Techniques d'enquête, décembre 1996
Vol. 22, n° 2, pp. 117-126
Statistique Canada

Estimation de la variance des estimateurs de calage: comparaison des méthodes du jackknife et de la linéarisation de Taylor

DIANA M. STUKEL, MICHAEL A. HIDIROGLOU et CARL-ERIK SÄRNDAL[1]

RÉSUMÉ

L'utilisation de données auxiliaires dans les méthodes d'estimation des enquêtes complexes, notamment l'Enquête sur la population active de Statistique Canada, ne cesse de se perfectionner. L'estimation par régression et l'estimation par itération du quotient étaient naguère les méthodes les plus courantes pour intégrer les données auxiliaires à l'estimation. Il arrivait toutefois que les poids associés à l'estimateur soient négatifs ou hautement positifs. Les progrès théoriques réalisés récemment par Deville et Särndal (1992) en vue de la construction de poids «restrictifs» que l'on peut assujettir à une valeur positive et à un plafond nous ont incités à étudier les propriétés des estimateurs en résultant. Nous examinons ici les propriétés de diverses méthodes servant à engendrer des poids de ce genre et la variance estimative correspondante. Nous nous intéresserons en particulier à deux méthodes d'estimation de la variance en recourant à une simulation de Monte Carlo articulée sur les données de l'Enquête sur la population active. Il s'agit en l'occurrence des méthodes du jackknife et de la linéarisation de Taylor. On en conclut que les estimateurs ponctuels et les estimateurs de la variance n'entraînent qu'un biais minime, même avec l'application de sérieuses «restrictions» aux poids finals.

MOTS CLÉS: Données auxiliaires; estimateurs par itération du quotient; estimateurs de régression; pondération restrictive.

1. INTRODUCTION

Les données auxiliaires comptent de nombreuses applications dans les enquêtes par échantillonnage. Un exemple typique est leur incorporation à l'estimation sous forme d'estimateurs de régression ou d'estimateurs par itération du quotient. Pour obtenir ces estimateurs, on multiplie le poids d'échantillonnage d'une unité par un facteur de correction, ce qui donne le poids final. Un inconvénient bien connu de l'estimateur de régession est que certains facteurs de correction peuvent être négatifs, si bien que le poids final le devient lui aussi. Par ailleurs, quelques facteurs de correction de l'estimateur par itération du quotient peuvent être très élevés et positifs, ce qui débouche sur des poids finals trop importants. Pour surmonter ces inconvénients, on peut recourir à une autre variété d'estimateurs, les «estimateurs de calage». Créés par Deville et Särndal (1992), ces estimateurs intègrent les données auxiliaires. Dans certains cas, on peut veiller à ce que les poids n'aient pas de valeur négative en spécifiant d'avance leurs limites inférieures et supérieures. On parvient aux poids «calés» en réduisant les fonctions qui mesurent l'écart entre les poids d'échantillonnage originaux et les poids finals calés, sous réserve de certaines contraintes de calage. Huang et Fuller (1978) ainsi que Singh et Mohl (1996) ont mis au point des estimateurs analogues qui conservent les propriétés précitées. Habituellement, les différences entre les estimations ponctuelles correspondant aux différentes fonctions de distance sont très faibles.

Depuis son lancement, l'Enquête sur la population active (EPA) de Statistique Canada a utilisé, à un moment ou à un autre la technique d'estimation de la variance de Taylor aussi bien que celle du jackknife, parallèlement à l'emploi d'estimateurs de régression et d'itération du quotient. Outre l'estimateur de régression en usage jusqu'à présent, l'EPA a récemment permis l'application d'autres estimateurs de calage en vue d'éliminer le problème d'éventuels poids négatifs. Il vaut donc la peine d'examiner le comportement de ces estimateurs ponctuels et des estimateurs de la variance de Taylor et jackknife correspondants, surtout quand ils tolèrent l'imposition de limites aux poids. Tel est le sujet principal de ce document. Disons d'emblée que les méthodes de Taylor et du jackknife présentent des avantages propres. La première exige considérablement moins de calculs que la seconde, mais il faut créer de nouvelles expressions pour chaque paramètre envisagé, fardeau paticulièrement lourd dans le cas des enquêtes à fins multiples dans lesquelles on est susceptible de s'intéresser à un grand nombre de paramètres. De son côté, la méthode du jackknife, n'exige pas l'établissement d'expressions encombrantes de la variance pour chaque nouveau paramètre; la fonction de l'estimateur ponctuel suffit.

Dans la partie 2, on expose les principes théoriques de l'estimation par calage et introduit une série de fonctions de distance connexes; la partie 3 traite de la variance des estimateurs de calage; on trouvera à la partie 4 les résultats d'une étude de simulation de Monte Carlo qui établit le biais des estimateurs ponctuels et des estimateurs de la variance de Taylor et jackknife correspondants (par rapport à une variance «réelle») pour diverses fonctions de distance issues de la théorie de calage. La partie 5 sert de conclusion.

[1] Diana M. Stukel, Division des méthodes d'enquêtes des ménages, et Michael A. Hidiroglou, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6; Carl-Erik Särndal, Département de Mathématiques et de Statistique, Université de Montréal, C.P. 6128, Succursale A, Montréal, P.Q., H3C 3J7.

2. FONCTIONS DE DISTANCE ET ESTIMATEURS DE CALAGE

Commençons par introduire l'idée fondamentale à la base de l'estimation de calage. Soit $U = \{1, ..., k, ..., N\}$, l'ensemble de N unités d'une population finie d'unités. Dans les enquêtes par échantillonnage, on désire souvent estimer les paramètres d'une population finie, par exemple les totalisations, les moyennes et les ratios. Pour plus de simplicité, nous ne nous intéresserons qu'aux totalisations, mais les idées exposées ici peuvent aisément s'étendre à d'autres paramètres. Supposons donc que l'objectif consiste à estimer le total de la population $Y = \sum_{k\in U} y_k$, où y_k représente la valeur de y, la variable à laquelle on s'intéresse pour la k-ième unité de population.

On prélève au hasard s unités de U selon un plan d'échantillonnage établi induisant les probabilités d'inclusion $\pi_k = P(k\in s)$. On suppose que ces dernières sont connues et positives. Soit $a_k = 1/\pi_k$, le poids d'échantillonnage associé à la k-ième unité. Enfin, supposons que les données auxiliaires prennent la forme des chiffres de population connus, pour une ou plusieurs variables auxiliaires.

