H.O. et Cochran, W.G., "On a Simple Procedure of Unequal Probability Sampling Without Replacement", *Journal of the Royal Statistical Society, Series B*, Vol. 27 (1962), pp. 482-491.

[9] Statistique Canada (Household Surveys Development Division), "Methodology of the Canadian Labour Force Survey (1976)", n° 71-526 au catalogue occasionel (publié en octobre 1977), pp. 33-38.

Annexe (A)

Aspects opérationnels de l'actualisation de l'échantillon au moyen de la procédure de Keyfitz

Prenons une strate ayant des unités N, avec des fractions de sondages inverses R_i; i = 1, 2, ..., N; définie selon (3.1), et avec le plan de remplacement décrit à la section 3 (page 8).

À un moment donné, les chiffres révisés des ménages sont obtenus, et les fractions de sondages inversés révisées R_i', i = 1, 2, ..., N, sont définis comme avant de sorte que $\sum_{i=1}^{N} R_i' = R$. Ensuite, les échantillons R uniques ordonnés selon les tailles révisées sont les suivants :

$$R_1'|R_1', \ (R_1'-1)|R_1', \ \ldots, \ R_N'|R_N', \ \ldots, \ 1|R_N' \ .$$

Par conséquent, au moment du renouvellement subséquent de l'échantillon, les probabilités de sélection des grappes doivent être ajustées de façon à ce qu'elles soient proportionnelles à la fraction de sondage inverse (FSI)révisée. Étant donné que nous avons le même nombre d'échantillons après l'actualisation que le nombre d'échantillons avant celle-ci, une simple mise en correspondance individuelle des échantillons avant l'actualisation avec les échantillons après l'actualisation peut être définie de manière à ce que :

i) les critères d'ajustement des probabilités de Keyfitz soient satisfaits;

ii) les échantillons postérieurs à l'actualisation puissent être limités à des logements qui n'ont pas été sélectionnés auparavant, pour lesquels, si la même grappe est conservée, une condition nécessaire, mais non suffisante est que :

$$x_i/R_i \geq x_i'/R_i', \qquad \text{(A.1)}$$

où $x_i|R_i$ est l'échantillon qui en aurait découlé s'il n'y avait pas eu d'actualisation et $x_i'|R_i'$ est l'échantillon postérieur à l'actualisation. Une autre condition concerne le choix de l'origine après l'actualisation et est abordée plus loin.

Une telle mise en correspondance (valeurs non exclusives) peut être effectuée comme suit :

a) Si $i \in D$, c'est-à-dire $R_i' < R_i$, alors les échantillons $R_i | R_i$, $(R_i - 1) | R_i$, … , $(R_i - R_i + 1) | R_i$ sont respectivement mis en correspondance dans les échantillons $R_i' | R_i'$, $(R_i' - 1) | R_i'$, … , $1 | R_i'$ et les échantillons $(R_i - R_i') | R_i$, $(R_i - R_i' - 1) | R_i$, … , $1 | R_i$ sont laissés temporairement sans correspondance.

b) Si $i \in l$, c'est-à-dire $R_i' \geq R_i$, alors les échantillons $R_i | R_i$, $(R_i - 1) | R_i$, … , $1 | R_i$ sont respectivement mis en correspondance dans les échantillons $R_i | R_i'$, $(R_i - 1) | R_i'$, … , $1 | R_i'$, en laissant les échantillons $R_i' | R_i'$, $(R_i' - 1) | R_i'$, … , $(R_i + 1) | R_i'$ disponibles.

c) Étant donné que $\sum_{i \in D} (R_i - R_i') = \sum_{i \in l} (R_i' - R_i) = f$, par exemple, les échantillons précédant l'actualisation et non mis en correspondance dans les grappes décroissantes peuvent être mis en correspondance de façon individuelle dans les échantillons disponibles après l'actualisation dans les grappes croissantes. Il y a f! mises en correspondance possibles. Idéalement, nous pourrions choisir cette mise en correspondance qui maximise l'intervalle de temps (c.-à-d. le nombre de périodes de renouvellement) avant que tout échantillon postérieur à l'actualisation revienne dans sa grappe correspondante d'avant l'actualisation et que l'on commence à réutiliser les logements. Cependant, l'évaluation de toutes les mises en correspondance f! ne sera pas toujours pratique, alors nous suggérons la procédure suivante :

