Techniques d'enquête, décembre 1996
Vol. 22, n° 2, pp. 127-138
Statistique Canada

Application de l'estimation par régression restreinte dans une enquête-ménage

BODHINI R. JAYASURIYA et RICHARD VALLIANT[1]

RÉSUMÉ

Dans cet article, les auteurs comparent empiriquement trois méthodes d'estimation – par régression, par régression restreinte et au moyen de la méthode dite de la personne principale – utilisées dans une enquête-ménage sur les dépenses de consommation. Les trois méthodes sont appliquées à la stratification a posteriori, qui est importante dans de nombreuses enquêtes-ménages afin de corriger le sous-dénombrement de la population cible. Dans les recensements externes, on dispose habituellement de chiffres de population pour des strates a posteriori pour les personnes, mais non pour les ménages. Si on a besoin d'estimations par ménage, on doit assigner un facteur de pondération unique à chaque ménage, tout en utilisant le nombre de personnes pour la stratification a posteriori. On y parvient facilement en employant des estimateurs de régression pour les totaux ou les moyennes, et en utilisant le nombre de personnes dans les données auxiliaires de chaque ménage. L'estimation par régression restreinte permet de mieux calculer les facteurs de pondération, car on contrôle les valeurs extrêmes et l'on peut obtenir des estimateurs présentant une variance moindre que les estimateurs de Horvitz-Thompson, tout en respectant les totaux de contrôle de la population. Les méthodes de régression permettent également d'utiliser des contrôles pour les chiffres au niveau des personnes et des ménages et pour les données auxiliaires quantitatives. Avec la méthode dite de la personne principale, les personnes sont classées dans les strates a posteriori, et les facteurs de pondération pour les personnes font l'objet d'un rajustement par quotient afin d'obtenir des totaux de contrôle de la population. De la sorte, chaque personne dans un ménage peut se voir attribuer un facteur de pondération différent. Le facteur de pondération associé à la «personne principale» est alors choisi comme facteur de pondération pour le ménage. Nous comparerons les moyennes calculées à partir des trois méthodes, ainsi que leurs erreurs-types estimées, pour un certain nombre de dépenses tirées de l'enquête sur les dépenses de consommation parrainée par le Bureau of Labor Statistics.

MOTS CLÉS: Calage; méthode de la personne principale; variance de répétition; régression restreinte.

1. INTRODUCTION

Dans les grandes enquêtes-ménages, le sous-dénombrement de la population cible constitue un problème de signal, qui est souvent attribuable à des taux de réponse différents parmi les sous-groupes de la population et à des lacunes dans la base de sondage. La stratification a posteriori est une méthode utilisée à l'étape de l'estimation afin de réduire les erreurs quadratiques moyennes basées sur l'information qui influe sur les variables de réponse. L'estimateur est construit de telle façon que le nombre total estimé de personnes s'inscrivant dans chaque strate a posteriori soit égal au chiffre de population véritable. Les chiffres de population par strate a posteriori proviennent habituellement d'un recensement externe portant sur le nombre de personnes, mais pas toujours sur le nombre de ménages. Si on a besoin d'estimations par ménage, il faut assigner un facteur de pondération unique à chaque ménage, tout en utilisant les chiffres de personnes pour la stratification a posteriori. On y parvient en estimant par régression les valeurs totales ou moyennes portant sur les chiffres de personnes dans les données auxiliaires de chaque ménage. L'estimation par régression restreinte contrôle les facteurs de pondération extrêmes, et peut fournir des estimateurs dont la variance est moindre que celles de l'estimateur de Horvitz-Thompson, tout en permettant de respecter les totaux de contrôle de la population. Une autre méthode utilisée dans certaines enquêtes est la méthode dite de la *personne principale* (PP) (Alexander 1987), dans laquelle le facteur de pondération du ménage est basé sur la personne désignée comme «personne principale» dans chaque ménage. Les personnes sont classées dans des strates a posteriori, et les facteurs de pondération des personnes font l'objet d'un rajustement des quotients afin d'obtenir des totaux de contrôle de la population, ce qui permet d'attribuer à chaque personne dans un ménage un facteur de pondération différent. Le facteur de pondération associé à la personne principale est ensuite assigné au ménage. Cette méthode spéciale est difficile à analyser théoriquement. Bien qu'on puisse facilement les corriger pour tenir compte du sous-dénombrement de la population, les estimateurs de régression qui font l'objet du présent article fournissent automatiquement un facteur de pondération pour les ménages qui n'est basé sur aucun de leurs membres particuliers. Lemaître et Dufour (1987) ont traité de l'utilisation, par Statistique Canada, de l'estimateur de régression dans un tel contexte.

On constate une utilisation croissante des estimateurs de régression dans les enquêtes, tant dans les travaux théoriques que dans les enquêtes réelles. Statistique Canada a incorporé l'estimateur de régression général dans son logiciel Système généralisé d'estimation (GES), qui est maintenant utilisé dans bon nombre de ses enquêtes (Estevao, Hidiroglou et Särndal 1995). Fuller, Loughin et Baker (1993) ont traité d'une application de cette méthode dans l'enquête américaine sur la consommation alimentaire (USDA). Un des attraits de

[1] Bodhini R. Jayasuriya et Richard Valliant, U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue, N.E., Room 4915, Washington, DC 20212, U.S.A.

l'estimation par régression est que bon nombre des techniques standard d'enquêtes, y compris l'estimateur de stratification a posteriori mentionné ci-dessus, sont des cas spéciaux des estimateurs de régression. En outre, les estimateurs de régression incorporent avec plus de souplesse les données auxiliaires que les autres méthodes courantes. Par exemple, dans une enquête-ménage, cette méthode accepte facilement les données auxiliaires au niveau des personnes et au niveau des ménages, qu'elles soient de nature qualitative ou quantitative. Parmi les travaux portant sur l'estimation par régression et la stratification a posteriori, mentionnons ceux de Bethlehem et Keller (1987), Casady et Valliant (1993), Deville et Särndal (1992), Deville, Särndal et Sautory (1993) et Zieschang (1990).

Dans la présente étude, nous comparons l'estimateur de régression avec l'estimateur PP actuellement utilisé au Bureau of Labour of Statistics (BLS). On peut écrire chaque estimateur sous forme d'une somme pondérée des valeurs des variables de réponse dans l'échantillon. Chaque facteur de pondération est ensuite interprété de façon classique comme étant le nombre de personnes dans la population qui aurait la valeur correspondante de la variable de réponse. Cette interprétation demande que chaque facteur de pondération soit égal ou supérieur à un. L'estimateur de régression par les moindres carrés ordinaires présente un désavantage, car il peut donner des facteurs de pondération non positifs. Un certain nombre d'auteurs ont suggéré des méthodes pour surmonter ce problème. La méthode la plus facile est probablement celle qui a été présentée par Deville et Särndal (1992), qui consiste à éliminer tout facteur de pondération négatif, et à éliminer également les facteurs de pondération extrêmes. Les estimateurs de régression restreinte, obtenus au moyen de ces nouveaux facteurs de pondération, sont également comparés à l'estimateur de régression original et à l'estimateur PP.