Un estimateur élémentaire de Y est l'estimateur de Horvitz-Thompson (HT)

$$\hat{Y}_a = \sum_{k\in s} a_k y_k.$$

L'estimateur HT ne peut intégrer les données auxiliaires qu'au stade du plan d'échantillonnage, mais il ne le fait pas nécessairement (cela dépend); on préférerait un autre estimateur en mesure d'intégrer les données auxiliaires au stade de l'estimation également. L'incorporation des données auxiliaires peut créer de nouveaux poids, notés w_k; $k\in s$. Par conséquent, le nouvel estimateur s'écrit:

$$\hat{Y}_w = \sum_{k\in s} w_k y_k. \tag{2.1}$$

L'approche retenue par Deville et Särndal (1992) ainsi que par Deville, Särndal et Sautory (1993) comporte la détermination des nouveaux poids $\{w_k: k\in s\}$ par leur rapprochement au maximum des poids d'échantillonnage originaux $\{a_k; k\in s\}$, au moyen d'une fonction de distance spécifiée. Les restrictions imposées aux nouveaux poids font en sorte qu'une fois appliqués à chacune des variables auxiliaires, ces derniers reproduiront le chiffre de population $\boldsymbol{X}$ connu . À savoir,

$$\sum_{k\in s} w_k \boldsymbol{x}_k = \boldsymbol{X} \tag{2.2}$$

est valable, ce qui entraîne un problème de minimisation sous contrainte. Ici, $\boldsymbol{x}_k' = (x_{1k}, x_{2k}, ..., x_{pk})$ représente un vecteur de longueur p incluant les valeurs des variables auxiliaires pour le k-ième individu, et les données auxiliaires d'une source extérieure sont résumées par le total vectoriel $\boldsymbol{X} = \sum_{k\in U} \boldsymbol{x}_k$. qui est connu.

$F^*(w_k, a_k)$ exprime la distance entre w_k et a_k. Deville et Särndal (1992) se bornent à discuter des fonctions de distance du genre $F^*(w_k, a_k) = a_k c_k F(w_k/a_k)$ où $w_k/a_k = g_k$, quotient du poids calé final sur le poids d'échantillonnage original, est appelé le «coefficient g». Ici c_k représente un poids positif connu, sans lien avec a_k; dans la pratique, on recourt souvent à la pondération $c_k = 1$. Notons que l'équation (2.1) peut aussi s'écrire de la façon suivante:

$$\hat{Y}_w = \sum_{k\in s} a_k g_k y_k.$$

On présume que F n'est pas négative et a une forme convexe, et que $F(1) = 0$, si bien que lorsque $w_k = a_k$, la distance entre les poids est nulle. D'autre part, F' doit être continue, biunivoque et en outre que, $F'(1) = 0$ et $F''(1) > 0$, de sorte que $w_k = a_k$ représente un minimum local. (Lire Deville, Särndal et Sautory 1993). La distance totale $\sum_{k\in s} a_k c_k F(w_k/a_k)$ est minimisée sous réserve de la contrainte (2.2). En d'autres termes, l'expression

$$\sum_{k\in s} a_k c_k F(w_k/a_k) - \lambda'\left(\sum_{k\in s} w_k \boldsymbol{x}_k - \boldsymbol{X}\right)$$

est minimisée par rapport à w_k, où λ est un vecteur p de multiplicateurs de Lagrange. En effectuant la différenciation par rapport à w_k, , en rendant l'équation égale à zéro et en résolvant pour w_k, on obtient les poids calés $w_k = a_k g_k = a_k g(\lambda'\boldsymbol{x}_k/c_k)$, où g représente la fonction inverse de f et $f(z) = dF(z)/dz$. Pour calculer w_k, il faut d'abord résoudre l'équation de calage qu'implique (2.2) afin d'obtenir λ, c'est-à-dire

$$\sum_{k\in s} a_k g(\lambda'\boldsymbol{x}_k/c_k)\boldsymbol{x}_k = \boldsymbol{X}. \tag{2.3}$$

Cette solution d'un (éventuel) système non linéaire de p équations à p inconnues pourrait nécessiter le recours à une méthode d'itération quelconque, celle de Newton-Raphson, par exemple.

Deville et Särndal (1992), Huang et Fuller (1978) ainsi que Singh et Mohl (1996) examinent plusieurs fonctions de distance. Les deux principales auxquelles nous nous attarderons sont la fonction de distance des moindres carrés généralisés (MCG) et la fonction de distance par itération du quotient (IQ) mentionnées par Deville et Särndal(1992).

La fonction de distance MCG se définit comme suit:

$$F^*(w_k,a_k) = F^*_{\mathrm{GLS}}(w_k,a_k)$$

$$= c_k(w_k - a_k)^2/a_k = a_k c_k (w_k/a_k - 1)^2. \tag{2.4}$$

Elle permet d'obtenir l'estimateur de régression généralisé (ERG) bien connu dont l'estimateur par ratio, l'estimateur de régression simple et l'estimateur de stratification a posteriori simple sont des cas particuliers. De (2.3), il s'ensuit que les poids calés représentant la fonction de distance MCG sont:

$$w_k = a_k g_k = a_k[1 + (\boldsymbol{X} - \hat{\boldsymbol{X}}_a)'\left(\sum_{j\in s} a_j \boldsymbol{x}_j \boldsymbol{x}_j'/c_j\right)^{-1}\boldsymbol{x}_k/c_k]$$

où $\hat{\boldsymbol{X}}_a = \sum_{k\in s} a_k \boldsymbol{x}_k$ représente l'estimateur HT de $\boldsymbol{X}$. On peut écrire l'estimateur de Y correspondant sous la forme habituelle de l'estimateur de régression

$$\hat{Y}_{w(\mathrm{GREG})} = \hat{Y}_a + (\boldsymbol{X} - \hat{\boldsymbol{X}}_a)'\hat{\boldsymbol{\beta}} \tag{2.5}$$

où

$$\hat{\boldsymbol{\beta}} = \left(\sum_{k\in s} a_k \boldsymbol{x}_k \boldsymbol{x}_k'/c_k\right)^{-1} \sum_{k\in s} a_k \boldsymbol{x}_k y_k/c_k. \tag{2.6}$$

L'estimateur de régression correspond donc à l'estimateur HT auquel s'ajoute un terme de correction. L'inconvénient de la fonction de distance MCG est qu'elle peut donner des poids négatifs, surtout si le système fait l'objet de contraintes trop nombreuses. Dans la pratique, les poids négatifs sont rares; cependant, il est préférable de les éliminer totalement car les interpréter pourrait s'avérer difficile.

La fonction de distance par itération du quotient (IQ) se définit comme suit:

$$\begin{aligned} F^*(w_k, a_k) &= F^*_{RR}(w_k, a_k) \\ &= c_k[w_k \log(w_k/a_k) - w_k + a_k] \\ &= a_k c_k[(w_k/a_k)\log(w_k/a_k) - (w_k/a_k) + 1]. \end{aligned} \tag{2.7}$$

On peut montrer que calculer les coefficients g au moyen de la fonction de distance IQ et des restrictions définies en (2.3) équivaut à utiliser l'algorithme d'ajustement proportionnel itératif (API) de Deming et Stephan (1940) afin de caler les valeurs marginales connues des tableaux de fréquence à deux dimensions ou plus. À l'inverse de la fonction de distance MCG, dont la solution est fermée, les équations de calage de la fonction de distance IQ ne peuvent être résolues que par itération. Divers logiciels le permettent; le logiciel CALMAR (lire Deville, Särndal et Sautory 1993), notamment, résoud les équations de calage pour la fonction de distance IQ par la méthode de Newton-Raphson plutôt qu'au moyen de l'algorithme API que proposent au départ Deming et Stephan. La fonction de distance IQ donne toujours des poids positifs; néanmoins, elle a pour défaut d'aboutir à des poids calés excessifs dans certains cas.