Supposons que $D = \{i_1', i_2', \ldots, i_d'\}$ définit l'ensemble de grappes décroissantes ordonnées selon des numéros de série croissants et que $v = \{v_1, v_2, \ldots, v_d\}$ désigne les changements correspondants de leur nombre d'échantillons. Définissons $l = \{i_1'', i_2'', \ldots, i_e''\}$ et $w = \{w_1, w_2, \ldots, w_e\}$ de manière analogue pour l'ensemble des grappes croissantes.

Pour chaque $\ell = 1, 2, \ldots, d$, la procédure décrite ci-dessous détermine une mise en correspondance commençant par la grappe décroissante i_ℓ'. L'intervalle de temps minimal au cours duquel un échantillon postérieur à l'actualisation sera réintroduit dans

sa grappe d'avant l'actualisation correspondante et commencera à réutiliser les logements est également obtenu pour chaque mise en correspondance. Si a_ℓ est l'intervalle de temps minimal pour la mise en correspondance ℓ, alors la mise en correspondance ℓ^* pour laquelle $a^*_\ell = \max\{a_1, a_2, \ldots, a_d\}$ est choisie. Pour un ℓ donné, la mise en correspondance est définie comme suit :

Trouver la première grappe $k_1 \in 1$ avec $i''_{k_1} > i'_\ell$; c'est-à-dire la première grappe croissante qui sera remplacée dans l'échantillon après la grappe i'_ℓ. Il y a v_ℓ échantillons non mis en correspondance dans la grappe décroissante i'_ℓ; mettre en correspondance tous ces échantillons dans la grappe croissante $i''_{k_1}, i''_{(k_1 + 1) \bmod e}, \ldots$ en épuisant w_{k_1} échantillons disponibles dans la grappe croissante i''_{k_1} avant de procéder à $i''_{(k_1 + 1) \bmod e}$ et de façon semblable pour $i''_{(k_1 + 1) \bmod e}, i''_{(k_1 + 2) \bmod e}, \ldots$ utiliser autant de grappes croissantes que nécessaire. Après avoir mis en correspondance les v_ℓ échantillons de la grappe décroissante i'_ℓ dans les grappes croissantes $i''_{k_1}, i''_{(k_1 + 1) \bmod e}, \ldots$, les chiffres correspondants des échantillons disponibles, c'est-à-dire $w_{k_1}, w_{(k_1 + 1) \bmod e}, \ldots$, sont ajustés. Ensuite, prendre la grappe décroissante $i'_{(\ell + 1) \bmod d}$ et trouver la première grappe $k_2 \in 1$ avec $i''_{k_2} > i'_{(\ell + 1) \bmod d}$ et comme il est mentionné précédemment, mettre en correspondance l'ensemble des $v_{(\ell + 1) \bmod d}$ échantillons non mis en correspondance dans la grappe décroissante $i'_{(\ell + 1) \bmod d}$ dans les échantillons disponibles des grappes croissantes $i''_{k_2}, i_{(k_2 + 1) \bmod e}, \ldots$. Répéter ce processus pour les grappes $i'_{(\ell + 2) \bmod d}, i'_{(\ell + 3) \bmod d}, \ldots, i'_{(\ell + d - 1) \bmod d}$.