Nous décrivons les trois différents estimateurs à la section 2. La section 3 présente une application de ces procédures, dans le cadre de l'enquête sur les dépenses de consommation faite par le BLS – le même contexte que celui qui est décrit par Zieschang (1990). Nous comparons les coefficients de variation pour un certain nombre de variables cibles de l'enquête, pour la population entière et pour un certain nombre de domaines. La section 4 est un résumé de nos conclusions.

2. ESTIMATEURS DE RÉGRESSION, DE CALAGE ET DE PERSONNE PRINCIPALE

Tout d'abord, présentons brièvement l'estimateur de régression. Un échantillon s de taille n est choisi dans une population finie U de taille N. Supposons que la probabilité de choisir la i-ième unité soit π_i. L'échantillon peut être à deux degrés, et l'unité peut être soit l'unité primaire d'échantillonnage, soit l'unité secondaire d'échantillonnage. Nous n'avons nul besoin ici de compliquer la notation en introduisant des indices inférieurs explicites pour chaque degré de l'échantillonnage. Dénotons la variable qui nous intéresse par y et supposons que sa valeur dans la i-ième unité, y_i, soit observée pour chaque $i \in s$. Supposons qu'il existe K variables auxiliaires $x_1, x_2, ..., x_K$ dont nous connaissons les valeurs pour chaque $i \in S$. Définissons $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{iK})'$, pour chaque $i \in U$, où x_{ik} représente la valeur de la variable x_k dans l'unité i. Représentons par $\boldsymbol{X} = (X_1, ..., X_K)'$ le vecteur de dimension K des totaux connus de la population pour les variables $x_1, x_2, ..., x_K$. L'estimateur de régression est donc motivé par le modèle de travail ξ suivant:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_K x_{iK} + \varepsilon_i \qquad (2.1)$$

pour $i = 1, ..., N$. Ici, $\beta_1, ..., \beta_K$ sont des paramètres inconnus du modèle. Les ε_i sont des erreurs aléatoires, pour lesquelles $E_\xi(\varepsilon_i) = 0$ et $\text{var}_\xi(\varepsilon_i) = \sigma_i^2$ pour $i = 1, ..., N$. Nous utilisons l'expression «modèle de travail» afin de souligner le fait que le modèle est probablement erroné jusqu'à un certain point. Dans l'enquête sur les dépenses de consommation (CE), l'unité d'analyse, indexée par i, est une unité de consommation (UC), qui est similaire à un ménage et qui est définie plus en détail à la section 3. La valeur y_i peut être les dépenses alimentaires totales par les UC et les x_{ik} peuvent être diverses caractéristiques des UC, par exemple le nombre de personnes d'âges différents ou le revenu de l'UC, qui peuvent influer sur les dépenses alimentaires de l'UC. La variance des dépenses peut dépendre de la taille de l'UC, de telle sorte qu'il serait raisonnable que l'expression σ_i^2 soit proportionnelle au nombre de personnes dans l'UC. Nous incluons une coordonnée à l'origine dans certains de nos modèles, en mettant la première variable auxiliaire x_1 égale à 1. On définit comme suit un estimateur de régression linéaire pour la population totale y:

$$\hat{y}_R = \hat{y}_\pi + (\boldsymbol{X} - \hat{\boldsymbol{x}}_\pi)' \hat{\beta} \qquad (2.2)$$

où $\hat{y}_\pi$ dénote le π-estimateur (ou estimateur de Horvitz-Thompson) de la population totale de y, c.-à-d.

$$\hat{y}_\pi = \sum_{i \in s} a_i y_i \qquad (2.3)$$

où $a_i = 1/\pi_i$. En outre, $\hat{\boldsymbol{x}}_\pi = (\hat{x}_{1\pi}, ..., \hat{x}_{K\pi})'$ est le vecteur des π-estimateurs des totaux de population pour les variables $x_1, x_2, ..., x_K$ et

$$\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_K)' = \left[\sum_{i \in s} \frac{a_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\sigma_i^2}\right]^{-1} \sum_{i \in s} \frac{a_i \boldsymbol{x}_i y_i}{\sigma_i^2}. \qquad (2.4)$$

Nous supposons que l'expression $\sum_{i \in s} a_i \boldsymbol{x}_i \boldsymbol{x}_i' / \sigma_i^2$ est non singulière. Même si le modèle (2.1) est infructueux jusqu'à un certain point, $\hat{y}_R / N$ est un estimateur cohérent avec le plan de sondage pour la moyenne $\overline{Y}$ de la population, peu importe que le modèle supposé soit vrai ou faux. Cela est évident d'après l'expression (2.2). Si $\hat{y}_\pi / N$ et $\hat{\boldsymbol{x}}_\pi / N$ sont des estimateurs convergent avec le plan de sondage de $\overline{Y}$ et de $\overline{\boldsymbol{X}}$, qui est le vecteur des moyennes de la population des variables auxiliaires, alors le deuxième terme dans $\hat{y}_R / N$ converge vers zéro, tandis que le premier converge vers $\overline{Y}$. Pour plus de détails, voir Särndal, Swensson et Wretman (1992).

On peut également exprimer l'estimateur de régression $\hat{y}_R$ sous forme d'une somme pondérée des valeurs y_i de l'échantillon, ce qui est une caractéristique souhaitable pour les enquêtes. On peut voir aisément que l'expression (2.2) peut être récrite sous la forme $\hat{y}_R = \sum_{i\in s} w_i y_i$ où

$$w_i = a_i\left[1 + (\boldsymbol{X} - \hat{\boldsymbol{x}}_\pi)'\boldsymbol{A}^{-1}\frac{\boldsymbol{x}_i}{\sigma_i^2}\right] \tag{2.5}$$

et où $\boldsymbol{A} = \sum_{i\in s} a_i \boldsymbol{x}_i \boldsymbol{x}_i'/\sigma_i^2$. Les coefficients de pondération dépendent de l'échantillon par l'intermédiaire des x_i qui sont dans l'échantillon, mais cela est également vrai pour bon nombre d'estimateurs d'enquête, y compris l'estimateur de stratification a posteriori. Toutefois, ces facteurs de pondération ne dépendent pas de la variable y particulière étudiée, ce qui signifie que l'on peut utiliser un ensemble de facteurs de pondération w_i pour toutes les estimations.

On calcule de manière triviale une moyenne par unité, soit: $\hat{\bar{y}}_R = \hat{y}_R/\hat{N}$ où $\hat{N} = \sum_{i\in s} w_i$. Si nous calculons les totaux des valeurs auxiliaires $\boldsymbol{x}_i$, alors

$$\begin{aligned}\sum_{i\in s} w_i \boldsymbol{x}_i' &= \sum_{i\in s}\left[a_i \boldsymbol{x}_i' + (\boldsymbol{X} - \hat{\boldsymbol{x}}_\pi)'\boldsymbol{A}^{-1}\frac{a_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\sigma_i^2}\right] \\ &= \boldsymbol{X}',\end{aligned} \tag{2.6}$$

c'est-à-dire que nous reproduisons les totaux connus de la population. Ceci est également une caractéristique de l'estimateur de stratification a posteriori.