Des poids négatifs comme ceux de la fonction de distance MCG et des poids positifs importants comme ceux de la fonction de distance IQ ne nous intéressent pas. On peut cependant définir une fonction de distance restrictive en vertu de laquelle les poids résultants w_k seront bornés. Il suffit d'imposer des restrictions à la fonction de distance $F(w_k/a_k)$, de sorte que les coefficients g, $g_k = w_k/a_k$ se retrouveront dans une fourchette préétablie. Pour cela, on fixe une limite inférieure L et une limite supérieure U en vertu desquelles $L < 1 < U$. Pour obtenir des poids positifs, on prend $L > 0$. Deville et Särndal (1992) présentent des variantes restrictives des deux fonctions de distance précitées, à savoir la fonction de distance MCG restrictive (MCGR) et la fonction de distance IQ restrictive (IQR) ou à logit. Huang et Fuller (1978) et Singh and Mohl (1996) proposent deux autres méthodes qui limitent les poids finals. Ces quatre fonctions de distance restrictives sont examinées plus loin. Singh et Mohl (1996) les analysent aussi en détail, mais sous un angle différent.

La fonction de distance MCG restrictive se définit ainsi:

$$F^*(w_k, a_k) = F^*_{\mathrm{RGLS}}(w_k, a_k) = \begin{cases} c_k(w_k - a_k)^2/a_k & \text{si} \quad L < w_k/a_k < U \\ \infty \quad \text{autrement.} \end{cases} \tag{2.8}$$

La fonction de distance IQ restrictive (ou à logit) s'écrit

$$F^*(w_k, a_k) = F^*_{\mathrm{RRR}}(w_k, a_k) =$$

$$\begin{cases} A^{-1} c_k[(w_k/a_k - L)\log[(w_k/a_k - L)/(1 - L)] \\ \quad + (U - w_k/a_k)\log[(U - w_k/a_k)/(U - 1)]] \\ \qquad\qquad\qquad \text{si} \quad L < w_k/a_k < U \\ \infty \quad \text{autrement} \end{cases} \tag{2.9}$$

où $A = (U - L)/\{(1 - L)(U - 1)\}$. La spécification $L = 0$, $U = \infty$ donne la fonction de distance IQ. On voit aisément que ces deux fonctions de distance partagent la propriété que les poids w_k correspondants satisfont à $L < w_k/a_k < U$.

Huang et Fuller (1978) suggèrent une méthode pour corriger les poids de régression afin que les contraintes de calage de l'équation (2.2) soient satisfaites et que les coefficients g restent voisins de un. Singh et Mohl (1996) montrent que leur méthode correspond à la minimisation d'une fonction de distance qui change d'une itération à l'autre. Ces deux auteurs modifient aussi la méthode originale afin de permettre l'application de restrictions arbitraires aux coefficients g, restrictions semblables à celles des fonctions de distance qui précèdent, et montrent que l'estimateur résultant est asymptotiquement équivalent à l'estimateur de régression. La fonction de distance modifiée de Singh et Mohl (HFM) s'écrit ainsi:

$$\begin{aligned} F^*(w_k^{(v-1)}, a_k) &= F^{(v)*}_{\mathrm{HFM}}(w_k^{(v-1)}, a_k) \\ &= (w_k^{(v-1)} - a_k)^2/a_k q_k^{(v-1)*}; \quad v = 1, 2, \ldots \end{aligned} \tag{2.10}$$

où $q_k^{(v-1)*} = q_k^{(v-1)} \ldots q_k^{(1)} q_k^{(0)}, q_k^{(0)} = 1$, et où v est le nombre d'itérations. Dans ce cas,

$$q_k^{(v-1)} = \begin{cases} 1 & \text{si} \quad \xi_k^{(v-1)} < .5 \\ 1 - \delta(\xi_k^{(v-1)} - .5)^2 & \text{si} \quad .5 \le \xi_k^{(v-1)} < 1 \\ (1 - \delta/4)/\xi_k^{(v-1)} & \text{si} \quad \xi_k^{(v-1)} \ge 1 \end{cases}$$

pour δ, choisi arbitrairement afin que $0 < \delta < 1$. Par ailleurs,

$$\xi_k^{(v-1)} = \begin{cases} (g_k^{(v-1)} - 1)/(L' - 1) & \text{si} \quad g_k^{(v-1)} \le 1 \\ (g_k^{(v-1)} - 1)/(U' - 1) & \text{autrement} \end{cases}$$

où $L' = \alpha L + 1 - \alpha$ et $U' = \alpha U + 1 - \alpha$ pour α, choisi arbitrairement afin que $0 < \alpha < 1$, et où L et U ont une valeur identique à celles des fonctions de distance restrictives qui précèdent. Les paramètres α et δ accélèrent la convergence de

l'algorithme d'itération utilisé pour parvenir à une solution. Singh et Mohl (1996) ont testé de façon empirique diverses valeurs pour ces paramètres à partir de vastes ensembles de données. Selon eux, les valeurs $\alpha = .67$ et $\delta = .8$ donnent de bons résultats dans la pratique. Enfin, le coefficient g de chaque itération est

$$g_k^{(\nu-1)} = 1 + (\boldsymbol{X} - \hat{\boldsymbol{X}}_w^{(\nu-2)})' \left(\sum_{j \in s} a_j q_j^{(\nu-2)*} \boldsymbol{x}_j \boldsymbol{x}_j' \right)^{-1} \boldsymbol{x}_k; \quad \nu = 2, 3, \ldots$$

où $\hat{\boldsymbol{X}}_w^{(\nu-2)} = \sum_{k \in s} w_k^{(\nu-2)} \boldsymbol{x}_k$; $\nu = 2, 3, \ldots$; où $w_k^{(\nu-2)} = a_k g_k^{(\nu-2)}$; $\nu = 2, 3, \ldots$; et où les valeurs de départ sont données par $g_k^{(0)} = 1$ et $w_k^{(0)} = a_k$.

Ces deux auteurs proposent aussi une nouvelle fonction de distance qui change d'itération en itération, baptisée fonction de distance à rétrécissement-minimisation (RM). Ils montrent que l'estimateur issu de cette fonction est lui aussi asymptotiquement équivalent à l'estimateur de régression. On l'exprime par:

$$F^*(w_k^{(\nu-1)}, a_k) = F_{SM}^{(\nu)*}(w_k^{(\nu-1)}, a_k) = (w_k^{(\nu-1)} - a_k^{(\nu-1)*})^2 / a_k^{(\nu-1)*}; \qquad \nu = 1, 2, \ldots \quad (2.11)$$

où

$$a_k^{(\nu-1)*} = \begin{cases} L' a_k & \text{si } w_k^{(\nu-1)} < L'' a_k \\ U' a_k & \text{si } w_k^{(\nu-1)} > U'' a_k \\ w_k^{(\nu-1)} & \text{autrement.} \end{cases} \qquad \nu = 2, 3, \ldots$$