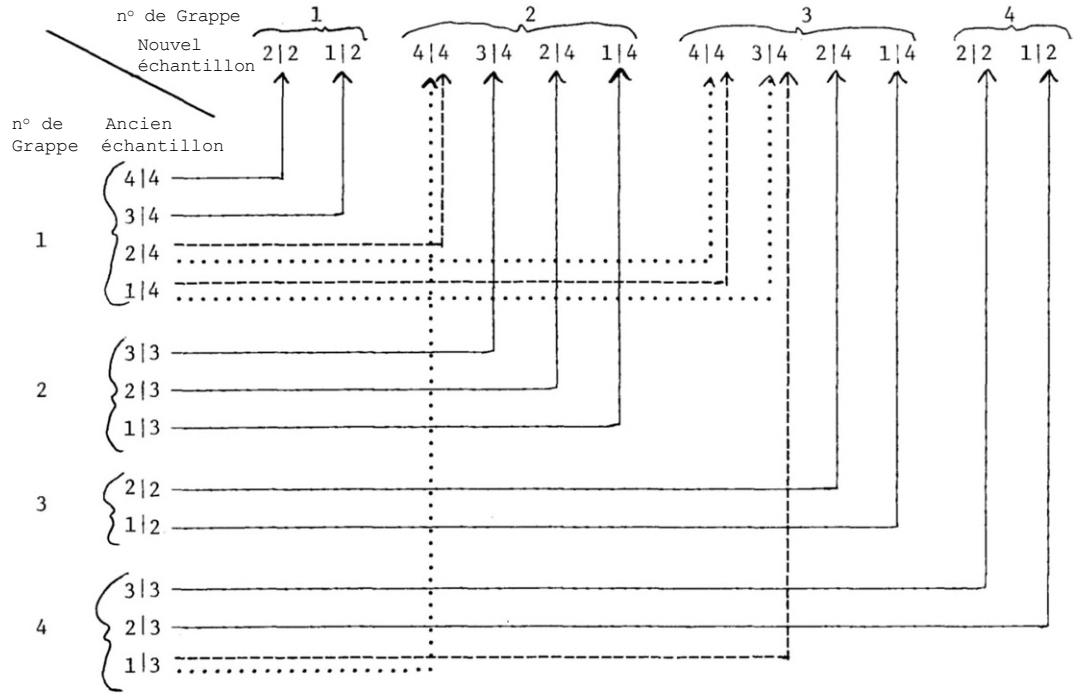
L'exemple suivant concernant le cas où nous avons quatre grappes avec les anciennes et la nouvelle FSI, tel que l'avons indiqué dans le tableau (A.1), illustre la procédure.

Tableau (A.1)

Grappe n°	Ancienne fraction de sondage inverse	Nouvelle fraction de sondage inverse
1	4	2
2	3	4
3	2	4
4	3	2
	12	12

L'ensemble des grappes décroissantes D = {1, 4} et les changements correspondants dans les FSI, c'est-à-dire V = {-2, -1}, et de même pour l'ensemble des grappes croissantes 1 = {2, 3}, W = {1, 2}. La figure (1) ci-dessous montre la mise en correspondance des échantillons antérieurs à l'actualisation dans les échantillons postérieurs à l'actualisation.

Mise en correspondance des échantillons antérieurs à l'actualisation dans les échantillons postérieurs à l'actualisation



Les lignes pleines correspondent aux échantillons avant l'actualisation mis en correspondance avec les échantillons après l'actualisation dans la même grappe, c'est-à-dire les cas où l'ancienne grappe sélectionnée est conservée. Les échantillons non mis en correspondance avant l'actualisation dans les grappes décroissantes peuvent être mis en correspondance avec les échantillons disponibles après l'actualisation dans les grappes croissantes à partir de la grappe décroissante 1 (lignes

brisées) ou à partir de la grappe décroissante 4 (lignes pointillées). L'intervalle de temps minimal pour la nouvelle sélection de logements est de 3 pour la mise en correspondance représentée par les lignes brisées et de 5 pour celle qui est représentée par les lignes pointillées. Dans la première mise en correspondance (lignes brisées), l'intervalle de temps minimal correspond à la mise en correspondance de l'échantillon antérieur à l'actualisation $1|3$ de la grappe 4 dans l'échantillon postérieur à l'actualisation $3|4$ de la grappe 3. Dans ce cas, après l'utilisation des échantillons $3|4$, $2|4$ et $1|4$ de la grappe 3, les logements de la grappe antérieure à l'actualisation, 4, seraient réutilisés avec l'échantillon $2|2$. Dans la deuxième mise en correspondance (lignes pointillées), l'intervalle de temps correspond à la mise en correspondance de l'échantillon antérieur à l'actualisation $1|4$ de la grappe 1 dans l'échantillon postérieur à l'actualisation $3|4$ de la grappe 3. Par conséquent, la deuxième mise en correspondance (lignes pointillées) sera utilisée.