L'estimateur de β dans l'expression (2.4) ne tient compte d'aucune corrélation entre les erreurs dans le modèle (2.1). Dans les populations en grappes, il peut y avoir une corrélation entre les unités qui sont géographiquement près les unes des autres, p. ex., les UC d'un même quartier. L'utilisation d'une matrice des covariances $\boldsymbol{V}$ complète peut donner un résultat encore plus optimal (p. ex., voir Casady et Valliant 1993, et Rao 1994). Bien que l'utilisation d'une matrice des covariances $\boldsymbol{V}$ complète puisse réduire la variance de $\hat{\beta}$, les éléments de $\boldsymbol{V}$ dépendront de la variable y particulière étudiée, et le calcul de $\boldsymbol{V}$ est habituellement pénible. Par conséquent, il est intéressant et commode de considérer le cas simple où $\boldsymbol{V} = \text{diag}(\sigma_i^2)$, qui donne l'expression (2.2). On notera que lorsque la variance du plan de sondage $\text{var}_p(\hat{y}_R)$ est estimée, il est nécessaire d'utiliser une méthode qui reflète de façon appropriée les grappes et les autres complexités du plan de sondage.

L'estimateur de régression a toutefois le désavantage que les facteurs de pondération peuvent avoir des valeurs beaucoup trop grandes ou faibles, voire négatives. Les estimateurs de calage restreint de Deville et Särndal (1992), que nous présentons maintenant, ajoutent des contraintes afin de limiter la taille des facteurs de pondération. Les estimateurs de calage sont formés par minimisation d'une distance donnée, F, entre un facteur de pondération initial et le facteur de pondération final, compte tenu de certaines contraintes. Ces contraintes peuvent comporter les variables auxiliaires disponibles, et ces dernières se trouvent donc incorporées dans l'estimateur. L'estimateur de régression présenté ci-dessus est un cas spécial de l'estimateur de calage dans lequel F est définie comme étant la fonction de distance par les moindres carrés généralisés,

$$F(w_i, a_i) = \frac{a_i c_i}{2}\left(\frac{w_i}{a_i} - 1\right)^2$$

pour $i = 1, ..., n$, les valeurs c_i étant un facteur de pondération positif connu (p. ex., si $c_i = \sigma_i^2$ ou $c_i = 1$) associé à l'unité i, et w_i le facteur de pondération final. La distance totale de de l'échantillon $\sum_{i\in s} F(w_i, a_i)$ est minimisée, selon les contraintes

$$\sum_{i\in s} w_i \boldsymbol{x}_i = \boldsymbol{X}. \tag{2.7}$$

Sous cette forme, on peut écrire les facteurs de pondération de l'estimateur de régression pour la population totale de y indiquée dans (2.5), comme suit :

$$w_i = a_i g(c_i^{-1}\lambda' \boldsymbol{x}_i) \tag{2.8}$$

pour $i = 1, ..., n$, où

$$g(u) = 1 + u, \tag{2.9}$$

pour $u \in \Re$ et λ est un multiplicateur de Lagrange calculé dans le processus de minimisation. La forme particulière de w_i avec $c_i = \sigma_i^2$ pour l'estimateur de régression a été donnée dans (2.5). Afin d'éliminer les extrêmes, on peut définir les facteurs de pondération en restreignant g de telle sorte que:

$$g(u) = \begin{cases} L & \text{si } u < L - 1 \\ 1 + u & \text{si } L - 1 \le u \le U - 1 \\ U & \text{si } u > U - 1. \end{cases} \tag{2.10}$$

Avec cette définition de g, les facteurs de pondération w_i satisfont à l'expression

$$L < w_i/a_i < U \tag{2.11}$$

pour $i = 1, ..., n$, de sorte que L et U puissent être choisis de façon à refléter l'écart voulu par rapport aux facteurs de pondération initiaux a_i. En choisissant $L > 0$, on s'assure que les facteurs de pondération sont positifs, et la valeur U est choisie de façon à être aussi petite qu'il convient afin d'empêcher des facteurs de pondération trop grands. Les facteurs de pondération obtenus par régression restreinte doivent être calculés de manière itérative; Stukel et Boyer (1992) ont publié un algorithme facilement programmable. Une autre méthode pour calculer les facteurs de pondération soumis à des restrictions consiste à utiliser la régression par crêtes, comme l'ont fait Bardsley et Chambers (1984).

Dans la plupart des enquêtes-ménages, la stratification a posteriori sert avant tout à corriger le sous-dénombrement de la population cible à cause de la base de sondage et de l'échantillon. Aux États-Unis, il y a peu de dénombrements fiables des ménages dans la population qui sont utilisables pour la stratification a posteriori. Par conséquent, on utilise habituellement les chiffres de personnes dans la population

pour calculer les totaux de contrôle des strates a posteriori. Ce désaccord entre l'unité d'analyse (le ménage) et l'unité de stratification a posteriori (la personne), lorsqu'une caractéristique du ménage présente un intérêt, a mené à l'élaboration de la méthode PP, utilisée dans les enquêtes CE (dépenses de consommation) et Current Population (enquêtes démographiques).

Dans la méthode PP décrite par Alexander (1987), on débute le processus de pondération en utilisant un facteur de pondération à base simple, a_i, que l'on corrige pour tenir compte des non-réponses. Le facteur de pondération corrigé est assigné à chaque personne dans le ménage, et les facteurs de pondération pour les personnes sont par la suite de nouveau corrigés afin que leur somme corresponde aux totaux de contrôle pour les personnes dans la population, par âge, race et sexe. Cette dernière correction peut donner à des personnes au sein d'un même ménage des pondérations différentes. On assigne ensuite au ménage le facteur de pondération de la personne désignée comme «personne principale» de ce ménage. Cette méthode comporte un certain élément d'arbitraire, et elle est difficile à analyser mathématiquement. Le but de cette recherche n'était pas de voir s'il était possible d'améliorer la méthode PP, mais plutôt d'utiliser la version actuelle de cette méthode comme base commode pour mesurer la performance d'autres estimateurs.

On peut formuler les estimateurs de régression (ordinaire et restreinte), de telle sorte que les totaux de contrôle pour les personnes dans la population soient satisfaits, toutes les personnes d'un même ménage ayant le même facteur de pondération, et sans qu'il soit nécessaire de choisir arbitrairement parmi les facteurs de pondération calculés pour les personnes afin d'assigner une pondération au ménage. On y parvient en définissant les variables auxiliaires au niveau du ménage. Par exemple, s'il y a trois strates d'âge a posteriori et que le ménage i compte 1, 0 et 2 personnes dans ces strates, le vecteur des données auxiliaires seraient $\boldsymbol{x}_i = (1, 0, 2)'$. On remarquera que cette formulation est différente de celle de Lemaître et Dufour (1987), qui ont défini les variables auxiliaires au niveau des personnes et ont assigné à chaque personne la moyenne des données pour le ménage, soit (1/3, 0, 2/3) dans l'exemple ci-dessus. Ces auteurs ont utilisé cette méthode par la «moyenne», parce qu'ils étaient intéressés à calculer des estimations à la fois pour les personnes (p. ex., le nombre de personnes employées) et pour les ménages (p. ex., les familles économiques). Nous n'avons besoin que d'un facteur de pondération pour le ménage, car nos variables cibles (c.-à-d. y), comme les dépenses pour le logement ou les services publics, sont recueillies au niveau du ménage.