Les termes des équations qui précèdent se définissent comme suit: $L' = \alpha L + (1 - \alpha)$, $U' = \alpha U + (1 - \alpha)$, $L'' = \eta L + (1 - \eta)$ et $U'' = \eta U + (1 - \eta)$ pour α et η, choisis arbitrairement respectant $0 < \alpha < \eta \leq 1$. Comme c'est le cas précédemment, les paramètres α et η accélèrent la convergence de l'algorithme d'itération permettant de parvenir à une solution; Singh et Mohl (1996) pensent que les valeurs $\alpha = .67$ et $\eta = .9$ donnent de bons résultats, en pratique. Enfin, $w_k^{(\nu-1)} = a_k g_k^{(\nu-1)}$; $\nu = 2, 3, \ldots$ où

$$g_k^{(\nu-1)} = \frac{a_k^{(\nu-2)*}}{a_k} \left[1 + (\boldsymbol{X} - \hat{\boldsymbol{X}}_w^{(\nu-2)})' \left(\sum_{j \in s} a_j^{(\nu-2)*} \boldsymbol{x}_j \boldsymbol{x}_j' \right)^{-1} \boldsymbol{x}_k \right]; \qquad \nu = 2, 3, \ldots$$

et où $\hat{\boldsymbol{X}}_w^{(\nu-2)}$ est défini comme précédemment. Les valeurs de départ sont données par $a_k^{(0)*} = a_k$ et $w_k^{(0)} = a_k$.

Une propriété de la fonction de distance Huang-Fuller modifiée et de la fonction de distance à rétrécissement-minimisation est que les contraintes de calage (équation (2.2)) sont respectées à chaque itération, tandis que les restrictions applicables à la fourchette de valeurs du coefficient g ne le sont qu'à la convergence. Avec la fonction de distance MCG restrictive et la fonction de distance IQ restrictive, les contraintes relatives à la fourchette de valeurs du coefficient g sont respectées à chaque itération, mais les restrictions de calage ne le sont qu'à la convergence. Or, il vaut souvent la peine de plafonner le nombre d'itérations nécessaires pour atteindre la convergence; on peut intégrer cette spécification à l'algorithme d'itération pour faciliter le calcul. Quand la limite supérieure est franchie à cause d'une convergence trop lente, l'application de l'algorithme prend fin prématurément. Quoi qu'il en soit, les contraintes de calage sont respectées avec la fonction de distance Huang-Fuller modifiée et sa fonction de distance à rétrécissement-minimisation. Pareillement, avec la fonction de distance MCG restrictive et la fonction de distance IQ restrictive ce sont les contraintes relatives à la fourchette de valeurs qu'on respecte.

Le comportement des coefficients g de certaines fonctions de distance a été abondamment étudié. Il suffit de lire, par exemple, Deville, Särndal et Sautory (1993). Stukel et Boyer (1992) montrent de façon empirique que les fonctions de distance MCG et IQ, et leurs variantes restrictives, dotées de larges limites, donnent des coefficients g dont la distribution suit assez bien la normalité pour un ensemble de données particulier. Lorsqu'on impose des limites plus étroites aux fonctions de distance restrictives, la distribution révèle cependant un «empilement» de coefficients g aux limites inférieure et supérieure. Malgré tout, les fonctions de distance restrictives semblent déboucher sur des estimations ponctuelles voisines de celles que donnent les fonctions de distance non restrictives, même avec une application très rigoureuse des limites, comme le montreront les résultats de notre étude empirique. Nous n'avons toutefois pas examiné le biais des estimateurs ponctuels et des estimateurs de variance quand on resserre à l'extrême les fonctions de distance. Notre analyse présente de l'intérêt pour les enquêtes comme l'EPA, dont le régime d'estimations existant a été élargi et qui permettent désormais à l'utilisateur de choisir entre les fonctions de distance MCG restrictive ou une fonction de distance à rétrécissement-minimisation, en plus de la fonction de distance MCG offerte jusqu'à présent.

3. ESTIMATION DE LA VARIANCE DES ESTIMATEURS DE CALAGE

La variance exacte de l'estimateur de calage $\hat{Y}_w$ est impossible à calculer, car l'estimateur ponctuel lui-même n'est pas linéaire. De plus, il n'existe pas de méthode non biaisée explicite pour estimer la variance. C'est pourquoi on recourt souvent à des méthodes à peu près non biaisées comme celle de Taylor et celle du jackknife, dans la pratique.

On se sert rarement de l'échantillonnage à tirage «avec remise» avec les plans d'échantillonnage stratifiés à degrés multiples en pratique, car on ne désire guère prélever la même unité plus d'une fois. Par conséquent, la majorité des enquêtes recourent à l'échantillonnage «sans remise», du moins lors de la première phase de l'échantillonnage. Malgré cela, si la fraction de la première phase de l'échantillonnage est faible (moins de 10%, par exemple), il pourrait s'avérer raisonnable d'utiliser une formule simplifiée, supposant un échantillonnage «avec remise», à la première étape, pour estimer la variance. Pareille simplification donne l'équation qui suit pour l'estimateur de régression généralisé (fonction de

distance MCG) avec un plan d'échantillonnage stratifié à degrés multiples:

$$\hat{V}_T^*(\hat{Y}_{w(\mathrm{GREG})}) = \sum_{h=1}^{L} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left[\sum_{k \in s_{hi}} a_{hik} e_{hik} - \frac{1}{n_h} \sum_{i=1}^{n_h} \sum_{k \in s_{hi}} a_{hik} e_{hik} \right]^2 \quad (3.1)$$

où s_{hi} est l'échantillon d'individus de la i-ième unité primaire d'échantillonnage (UPÉ) et de la h-ième strate, a_{hik} est le poids d'échantillonnage original du plan d'échantillonnage pour l'individu k de l'UPÉ i et de la strate h, et n_h est le nombre d'UPÉ échantillonnées dans la strate h. De plus, $e_{hik} = y_{hik} - \boldsymbol{x}'_{\boldsymbol{hik}}\hat{\beta}$ donne la valeur résiduelle associée à l'estimateur de régression, où $\hat{\beta} = (\sum_{hik \in s} a_{hik} \boldsymbol{x}_{\boldsymbol{hik}} \boldsymbol{x}'_{\boldsymbol{hik}} / c_{hik})^{-1} \sum_{hik \in s} a_{hik} \boldsymbol{x}_{\boldsymbol{hik}} y_{hik} / c_{hik}$. La formule «avec remise» donnée en (3.1) surestime la variance réelle de nombreux plans d'échantillonnage (lire Särndal, Swensson et Wretman (1992, partie 4.6)). Notons cependant que sur le plan technique, cet estimateur simplifié de la variance ne correspond *pas* à l'estimateur de Taylor, même si on l'appelle souvent ainsi pour des raisons historiques. C'est pourquoi nous en ferons autant.