Dans le schéma de mise en correspondance ci-dessus, il apparaît clairement que :

i) Les grappes sont sélectionnées selon une probabilité proportionnelle à leur FSI révisée, au besoin.
ii) La probabilité est la même pour chaque échantillon postérieur à l'actualisation, de sorte que, selon le mode de renouvellement, ces probabilités seront préservées.
iii) Les conditions de Keyfitz sur le rejet et la conservation des grappes se vérifient.
iv) La condition nécessaire pour éviter une nouvelle sélection de logements se vérifie également.

Maintenant que nous avons déterminé l'échantillon postérieur à la mise à jour dans le processus de mise en correspondant précédent, il reste à déterminer les origines aléatoires après l'actualisation. Les trois éventualités suivantes se présentent :

i) Au moment de l'actualisation, l'ancienne grappe est rejetée et une nouvelle grappe i est sélectionnée. Ensuite, une origine aléatoire r_i', $1 \leq r_i' \leq R_i'$ est choisie, et si l'échantillon à

introduire est $j \mid R_i'$, alors les échantillons systématiques déterminés par les origines r_i', $(r_i' + 1) \bmod R_i'$, …, $(r_i' + j - 1) \bmod R_i'$ sont associés aux échantillons $j \mid R_i'$, $(j-1) \mid R_i'$, …, $1 \mid R_i'$, respectivement.

ii) La grappe i sélectionnée précédemment est conservée et $R_i' = R_i$. Dans ce cas, la séquence de renouvellement dans la grappe i reste inchangée.

iii) La grappe sélectionnée précédemment est conservée et $R_i' \neq R_i$. Dans ce cas, une mise en correspondance des anciennes origines dans les nouvelles origines est requise, de sorte que la probabilité globale pour chaque nouvelle origine soit égale à $1/R_i'$ et que le nombre de logements à utiliser dans les origines postérieures à l'actualisation ne dépasse jamais le nombre de logements utilisés avant l'actualisation. La première condition permet d'assurer une sélection impartiale des origines, tandis que la deuxième condition nous permet de réorganiser les logements, comme il est décrit plus loin, de sorte à empêcher de nouvelles sélections de logements.

Soit $\Pr(s \rightarrow s')$ la probabilité que l'origine antérieure à l'actualisation $s (s = 1, 2, \ldots, R_i)$ soit mise en correspondance avec l'origine postérieure à l'actualisation $s' (s' = 1, 2, \ldots, R_i')$. Nous devons alors déterminer une matrice P $R_i \times R_i'$, de sorte que $\Pr(s \rightarrow s')$ soit donné par $P_{ss'}$, où :

$$\sum_{s'=1}^{R_i'} P_{ss'} = 1 \quad \text{pour tous les } s$$

$$\sum_{s=1}^{R_i} \frac{1}{R_i} P_{ss'} = \frac{1}{R_i'} \quad \text{pour tous les } s',$$

et que la condition nécessaire pour empêcher la nouvelle sélection de logements se vérifie également. Pour ce faire, on peut déterminer une matrice A $R_i \times R_i'$, de sorte que :

$$\sum_{s'=1}^{R_i'} a_{ss'} = R_i' \quad \text{pour tous les } s \qquad \text{(A.2)}$$

$$\sum_{s=1}^{R_i} a_{ss'} = R_i \quad \text{pour tous les } s', \tag{A.3}$$

et attribuer les valeurs maximales possibles aux éléments de la matrice A dans l'ordre $a_{11}, a_{12}, \ldots, a_{1R_i'}, a_{21}, \ldots, a_{R_i 1}, a_{R_i 2}, \ldots, a_{R_i R_i'}$ sous réserve des contraintes (A.2) et (A.3). Alors $\Pr(s \to s')$ est simplement donné par $a_{ss'} / R_i'$ c'est-à-dire que la matrice P sera définie comme suit :

$$P = \frac{1}{R_i'} A \tag{A.4}$$

Les probabilités $P_{ss'}(s = 1, 2, \ldots, R_i, s' = 1, 2, \ldots, R_i')$ définies par (A.4) mettront toujours en correspondance l'ancienne origine avec la plus grande probabilité admissible dans la plus petite nouvelle origine à chaque étape en commençant par l'ancienne origine 1, puis l'ancienne origine 2, et ainsi de suite jusqu'à l'ancienne origine R_i.