3. APPLICATION

Nous comparons les trois estimations (c.-à-d., régression, régression restreinte avec $L = 0.5$ et $U = 4$, et personne principale) en les appliquant aux moyennes estimées et à leurs erreurs-types estimées pour un certain nombre de dépenses tirées de l'enquête CE, parrainée par le Bureau of Labour Statistics.

L'enquête CE recueille de l'information sur les habitudes de dépenses des consommateurs américains et le coût de la vie pour ceux-ci. L'enquête comporte deux parties, soit une entrevue trimestrielle et une enquête hebdomadaire par journal. L'enquête par entrevue recueille des données détaillées sur les types de dépenses dont il est probable que les répondants puissent se souvenir pour une période de trois mois ou plus (p. ex., propriété, automobiles, gros électroménagers) et qui représentent de 60 à 70% des dépenses totales du ménage. L'enquête par journal est réalisée à domicile par la famille enquêtée pendant deux période consécutives d'une semaine, et elle recueille des données sur toutes les dépenses de la famille pendant cette période. L'échantillon est choisi en deux étapes, avec les unités primaires d'échantillonnage géographiques à la première étape, et les ménages à la seconde.

Nous avons évalué les estimateurs décrits ci-dessus pour un certain nombre de dépenses tirées de l'enquête par entrevue. Nous avons utilisé les données recueillies pendant le deuxième trimestre de 1992, ce qui représentait $n = 5156$ UC. L'unité primaire d'analyse, dans l'enquête CE, est l'unité de consommation, c'est-à-dire une famille économique au sein d'un ménage. Une unité de consommation (UC) consiste en personnes au sein du ménage qui partagent des dépenses. Ainsi, il peut y avoir plus d'une UC dans un ménage.

Nous avons étudié cinq ensembles différents de variables auxiliaires (les $\boldsymbol{x}_i$ dans la notation utilisée à la section 2). Nous avons choisi ces six variables en testant l'adéquation du modèle (2.1) avec les dépenses choisies pour différentes combinaisons des variables auxiliaires disponibles. Nous avons déterminé les combinaisons de variables auxiliaires pour lesquelles chaque coefficient de régression estimé était significatif dans une régression par les moindres carrés ordinaires, au niveau de 5%. Une étape simple mais importante qui a grandement amélioré l'ajustement avec le modèle a été d'inclure une coordonnée à l'origine. Nous avons également incorporé dans le choix des variables auxiliaires le fait que nous savions que l'enquête présentait un sous-dénombrement plus important des Noirs que des non-Noirs, et qu'il fallait en tenir compte par stratification a posteriori. Nous avons considéré cette méthode de choix des variables comme exploratoire et, par conséquent, nous avons étudié certains nombres de combinaisons afin de déterminer quel ensemble donnait les meilleurs estimateurs des dépenses moyennes. Nous avons inclus les 56 strates a posteriori basées sur l'âge, la race et le sexe, couramment utilisées dans l'enquête CE. (Ces 56 strates sont habituellement réduites, dans les activités réelles liées à l'enquête CE, en raison de la taille trop petite des échantillons dans certaines cellules.) Parmi les autres variables qui étaient statistiquement significatives dans diverses combinaisons, mentionnons la région (Nord-Est, Middle West, Sud et Ouest), l'urbanité (urbain/rural) par région, l'âge de la personne de référence de l'UC (< 25, 25-34, 35-44, 45-64, 65+), le type de ménage (propriétaire/ locataire), le revenu de l'UC avant impôt et les 56 strates a posteriori groupées par sexe et certaines par catégories d'âge, afin de former 10 catégories d'âge/race. À partir de cette information, nous avons calculé les facteurs de pondération (2.8) en utilisant les valeur g données dans l'expression (2.9), c.-à-d. les valeurs regwts, et

Tableau 1
Facteurs de pondération et variables auxiliaires correspondantes

Facteur de pondération	Variables auxiliaires	K
regwts0	Âge/race/sexe	56
regwts1	Coordonnée à l'origine, âge/race/sexe, région, urbain × région	18
regwts2	Coordonnée à l'origine, âge/race/sexe, région, urbain × région, âge de la personne de référence, type de ménage, revenu familial avant impôt	24
calwts0	Âge/race/sexe	45
calwts1	Coordonnée à l'origine, âge/race/sexe, région, urbain × région	18
calwts2	Coordonnée à l'origine, âge/race/sexe, région, urbain × région, âge de la personne de référence, type de ménage, revenu familial avant impôt	24
calwts3	Coordonnée à l'origine, âge/race/sexe, région, urbain × région, revenu familial avant impôt (tronqué à 500 000 $)	19
calwts4	Coordonnée à l'origine, âge/race/sexe, région, urbain × région, âge de la personne de référence, type de ménage	23
PP	Âge/race/sexe	56[1]

[1] L'ensemble initial de 56 strates est habituellement ramené à 30-40, en raison de la faible taille des échantillons dans certaines cellules.

(2.10), c.-à-d. les valeurs calwts. Pour les facteurs de régression ordinaire et les facteurs de régression restreinte, nous avons mis la valeur a_i égale au facteur de pondération de base corrigé, c.-à-d. $1/\pi_i$ fois un facteur de correction tenant compte des non-réponses. Pour que la matrice $\boldsymbol{A}$ présentée à la section 2 soit non-singulière, nous avons omis dans chaque $\boldsymbol{x}_i$ une des catégories dans certaines variables auxiliaires, comme la région. Pour l'application dont il est fait état ici, les totaux de population nécessaires pour évaluer $\boldsymbol{X} = (X_1, ..., X_K)'$ ont été tirés du *Statistical Abstract of the United States* (1993), dont les sources sont le Recensement de 1990 et les *Current Population Reports* publiés par le *U.S. Bureau of the Census*. Lorsque nous utilisons une coordonnée à l'origine, le total de contrôle approprié pour cette variable est le nombre d'UC dans la population, pour lequel nous avons utilisé l'estimation PP à titre de valeur substitut. Les combinaisons de variables auxiliaires utilisées pour former les différents facteurs de pondération sont présentées au tableau 1. C'est le facteur regwts0 qui utilise le plus grand nombre de strates a posteriori, soit 56 pour l'âge, la race et le sexe. Ces 56 strates sont le point de départ pour la méthode PP, mais elles sont habituellement ramenées à 30 ou 40 strates en raison de la petite taille de certaines cellules. Lorsque nous avons calculé le facteur calwts0, ces 56 strates a posteriori ont été ramenées à 45, car les contraintes imposées par les bornes L et U pouvaient être sources de singularité dans l'algorithme basé sur la matrice.