Hidiroglou, Fuller et Hickman (1980) proposent une amélioration à l'équation (3.1) où le coefficient g est inclus à la formule de la variance (rappelons que $w_{hik} = a_{hik} g_{hik}$). Le résultat est le suivant:

$$\hat{V}_T(\hat{Y}_{w(\mathrm{GREG})}) = \sum_{h=1}^{L} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left[\sum_{k \in s_{hi}} w_{hik} e_{hik} - \frac{1}{n_h} \sum_{i=1}^{n_h} \sum_{k \in s_{hi}} w_{hik} e_{hik} \right]^2. \quad (3.2)$$

Särndal (1982) suggère une équation semblable à l'équation (3.2) pour un échantillonnage à deux degrés, mais avec des estimateurs de la variance de type Yates-Grundy. De leur côté, Deville et Särndal (1992) montrent qu'une fonction de distance qui obéit à un jeu de conditions générales donnera un estimateur asymptotiquement équivalent à celui obtenu grâce à la fonction de distance MCG, c'est-à-dire $\hat{Y}_{w(\mathrm{GREG})}$, indiqué en (2.5). Singh et Mohl (1996) vont un peu plus loin et y intègrent la fonction de distance Huang-Fuller modifiée et la fonction de distance à rétrécissement-minimisation. La variance asymptotique de l'estimateur de calage $\hat{Y}_w$ peut donc être considérée comme étant approximativement égale à l'estimateur $\hat{Y}_{w(\mathrm{GREG})}$. Cette dernière observation nous amène à une méthode permettant d'estimer la variance de Taylor commune à tous les estimateurs de calage, c'est-à-dire d'estimer la variance de $\hat{Y}_w$ en modifiant l'estimateur de variance de Taylor utilisé pour $\hat{Y}_{w(\mathrm{GREG})}$ plutôt qu'en effectuant une nouvelle dérivation de la formule de Taylor pour chaque fonction de distance. Pour obtenir l'estimateur de la variance d'une autre fonction de distance que la fonction MCG, on peut donc recourir à l'équation (3.2) et remplacer les poids finals $\{w_{hik}\}$ de la fonction de distance MCG par ceux de la fonction de distance à laquelle on s'intéresse.

Pour trouver l'estimateur de la variance de $\hat{Y}_w$, sans tenir compte de la fonction de distance servant à établir les poids finals calés, on recourt couramment à la méthode du jackknife. Voici la formule de la variance applicable à un plan d'échantillonnage stratifié à degrés multiples avec remise au premier degré:

$$\hat{V}_J(\hat{Y}_w) = \sum_{h=1}^{L} \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} (\hat{Y}_w(hi) - \hat{Y}_w)^2 \quad (3.3)$$

où $\hat{Y}_w(hi)$ correspond souvent à «l'estimateur de répétition», à savoir le reste de l'échantillon une fois que l'UPÉ i a été retirée de la strate h. On obtient $\hat{Y}_w(hi)$ en calculant de nouveau $\hat{Y}_w$ après retrait de la i-ième UPÉ de la h-ième strate avec $h = 1, ..., L$; $i = 1, ..., n_h$, c'est-à-dire quand on modifie les poids d'échantillonnage originaux pour refléter l'exclusion de l'UPÉ en question et recalcule les coefficients g d'après l'échantillon réduit ou la répétition. Enfin, on obtient l'estimateur jackknife en retirant successivement les UPÉ une à la fois, en calculant l'estimateur de répétition correspondant, puis en érigeant l'estimateur final grâce à (3.3). L'estimateur de la variance jackknife indiqué en (3.3) est le plus conservateur des quatre cités par Wolter (1985) dans sa longue dissertation de la question.

Il vaut la peine de souligner que Yung et Rao (1996) parviennent à l'équation (3.2) comme approximation de l'estimateur de la variance jackknife (3.3) pour l'estimateur de régression généralisé. Ces auteurs appellent l'équation (3.2) «estimateur de la variance jackknife linéarisé». Leur simulation révèle que le biais (conditionnel et inconditionnel) de l'estimateur de la variance de Taylor (équation (3.1)), de l'estimateur de la variance jackknife linéarisé (équation (3.2)) et de l'estimateur de la variance jackknife (équation (3.3)) ont un comportement semblable. Bien que leur simulation insiste sur les estimateurs de la variance de l'estimateur de régression généralisé sans restrictions, notre propre simulation, qui suit, porte sur les estimateurs de la variance de l'estimateur de régression généralisé et des estimateurs reposant sur d'autres fonctions de distance restrictives et non restrictives.

4. SIMULATION DE MONTE CARLO

4.1 Description

Pour comparer la performance des estimateurs de calage à celle des estimateurs de la variance de Taylor et jackknife correspondants, nous avons entrepris une simulation de Monte Carlo dans laquelle nous avons examiné les propriétés fréquentistes de leur échantillon tiré d'une population finie.

Les données de l'Enquête sur la population active (EPA) de décembre 1990 pour Terre-Neuve nous ont servi à créer une population finie d'où a été prélevé de façon répétitive un échantillon. L'EPA est l'enquête-ménage par échantillonnage permanente la plus importante entreprise par Statistique Canada. Les données mensuelles sur le marché du travail sont recueillies dans le cadre d'un plan d'échantillonnage complexe à degrés multiples qui comporte plusieurs niveaux de stratification. On trouvera une description détaillée du plan d'échantillonnage de cette enquête avant sa restructuration

survenue en 1991, dans Singh, Drew, Gambino et Mayda (1990). Brièvement, précisons que les provinces sont stratifiées en «régions économiques», soit de vastes zones de structure économique analogue. Terre-Neuve compte quatre régions économiques. Ces dernières sont subdivisées en «unités autoreprésentatives» (UAR) et «unités non autoreprésentatives» (UNAR), elles-mêmes scindées en strates de niveau inférieur. Les UAR correspondent aux agglomértions de plus de 15,000 habitants comme St. John's et Cornerbrook, à Terre-Neuve. Au niveau le plus détaillé, Terre-Neuve compte 45 strates, comprenant chacune moins de six unités primaires d'échantillonnage (UPÉ), un nombre insuffisant pour l'échantillonnage dans le cadre de notre simulation. Les 45 strates ont donc été regroupées en 18, renfermant chacune de six à 18 UPÉ. Le regroupement n'a pas modifié la région économique, ni les régions métropolitaines de recensement (RMR) de St. John's et de Cornerbook.

Pour la simulation de Monte Carlo, on a tiré $R = 4{,}000$ échantillons d'environ 1,000 unités dans la «population» de Terre-Neuve (égale à 9,152) en vertu d'un plan d'échantillonnage à deux degrés. Des strates groupées venant des UNAR, on a sélectionné deux UPÉ par échantillonnage avec remise avec probabilité proportionnelle à la taille au premier degré. La taille de l'échantillon correspondait au nombre de logements de l'UPÉ. Au deuxième degré, on a retenu un logement sur cinq dans les UPÉ sélectionnées, par échantillonnage aléatoire simple à tirage sans remise. Pour les strates groupées appartenant à des UAR, on a choisi trois UPÉ par échantillonnage avec remise à la première étape avec probabilité proportionnelle à la taille. Au second degré, tous les logements des UPÉ échantillonnées ont été retenus, si bien que cette partie du plan d'échantillonnage se bornait à un échantillonnage de grappes complètes. Cette solution s'est avérée nécessaire faute d'un nombre suffisant de logements par UPÉ pour procéder à un sous-échantillonnage dans les UAR. Le choix de deux UPÉ dans les strates UNAR plutôt que trois, comme dans les strates UAR, est attribuable au fait qu'en général, les premières comptent moins d'UPÉ que les secondes. Dans l'ensemble, 47 UPÉ ont été échantillonnées. Dans l'un ou l'autre cas (UNAR ou UAR), tous les habitants des logements se sont tous retrouvés dans l'échantillon. Bien que ce plan soit un amalgame des plans à un et à deux degrés, nous le considérerons comme un plan d'échantillonnage à deux degrés pour plus de commodité.