La matrice A qui définit la mise en correspondance du cas $R_i = 6$ et $R_i' = 7$ est donnée dans le tableau (A.2).

Tableau (A.2)
Matrice A pour obtenir la probabilité de l'origine après l'actualisation compte tenu de l'origine avant l'actualisation

Origine avant l'actualisation	Origine après l'actualisation						
	1	**2**	**3**	**4**	**5**	**6**	**7**
1	6	1	0	0	0	0	0
2	0	5	2	0	0	0	0
3	0	0	4	3	0	0	0
4	0	0	0	3	4	0	0
5	0	0	0	0	2	5	0
6	0	0	0	0	0	1	6

A partir du tableau précédent, nous trouvons $\Pr(1 \to 1) = \frac{6}{7}$, $\Pr(1 \to 2) = \frac{1}{7}$ etc. Nous pouvons facilement vérifier que si la mise en correspondance du cas $R_i = 6$, $R_i' = 7$ est donné par la matrice A ci-dessus, alors la

mise en correspondance du cas $R_i = 7$ et $R_i' = 6$ sera donné par A^T où A^T est la transposition de la matrice A, et cela est vrai en général.

Il est facile de vérifier que la mise en correspondance des origines avant l'actualisation avec les origines après l'actualisation, combinée avec la mise en correspondance antérieure des origines après l'actualisation, garantit que le nombre de logements à utiliser après l'actualisation dans les grappes conservées est inférieur ou égal au nombre de logements inutilisés avant l'actualisation. Il suffit essentiellement d'ajuster l'ordre des logements de façon à ce que les logements déjà sélectionnés apparaissent tous dans les origines postérieures à l'actualisation qui ne seront pas utilisées.

Avant d'envisager l'ajustement de l'ordre, il convient de mentionner que dans tous les cas pour les futures grappes introduites dans l'échantillon après l'actualisation, une origine r_i', $1 \leq r_i' \leq R_i'$ est choisie au hasard et un calendrier de renouvellement comprenant une séquence d'échantillons systématiques est déterminé de la même façon qu'avant l'actualisation.

Ajustement de l'ordre des logements

La FSI de la grappe, R_i, et le nombre de logements N_{it} dans la grappe i au temps t déterminent le nombre de logements qui seront sélectionnés dans chaque origine dans la grappe. Si $b_{it} = \left[\frac{N_{it}}{R_i}\right]$ et $Q_{it} = N_{it} - R_i \,.\, b_{it}$, alors les Q_{it} premières origines ont $b_{it} + 1$ logements et les autres ont b_{it} logements. Un modèle ou une matrice incomplète sont définis par N_{it} et R_i, comme il est illustré à la page suivante, pour le cas $N_{it} = 16$, $R_i = 6$.

Figure (2)

Origines	1	2	3	4	5	6
Logements	X	X	X	X	X	X
	X	X	X	X	X	X
	X	X	X	X		

Normalement, les logements sont inscrits par rangée dans ce modèle, à savoir :

Figure (3)

Origines	**1**	**2**	**3**	**4**	**5**	**6**
Logements	1	2	3	4	5	6
	7	8	9	10	11	12
	13	14	15	16		

de sorte que les logements 1, 7 et 13 soient sélectionnés avec l'origine 1, etc. Les nouveaux logements sont ajoutés par rangée, ce qui augmente la taille de la matrice. Si la FSI devient R_i' lors de l'actualisation avec une origine après l'actualisation de r_i', alors le nouvel ordre fonctionnerait comme suit.