3.1 Comparaison des facteurs de pondération

Nous avons effectué diverses comparaison des facteurs de pondération calculés au moyen des différentes méthodes, mais nous ne pouvons n'en mentionner que quelques-unes ici. La figure 1 représente graphiquement les facteurs de pondération PP, regwts0, calwts0 et calwts1, par rapport aux facteurs de pondération de base corrigés. Pour les facteurs PP et regwts0, la correction correspondant au passage de a_i à w_i est beaucoup plus variable que pour calwts0 et calwts1, lesquels facteurs sont soumis aux contraintes $L = 0.5$ et $U = 4$. Une grande variabilité parmi les w_i peut donner des estimations des dépenses présentant une variance élevée, et des intervalles de confiance médiocres, car la normalité d'un grand échantillon pourrait ne pas être préservée. Bien que l'expression (2.11) signifie que $a_i/2 < w_i < 4a_i$ pour chaque i pour les facteurs calwts, la partie droite inférieure de la figure 1 montre que le facteur calwts1 satisfait à l'intervalle $a_i/2 < w_i \leq 2a_i$, pour chaque i. Ainsi, en mettant $U = 2$ ou 3, cela aurait peu d'effet sur le facteur calwts1. Le facteur calwts0 aurait été légèrement affecté si on avait mis $U = 2$, car quelques points se trouvaient à l'extérieur de la ligne de référence supérieure. Les deux parties supérieures indiquent que les facteur PP et regwts0 ne se conforment pas à la contrainte $a_i/2 < w_i < 2a_i$.

Dans cet exemple d'application, nous nous sommes peu préoccupé des facteurs de régression négatifs. Dans tout l'échantillon, une seule UC présentait une valeur négative pour les facteurs regwts1 et regwts2, tandis que le facteur regwts0 n'en présentait aucune. Toutefois, dans les répétitions utilisées pour l'estimation de la variance, décrite à la section 3.2, deux ou trois UC présentaient des facteurs de pondération négatifs dans bon nombre de répétitions, de sorte que l'utilisation de la contrainte L était plus importante dans ce cas.

3.2 Précision des estimations à partir des méthodes différentes

Bien qu'il soit instructif de comparer les facteurs de pondération, les méthodes doivent, en définitive, être jugées sur le niveau des moyennes estimées pour l'UC et de leur précision. Les erreurs-types de ces estimateurs ont été calculées par la méthode du demi-échantillonnage compensé (*balanced half-sampling* – BHS) utilisant 44 répétitions, comme on le fait actuellement dans l'enquête CE pour l'estimateur PP. L'estimateur BHS est construit de telle sorte qu'il reflète la stratification et la répartition en grappes qui sont utilisées dans l'enquête CE. La construction du demi-échantillon suit une méthode prescrite (McCarthy 1969) afin de contenir la moitié des unités primaires d'échantillonnage dans une enquête. En définissant la moyenne par UC, basée sur les UC dans le demi-échantillon α, comme étant $\hat{\bar{y}}_{R(\alpha)}$ et en définissant la moyenne pour l'échantillon complet comme étant $\hat{\bar{y}}_R$, l'estimation BHS de la variance est $V_{BHS}(\hat{\bar{y}}_{R(\alpha)}) = \sum_{\alpha=1}^{44}(\hat{\bar{y}}_{R(\alpha)} - \hat{\bar{y}}_R)^2/44$. Pour calculer chaque $\hat{\bar{y}}_{R(\alpha)}$, les mêmes étapes d'estimation utilisées pour l'échantillon complet sont répétées pour les UC faisant partie du demi-échantillon. Comme les estimations des dépenses tirées de l'enquête CE sont publiées pour divers domaines d'intérêt, nous avons calculé les moyennes et les erreurs-types pour quelques domaines choisis. Pour chacun d'entre eux, nous avons calculé le coefficient de variation (c.v.), et ensuite son quotient au c.v. du facteur de pondération PP estimé.

Pour chaque type de facteur de pondération, si le quotient de chaque c.v. des dépenses sur celui des facteurs PP est

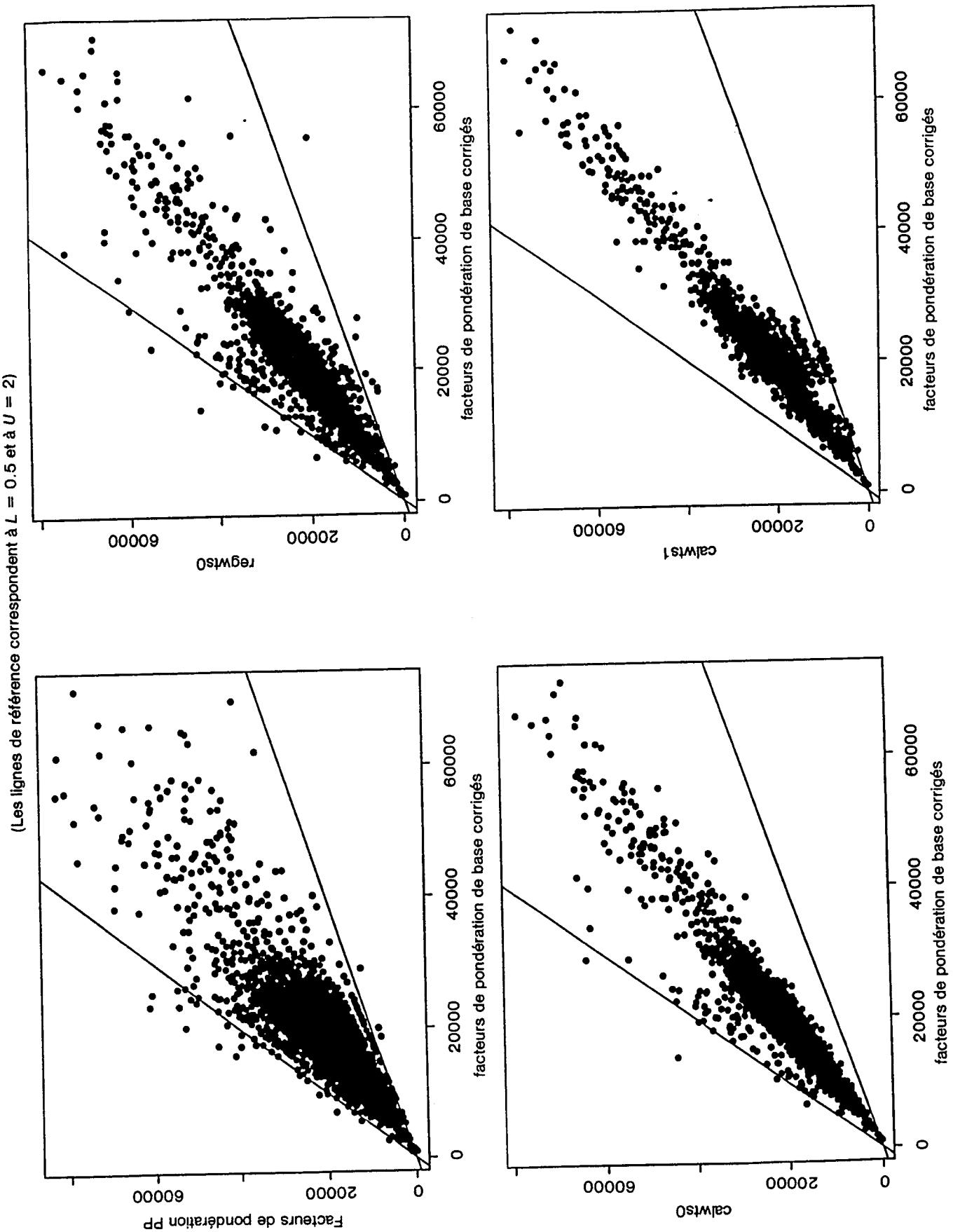

Figure 1. Graphiques de quatre ensembles de facteurs de pondération en fonction des facteurs de pondération de base corrigés