Nous nous sommes intéressés à Y, le nombre total de chômeurs. Cette valeur a été obtenue par application de l'équation $Y = \sum_{k \in U} y_k = \sum_{k=1}^{9152} y_k$ à la population fixée, où $y_k = 1$ si l'individu k est au chômage et est égal à 0 dans les autres cas. Pour chacun des $R = 4{,}000$ échantillons, on a calculé $\hat{Y}_w$, soit le nombre total estimatif de chômeurs $\hat{Y}_w = \sum_{k \in s} w_k y_k$. La valeur $\{w_k : k \in s\}$ a été déterminée au moyen des six fonctions de distance dont on a déjà parlé:

(1) la fonction de distance généralisée des moindres carrés (MCG) (équation (2.4)),
(2) la fonction de distance par itération du quotient (IQ) (équation (2.7)),
(3) la fonction de distance MCG restrictive (MCGR) (équation (2.8)),
(4) la fonction de distance IQ restrictive (IQR) ou à Logit (équation (2.9)),
(5) la fonction de distance Huang-Fuller modifiée (HFM) ($\alpha = 0.67$, $\delta = 0.8$) (équation (2.10)), et
(6) la fonction de distance à rétrécissement-minimisation (RM) ($\alpha = 0.67$, $\eta = 0.9$) (équation (2.11)).

Les quatre ensembles de limites qui suivent ont été appliqués aux quatre dernières fonctions de distance pour restreindre la minimisation (i) $L = 0$, $U = 4$; (ii) $L = 0.4$, $U = 2$; (iii) $L = 0.68$, $U = 1.6$ et (iv) $L = 0.8$, $U = 1.3$. On a ainsi obtenu 18 estimateurs ponctuels. Pour chacun d'eux, le calage reposait sur les données auxiliaires issues des projections provinciales du recensement pour dix catégories âge/sexe mutuellement exclusives et exhaustives (catégories: < = 14, 15-24, 25-44, 45-64, > = 65, pour chaque sexe) et les quatre régions économiques de Terre-Neuve. Les données auxiliaires sur chaque individu formaient donc un vecteur de 14 unités comptant exactement deux fois la valeur un et 12 fois la valeur zéro. On a cependant dû réduire les dimensions du vecteur à 13 quand on a recouru à la méthode de Newton-Raphson pour résoudre l'équation (2.3). Parallèlement, on a postulé que $c_k = 1$ pour les quatre premières fonctions de distance.

Nous avons calculé l'estimateur de la variance jackknife de l'équation (3.3) pour chacun des $R = 4{,}000$ échantillons et des 18 estimateurs ponctuels. Nous avons aussi calculé l'estimateur de la variance de Taylor de l'équation (3.2) et procédé à la modification suggérée à la partie 3 pour les autres fonctions de distance que la fonction MCG. Notons que puisqu'on a recouru à un échantillonnage avec remise avec probabilité proportionnelle à la taille plutôt qu'à un échantillonnage sans remise au premier degré, l'estimateur de variance de l'équation (3.2) convenait tout à fait à la simulation. Enfin, on s'est servi de la formule (3.1) avec la fonction de distance MCG seulement, afin de voir ce qui se produirait si on omettait les coefficients g dans l'estimateur de la variance.

Plusieurs propriétés fréquentistes ont été examinées pour chacune des six fonctions de distance précitées. Les voici.

(A) On a estimé le biais relatif en pour cent du nombre estimatif de chômeurs (par rapport à l'ensemble de la population) au moyen de l'équation:

$$\frac{E_M(\hat{Y}_w) - Y}{Y} * 100 \qquad (4.1)$$

où

$$E_M(\hat{Y}_w) = \frac{1}{R} \sum_{r=1}^{R} \hat{Y}_{w_r}$$

représente l'espérance de Monte Carlo de l'estimateur ponctuel $\hat{Y}_w$ pour R échantillons et $\hat{Y}_{w_r}$, la valeur de $\hat{Y}_w$ pour l'échantillon r.

(B) Le biais relatif en pour cent de l'estimateur de la variance de Taylor/jackknife (par rapport à la variance réelle) est estimé au moyen de l'équation

Tableau 1
Biais relatif en pour cent des estimateurs ponctuels, et biais relatif en pour cent et coefficients de variation en pour cent des estimateurs de la variance de Taylor et jackknife (échantillon d'environ 1000)

Fonction de distance		Biais relatif en % estimateur ponctuel	Biais relatif en % variance Taylor	Biais relatif en % variance jackknife	CV en % variance Taylor	CV en % variance jackknife	Nombre d'échantillons écartés (sur 4000)
MCG (régression)		.11	-6.01 (éq 3.1) -5.82 (éq 3.2)	-1.73	60.79 (éq 3.1) 59.60 (éq 3.2)	62.86	0
MCG restrictive	($L = 0$, $U = 4$)	.11	-5.82	-1.73	59.60	62.86	0
	($L = .4$, $U = 2$)	.10	-5.36	-1.27	59.93	63.21	32
Itération du quotient		.52	-6.20	0.84	59.45	63.35	0
IQ restrictive	($L = 0$, $U = 4$)	.50	-6.09	-0.31	59.48	63.47	0
	($L = .4$, $U = 2$)	.46	-5.69	-0.39	59.81	64.21	32
Huang-Fuller modifiée	($L = 0$, $U = 4$)	.11	-5.82	-1.73	59.60	62.86	0
	($L = .4$, $U = 2$)	.10	-5.36	-1.20	59.94	63.27	32
Rétrécissement-minimisation	($L = 0$, $U = 4$)	.11	-5.82	-1.73	59.60	62.86	0
	($L = .4$, $U = 2$)	.10	-5.36	-1.27	59.94	63.25	32

$$\frac{(E_M(\hat{V}(\hat{Y}_w)) - V_{\text{true}})}{V_{\text{true}}} * 100 \qquad (4.2)$$

où

$$E_M(\hat{V}(\hat{Y}_w)) = \frac{1}{R}\sum_{r=1}^{R} \hat{V}_r(\hat{Y}_w)$$

et

$$V_{\text{true}} = \frac{1}{R}\sum_{r=1}^{R} (\hat{Y}_{w_r} - E_M(\hat{Y}_w))^2$$

et $\hat{V}_r(\hat{Y}_w)$ correspond à la valeur de $\hat{V}(\hat{Y}_w)$ (Taylor ou jackknife) de l'échantillon r.