Les logements dans les origines inutilisées sont énumérés par colonne de gauche à droite à partir du modèle ci-dessus; supposons qu'il y a L_i de ces logements. Un nombre aléatoire ℓ_i; $1 \leq \ell_i \leq L_i$ est déterminé. Ensuite, dans l'ordre ℓ_i, $(\ell_i + 1) \bmod L_i, \ldots, (\ell_i + L_i - 1) \bmod L_i$, les logements inutilisés sont inscrits par colonne dans le modèle sous la nouvelle origine de la FSI en commençant par la colonne r_i' et en revenant à la première colonne du modèle une fois la dernière colonne atteinte. Si nous prenons les autres origines dans l'ordre dans lequel elles ont été utilisées, les logements sont inscrits de la même façon à partir de la position suivant le dernier logement inutilisé.

Pour illustrer cela, supposons qu'à $t = 1$, la grappe i avec $R_i = 6$, $r_i = 1$ a été sélectionné avec l'échantillon $6|6$, et que $N_{i1} = 16$. À $t = 4$, l'échantillon est actualisé, de sorte que $r_i^* = 4$, où r_i^* est l'origine qui en aurait résulté s'il n'y avait pas eu d'actualisation. Supposons que nous avons $R_i' = 7$, alors les mises en correspondance requises précisent respectivement que : i) l'échantillon après l'actualisation devrait être $3|7$; et ii) l'origine après l'actualisation devrait être $r_i' = 4$ avec une probabilité de 4/7 et $r_i' = 5$ avec une probabilité de 3/7. Disons que nous avons $r_i' = 4$. D'après la figure (3), les logements sous dans anciennes origines non utilisées (c.-à-d. les origines 4, 5 et 6) sont {4, 10, 16, 5, 11, 6, 12}. Supposons $\ell_i = 3$, nous obtiendrons alors l'ajustement de l'ordre suivant.

Figure (5)

Nouvelles origines	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Logements	7	8	9	16	11	12	10
	13	14	15	5	6	4	1
	2	3					

Après avoir utilisé les origines 4, 5 et 6, le renouvellement s'effectue dans la grappe subséquente.

Il convient de mentionner que si r_i' avait été choisi comme nombre entier aléatoire entre 1 et R_i', nous aurions pu avoir $r_i' = 1$, auquel cas, selon l'origine après l'actualisation 1, 2, 3, un total de 8 logements doivent être sélectionnés alors que $L_i = 7$; autrement dit, une nouvelle sélection de logements aurait eu lieu.

Il peut être démontré dans l'exemple ci-dessus que la procédure d'ajustement de l'ordre est légèrement biaisée pour la sélection au niveau du logement. Compte tenu de l'échantillon avant l'actualisation 3|6, les origines inutilisées peuvent être {1, 2, 3}, {2, 3, 4}, {3, 4, 5}, {4, 5, 6}, {5, 6, 1}, ou {6, 1, 2}, avec une probabilité égale de savoir où r_i^* correspond à la première origine dans chaque cas. Pour $N_{i1} = N_{i4} = 16$, les logements sous chacune de ces origines sont tous déterminés. La mise en correspondance des origines lors de l'actualisation prend : $r_i^* = 1$ à $r_i' = 1$ avec une probabilité de 6/7 et à $r_i' = 2$ avec une probabilité de 1/7, après quoi, dans chaque cas, 3 logements sur les 9 logements sous une origine avant l'actualisation {1, 2, 3} seront sélectionnés avec une probabilité égale; $r_i^* = 2$ à $r_i' = 2$ avec une probabilité de 5/7, après quoi, 3 logements sur 9 sont sélectionnés avec une probabilité égale; et $r_i^* = 2$ à $r_i' = 3$ avec une probabilité de 2/7, après quoi, 2 des 9 logements sont sélectionnés avec une probabilité égale, etc. Les probabilités globales au temps t = 4 sont {0,14484, 0,14749, 0,14749, 0,13955, 0,13690, 0,13690} pour les logements sous les origines avant l'actualisation {1, 2, ..., 6} respectivement; tandis que sous la nouvelle FSI de 7, les probabilités après l'actualisation des logements devraient chacune être égales à $1/7 \doteq 0{,}14286$. Étant donné le choix entre les risques inhérents au fardeau du répondant découlant des nouvelles sélections de logements et le léger biais de sélection au niveau du logement en raison de l'ajustement de l'ordre, cette dernière procédure a été jugée préférable.