inférieur à 1, il y a alors amélioration par rapport à l'estimation PP, car, pour tous les facteurs de pondération, les estimations moyennes et les dépenses étaient très près de celles obtenues avec les estimations PP. Nous avons calculé les quotients des c.v. et les quotients des moyennes pour chacun des ensembles de facteurs de pondération décrits au tableau 1, pour chacune des dépenses choisies et pour chacun des domaines suivants :

(1) Âge de la personne de référence: < 25, 25-34, 35-44, 45-54, 55-64, 65+;
(2) Région: Nord-Est, Mid-Ouest, Sud, Ouest;
(3) Taille de l'UC: 1, 2, 3, 4, 5+;
(4) Composition des ménages: mari et femme seulement, mari et femme + enfants, mari et femme + autres, un parent + au moins un enfant < 18, personne seule et autres UC;
(5) Type de ménage: propriétaire, locataire;
(6) Race de la personne de référence: Noir, non-Noir.

Nous discuterons seulement des domaines (1) à (3) ici. En outre, les quotients pour toutes les UC, c'est-à-dire le total pour l'ensemble des domaines, ont été calculés pour chaque dépenses, et il sont présentés au tableau 2. Pour la catégorie «Toutes les dépenses», les facteurs regwts2, calwts2 et calwts3, avec des quotients de 0.79, 0.78 et 0.75 respectivement, présentent une réduction substantielle du c.v. par rapport à la méthode PP. Pour les dépenses ventilées, les facteurs regwts1 et calwts1 présentent une amélioration raisonnablement uniforme par rapport à la méthode PP, sans les pertes encourues par certains des autres facteurs pour des dépenses comme l'ameublement, les assurances personnelles et les pensions, ainsi que pour la sous-catégorie pensions et sécurité sociale.

Tableau 2
Quotients des c.v. de la méthode PP sur les c.v. obtenus par les différentes méthodes de pondération.
Le quotient minimal dans chaque rangée est en gris

Dépense	regwts			calwts				
	0	1	2	0	1	2	3	4
Toutes les dépenses	0.98	0.90	0.79	0.98	0.90	0.78	0.75	0.87
Logement	0.93	0.85	0.75	0.93	0.85	0.74	0.72	0.84
Services publics	1.08	1.03	0.94	1.07	1.03	0.88	0.91	0.92
Ameublement	1.08	1.21	3.52	1.06	1.21	2.58	2.57	1.17
Gros électroménagers	1.08	1.06	1.04	1.06	1.08	1.09	1.00	1.03
Tous les véhicules	0.90	0.89	0.98	0.91	0.89	0.98	0.97	0.90
Automobiles et camions neufs	0.95	0.91	1.01	0.96	0.91	1.02	1.02	0.91
Automobiles et camions usagés	0.98	0.94	0.96	0.97	0.94	0.97	0.96	0.95
Essence, huile moteur	1.17	1.11	1.03	1.12	1.10	0.99	0.94	1.10
Soins de santé	1.05	0.97	0.86	1.07	0.97	0.85	0.87	0.94
Éducation	0.92	0.93	1.04	0.91	0.93	1.06	1.07	0.88
Contributions pécuniaires	1.01	1.02	1.28	1.01	1.02	1.30	1.29	1.03
Assurance personnelle, pensions	1.00	0.97	1.64	1.01	0.98	1.24	0.98	0.95
Assurances-vie et autres assurances personnelles	1.08	1.02	1.53	1.08	0.98	1.38	1.33	1.01
Pensions, sécurité sociale	1.00	0.99	1.75	1.01	0.99	1.34	1.06	0.97

Les figures 2-4 sont la représentation graphique par treillis (Cleveland 1993) des valeurs c.v. et des quotients des moyennes pour calwts0 et calwts1. Le facteur calwts0 est représenté, car c'est le calage qui équivaut le plus à la méthode actuelle de stratification a posteriori. Le facteur calwts1 semble être la meilleure des diverses méthodes que nous avons examinées, au sens où elle améliore l'estimation de la rubrique «Toutes les dépenses», tout en offrant un rendement uniforme pour les différentes catégories de dépenses. Dans chaque partie des tracés, une ligne de référence verticale est tracée au point 1, c'est-à-dire le point d'égalité entre les résultats de calage et ceux obtenus pour la méthode PP. La rangée inférieure sur chaque tracé représente les quotients des moyennes des facteurs calwts0 et calwts1 sur les moyennes PP, et illustre le fait que, hormis quelques exceptions, les niveaux des moyennes obtenus par les deux méthodes de régression restreinte sont à peu près les mêmes que pour la méthode PP.

Les deux méthodes de calage choisies améliorent, dans l'ensemble, les c.v. par rapport à la méthode PP, c.-à-d. que les quotients c.v. tendent à être inférieurs à 1 pour la plupart des domaines et des dépenses, le facteur calwts1 étant quelque peu meilleur à cet égard que le facteur calwts0, Pour les domaines d'âge de référence < 25 et 65+, par exemple, 12 des 15 catégories de dépenses présentent des quotients calwts1 inférieurs à 1. Pour les tailles des UC comprises entre 1 et 4, le nombre de quotients c.v. égaux ou inférieurs à 1 est de 12, 9, 9 et 11. Il y a bien évidemment des exceptions. Pour la région du Sud, seulement six des 15 catégories de dépenses présentent, pour le facteur calwts1, des quotients c.v. égaux ou inférieurs à 1.

Les facteurs calwts2 et calwts3, qui utilisent le revenu familial avant impôt comme l'une des variables auxiliaires, présentent un rendement quelque peu erratique dans les différentes catégories, présentant parfois une amélioration importante par rapport à la méthode PP, mais, dans d'autres cas, des pertes majeures. Ce comportement est relié à la nature même de la variable revenu familial. Pour l'ensemble complet de données composé de 5,156 UC, le revenu avant impôt était positif pour 4,698 UC, nul pour 450 UC et négatif pour 8 UC. Les valeurs nulles représentent des quotients incomplets sur le revenu, tandis que les valeurs négatives sont associées à des familles pour lesquelles des pertes commerciales avaient été ajoutées à d'autres revenus. Dans un cas comme dans l'autre, ces UC entachent l'utilité de cette variable pour la prévision des dépenses. Peut-être qu'en utilisant une autre mesure du revenu, combinée avec l'imputation des postes pour les revenus manquants, on pourrait améliorer la performance des facteurs calwts2 et calwts3 pour les estimations dans les différents domaines.