(C) Le coefficient de variation en pour cent de l'estimateur de la variance de Taylor/jackknife (par rapport à la variance réelle) est estimé par

$$\frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R} (\hat{V}_r(\hat{Y}_w) - V_{\text{true}})^2}}{V_{\text{true}}} * 100 \qquad (4.3)$$

c'est-à-dire l'erreur quadratique moyenne de l'estimateur de la variance divisée par la variance réelle en pour cent. Bien que la plupart des études portent sur le *biais* des estimateurs de la variance, il vaut la peine d'examiner le *coefficient de variation* des estimateurs de la variance pour savoir dans quelle mesure les estimations de la variance fluctuent elles-mêmes.

Remarquons qu'il aurait pu s'avérer préférable de comparer les valeurs à l'«erreur quadratique moyenne réelle» plutôt qu'à la «variance réelle» aux équations (4.2) et (4.3). Toutefois, dans notre simulation, le biais relatif était si faible que l'écart entre les deux comparaisons est à toute fin pratique négligeable.

Enfin, pour déterminer si le nombre d'échantillons retenu était approprié, on a calculé l'erreur de Monte Carlo en prenant le coefficient de variation en pour cent de $E_M(\hat{V}(\hat{Y}_w))$ donné par:

$$\frac{\sqrt{\frac{1}{R^2}\sum_{r=1}^{R} [\hat{V}_r(\hat{Y}_w) - E_M(\hat{V}(\hat{Y}_w))]^2}}{E_M(\hat{V}(\hat{Y}_w))} * 100. \qquad (4.4)$$

Avec $R = 4{,}000$, l'erreur de Monte Carlo était constamment faible (entre 0.99% et 3.60%) pour les méthodes du jackknife et de Taylor, signe que les résultats sont stables.

4.2 Résultats de l'étude

Le tableau 1 indique le biais relatif en pour cent de l'estimateur ponctuel (équation (4.1)) et des estimateurs de la variance de Taylor et jackknife (équation (4.2)), ainsi que les coefficients de variation en pour cent des mêmes estimateurs (équation (4.3)). Le biais relatif en pour cent des estimations ponctuelles (colonne deux) est négligeable, puisqu'il varie entre 0.10% et 0.52%, soit beaucoup moins que 1% dans tous les cas. Le fait que les estimations ponctuelles présentent un biais similaire est raisonnable, étant donné l'équivalence

Tableau 2
Biais relatif en pour cent des estimateurs ponctuels, et biais relatif en pour cent et CV en pour cent des estimateurs de la variance de Taylor et jackknife (échantillon d'environ 2000)

Fonction de distance		Biais relatif en % estimateur ponctuel	Biais relatif en % variance Taylor	Biais relatif en % variance jackknife	CV en % variance Taylor	CV en % variance jackknife	Nombre d'échantillons écartés (sur 4000)
MCG (régression)		.02	-2.71 (éq 3.1) -2.61 (éq 3.2)	-1.43	23.03 (éq 3.1) 22.84 (éq 3.2)	23.29	0
MCG restrictive	($L = 0$, $U = 4$)	.02	-2.61	-1.43	22.84	23.29	0
	($L = .4$, $U = 2$)	.02	-2.61	-1.43	22.84	23.29	0
	($L = .68$, $U = 1.6$)	.02	-2.61	-1.44	22.84	23.29	0
	($L = .8$, $U = 1.3$)	.02	-2.75	-1.56	22.70	23.15	118
Itération du quotient		.25	-2.75	-1.15	22.84	23.43	0
IQ restrictive	($L = 0$, $U = 4$)	.17	-2.67	-1.36	22.84	23.30	0
	($L = .4$, $U = 2$)	.16	-2.70	-1.42	22.84	23.29	0
	($L = .68$, $U = 1.6$)	.31	-2.77	-0.49	22.83	24.20	0
	($L = .8$, $U = 1.3$)	.27	-2.91	*	22.70	*	118
Huang-Fuller modifiée	($L = 0$, $U = 4$)	.02	-2.61	-1.43	22.84	23.29	0
	($L = .4$, $U = 2$)	.02	-2.61	-1.43	22.84	23.29	0
	($L = .68$, $U = 1.6$)	.02	-2.61	-1.44	22.84	23.29	0
	($L = .8$, U = 1.3)	.02	-2.58	-1.36	22.73	23.18	116
Rétrécissement-minimisation	($L = 0$, $U = 4$)	.02	-2.61	-1.43	22.84	23.29	0
	($L = .4$, $U = 2$)	.02	-2.61	-1.43	22.84	23.29	0
	($L = .68$, $U = 1.6$)	.02	-2.61	-1.44	22.84	23.29	0
	($L = .8$, $U = 1.3$)	.02	-2.61	-1.24	22.73	23.63	118

asymptotique des estimateurs de calage par rapport à l'estimateur de régression.

La troisième colonne indique le biais relatif en pour cent de l'estimateur de la variance de Taylor. Dans ce cas, on sous-estime toujours la variance réelle, mais jamais de plus de 6.2%. En ce qui concerne l'estimateur de régression, l'inclusion du coefficient g à la formule de la variance (équation (3.1) ou équation (3.2)) ne paraît guère faire de différence; le biais ne s'améliore que légèrement lorsqu'on ajoute le coefficient g (-5.82% contre -6.01%). L'estimateur de la variance jackknife (quatrième colonne), par contre, donne constamment de meilleurs résultats que l'estimateur de la variance de Taylor. Ainsi, il sous-estime presque toujours la variance réelle, mais de moins de 2% dans tous les cas.

Toutes les fonctions de distance aboutissent à une solution, mais la fonction MCG exige un algorithme d'itération. À cause de cela, quelques-uns des 4,000 échantillons posent des problèmes de convergence, surtout aux limites extrêmes du coefficient g. On a donc rejeté les échantillons pour lesquels l'algorithme ne débouchait pas sur la convergence. Par conséquent, ces échantillons n'ont pas contribué aux diverses mesures de Monte Carlo. La dernière colonne du tableau 1 indique le nombre d'échantillons écartés. Aux limites extrêmes ($L = 0.68$, $U = 1.6$ et $L = 0.8$, $U = 1.3$), on a dû en rejeter tant (de 231 à 234 pour $L = 0.68$, $U = 1.6$ et de 1,562 à 1,602 pour $L = 0.8$, $U = 1.3$) que les résultats ne nous ont pas paru fiables. Ils ne sont donc pas signalés. Ces limites plus étroites ne sont néanmoins pas dépourvues d'intérêt. C'est la raison pour laquelle on a repris la simulation en doublant à peu près la taille de l'échantillon (hausse d'environ 1,000 à 2,000). Remarquons que Deville et Särndal (1992) montrent que pour toutes les fonctions de distance, la convergence est atteinte avec probabilité égale à un si la taille de l'échantillon est suffisante.