Compte tenu de tout ce qui précède, les facteurs regwts1, calwts1 et calwts4 sont des choix efficaces pour ce type d'application. Le facteur calwts1 a l'avantage, par rapport à regwts1, de ne pas offrir de valeur négative. Comme le calcul de calwts4 nécessite 23 variables auxiliaires, contre seulement 18 pour le facteur calwts1, ce dernier représente le choix le plus économique. À la suite de l'analyse que nous venons de présenter, nous avons effectué une étude similaire en utilisant

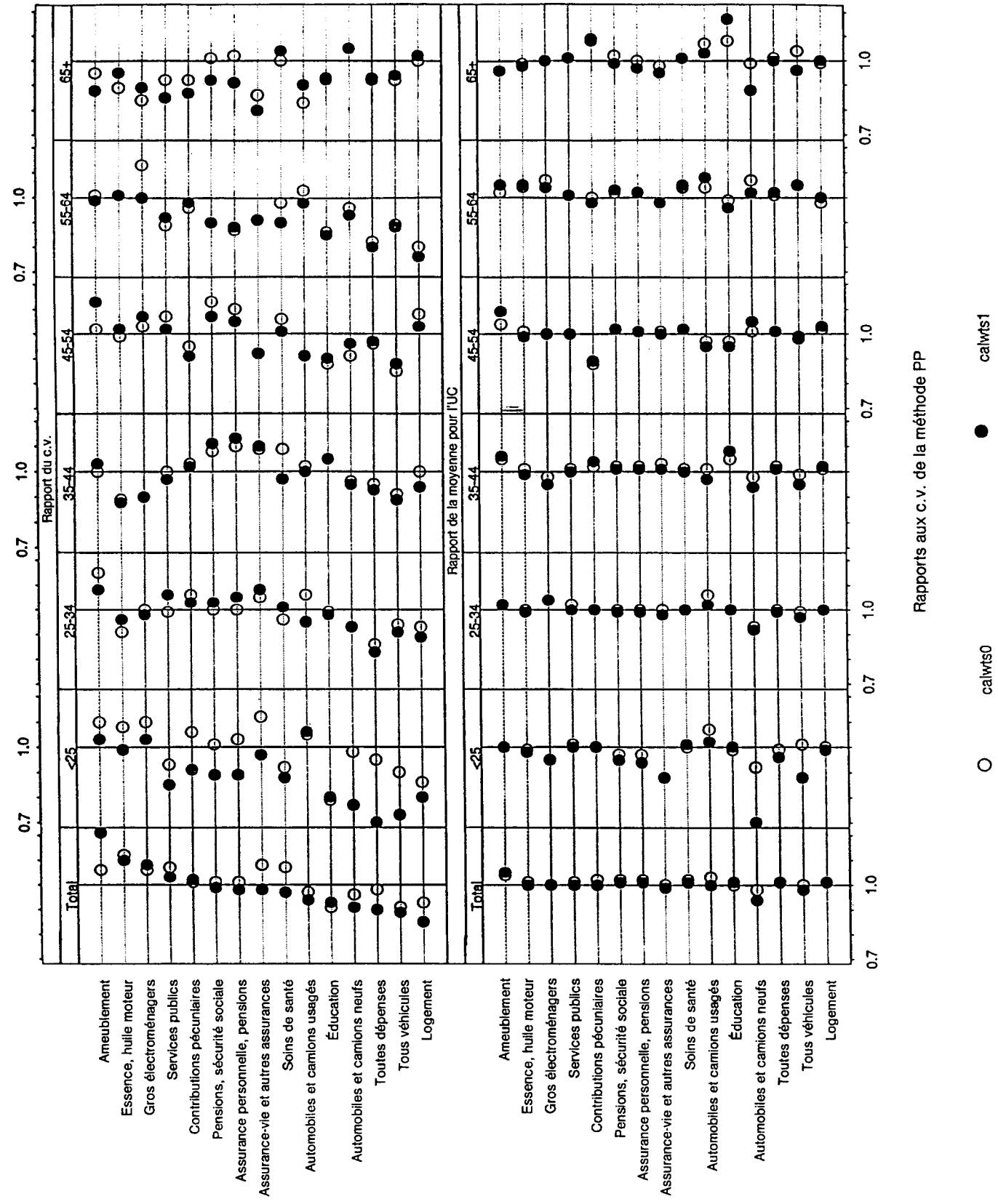

Figure 2. Rapports des c.v. et des moyennes obtenus pour deux méthodes de pondération aux c.v. de la méthode PP, selon l'âge de la personne de référence