Les colonnes cinq et six du tableau 1 indiquent les CV en pour cent des estimateurs de la variance de Taylor et jackknife. Les coefficents de variation de toutes les fonctions de distance se ressemblent, puisque leurs valeurs fluctuent de 59.45% à 64.21%. Toutefois, ceux de la méthode jackknife sont toujours un peu au-dessus de ceux de la méthode de Taylor. Sans nier leur importance, on a déjà observé des coefficients de variation d'un tel ordre de grandeur dans d'autres simulations se rapportant à la variance. Lire par exemple Kovačević, Yung et Pandher (1995). Nous voulions cependant savoir si les principaux résultats concernant le biais des estimateurs de la variance tiendraient toujours advenant une réduction des coefficients de variation. À la suggestion d'un examinateur, nous avons donc effectué une autre simulation en portant le nombre d'UPÉ de 47 à 83, sachant que les CV des estimateurs de la variance sont à peu près

inversement corrélés au nombre d'UPÉ tirées. On a relevé e nombre d'UPÉ en faisant en sorte qu'il y ait autopondération du plan d'échantillonnage; cette approche semble avoir eu les plus grands effets sur la réduction des CV. Le second degré du plan d'échantillonnage n'a pas été modifié. La deuxième simulation a eu pour avantage secondaire de doubler approximativement la taille de l'échantillon, ce qui nous a permis de surmonter les difficultés de convergence mentionnées au paragraphe précédent.

Les résultats de la deuxième simulation apparaissent au tableau 2. La dernière colonne indique la diminution du nombre d'échantillons écartés en raison d'un problème de convergence. Les cinquième et sixième colonnes montrent que les CV ont sensiblement diminué pour se situer entre 22.70% et 24.2%, les estimateurs jackknife se caractérisant toujours par une valeur légèrement plus élevée. Comme c'était le cas auparavant, le biais relatif en pour cent de l'estimateur ponctuel reste négligeable, puisqu'il se maintient toujours nettement sous un pour cent. Lors de la simulation antérieure, le biais relatif en pour cent de l'estimateur de Taylor s'établissait constamment autour de -6%; lors de la nouvelle simulation, il fluctue toujours autour de -3%, ce qui signifie une fois de plus que la variance réelle a été sous-estimée. Encore une fois, on note très peu de variation dans le biais résultant des équations (3.1) et (3.2) pour la fonction de distance MCG. Le biais relatif en pour cent de l'estimateur jackknife (toujours d'environ -1.5%) est constamment inférieur en valeur absolue à celui de l'estimateur de Taylor (valeur absolue). Il existe cependant un cas (IQ restrictive (L = 0.8, U = 1.3)) où l'estimateur jackknife soulève un problème de convergence; on n'a pas tenu compte de ces résultats, identifiés par un «*». Fait surprenant, on ne remarque pratiquement aucun changement de biais pour les fonctions de distance restrictives avec les estimateurs de Taylor et jackknife, car les limites deviennent successivement plus étroites. En réalité, le biais relatif en pour cent des diverses fonctions de distance, tant pour l'estimateur de Taylor que pour l'estimateur jackknife, ne paraît guère varier. Signalons qu'à la deuxième simulation, l'erreur de Monte Carlo variait entre 0.37% et 2.13%.

5. CONCLUSIONS

Dans ce document, nous avons étudié le comportement des estimateurs ponctuels et des estimateurs de la variance de Taylor et jackknife correspondants pour plusieurs fonctions de distance auxquelles la théorie de calage donne accès. On a particulièrement insisté sur les fonctions de distance qui restreignent la fourchette de valeurs des coefficients g, donc éliminent un poids final éventuellement négatif ou positif mais très élevé. Tous les estimateurs ponctuels examinés se caractérisaient par un biais négligeable.

Les estimateurs jackknife et de Taylor révèlent une faible sous-estimation de la variance réelle, bien que le biais de l'estimateur jackknife soit toujours inférieur à celui de l'estimateur de Taylor (valeur absolue). Le résultat le plus étonnant demeure que, pour les estimateurs de Taylor et jackknife, le biais ne varie pratiquement pas, aussi bien quand on limite à l'extrême le coefficient g que dans le cas de limites moins étroites. Il conviendrait néanmoins de se montrer prudent lorsqu'on utilise des limites extrêmes, étant donné les difficultés de convergence qui peuvent surgir, surtout lorsqu'on recourt à la méthode jackknife pour estimer la variance, les estimateurs ponctuels devant être recalculés constamment. Si l'usage des fonctions de distance restrictives a pour but principal d'éliminer un poids éventuellement négatif ou positif très élevé, il suffirait d'imposer des limites modestes aux coefficients g.

En guise de remarque finale, soulignons qu'il est intéressant de constater que l'estimateur de jackknife a réclamé environ 97% du temps de calcul alors que la méthode de Taylor linéalisée n'en a exigé que les 3% restants. Pareille différence, qu'on pourrait qualifier d'extrême, dans le temps nécessaire au calcul pourrait avantager la méthode de Taylor si on a besoin de mesurer précisément un grand nombre de dimensions. Face aux progrès récents réalisés au niveau de l'efficacité des calculs relatifs à l'estimateur de la variance jackknife (programme WESVARPC (1995), par exemple), il se pourrait qu'on puisse compenser ce déséquilibre. Malgré cela, rappelons qu'à l'heure actuelle, le programme WESVARPC n'a amélioré l'efficacité de calcul que des plans d'échantillonnage à deux UPÉ par strate dont les estimateurs de stratification a posteriori ne présentaient qu'une dimension.

En conclusion, nos travaux ne démontrent pas de façon concluante la supériorité d'un estimateur de la variance quelconque et indiquent que les deux estimateurs fonctionnent raisonnablement bien pour toutes les fonctions de distance. Il revient donc à l'utilisateur de déterminer la combinaison de variance et de fonction de distance qui fonctionne le mieux, compte tenu des exigences du système.

REMERCIEMENTS

Les auteurs tiennent à remercier Chris Mohl qui leur a fourni une partie des codes informatiques utilisés lors de la simulation. Ils aimeraient également remercier le rédacteur en chef associé et deux examinateurs pour leurs remarques utiles sur la première version du document.

BIBLIOGRAPHIE

DEMING, W.E., et STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.

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

DEVILLE, J.-C., SÄRNDAL, C.-E., et SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

HIDIROGLOU, M.A., FULLER, W.A., et HICKMAN, R.D. (1980). SUPERCARP, Department of Statistics, Iowa State University, Ames, Iowa.

HUANG, E.T., et FULLER, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section, American Statistical Association*, 300-305.

KOVAČEVIĆ, M.S., YUNG, W., et PANDHER, G.S. (1995). Estimating the sampling variances of measures of income inequality and polarization – An empirical study. Documents de travail de la Direction de la méthodologie, HSMD-95-007E, Statistique Canada.

SÄRNDAL, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning and Inference*, 7, 155-170.

SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

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

SINGH, M.P., DREW, J.D., GAMBINO, J.G., et MAYDA, F. (1990). *Méthodologie de l'enquête sur la population active du Canada: 1984-1990.* N° 71-526 au catalogue, Statistique Canada.

STUKEL, D.M., et BOYER, R. (1992). Calibration estimation: An application to the Canadian Labour Force Survey. Documents de travail de la Direction de la méthodologie, SSMD-92-009E, Statistique Canada.

WESVARPC (1995). Westat Inc., Rockville, Maryland.

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

YUNG, W., et RAO, J.N.K. (1996). Linéarisation des estimateurs de variance jackknife dans un échantillonnage stratifié à degrés multiples. *Techniques d'enquête*, 22, 23-31.