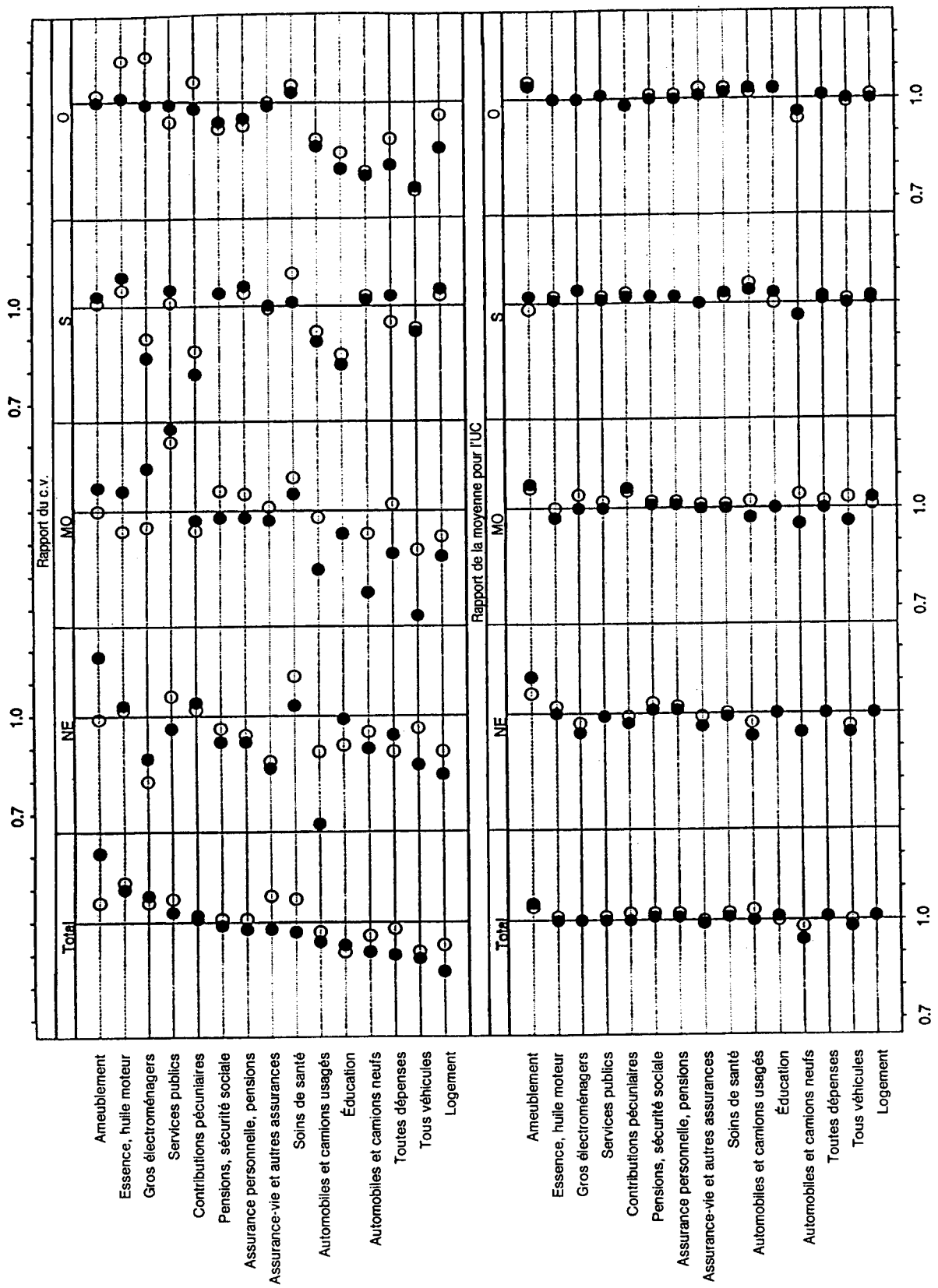

Figure 3. Rapports des c.v. et des moyennes obtenus par deux méthodes de pondération aux c.v. de la méthode PP, selon la région

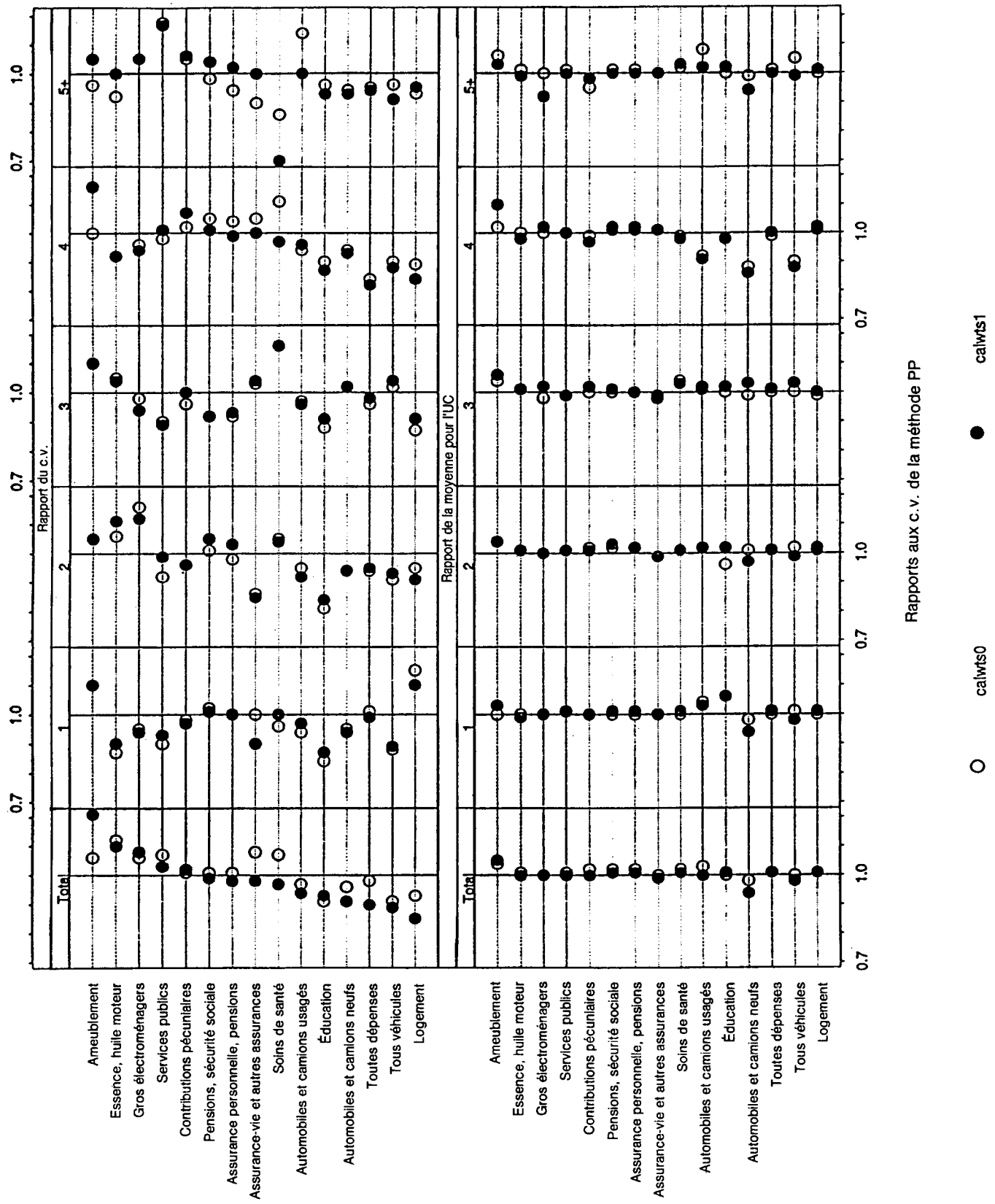

Figure 4. Rapports des c.v. et des moyennes obtenus par deux méthodes de pondération aux c.v. de la méthode PP, selon la taille de l'UC

les données pour une année complète provenant des enquêtes par entrevue et journal, pour 1990. Les résultats ont été similaires à ceux que nous venons de présenter, et nous avons adopté un ensemble final de 24 variables auxiliaires, basé sur le nombre de personne par âge, race, sexe, région, urbain × région, sur le nombre d'UC par type de ménage et sur l'utilisation d'une coordonnée à l'origine. On procède actuellement à la conversion des estimations pour l'enquête CE, afin d'utiliser la régression restreinte pour celle-ci.

4. CONCLUSION

L'objectif de cette étude était d'explorer des méthodes permettant de calculer les facteurs de pondération pour les ménages, facteurs qui ne dépendraient pas de la pondération accordée à un seul membre du ménage. Nous avons présenté différents types de facteurs de pondération basés sur l'estimation par régression, et nous avons évalué leurs mérites relatifs. L'estimation par régression incorpore les méthodes actuelles de stratification a posteriori de l'enquête, dans lesquelles la somme pondérée des personnes dans chaque strate a posteriori doit être égale aux chiffres correspondants fournis par un recensement indépendant. On y parvient en utilisant des variables auxiliaires qui sont incorporées dans le modèle de régression. Celui-ci produit également automatiquement, pour chaque ménage de l'échantillon, un facteur de pondération qui ne dépend d'aucun de ses membres.

Nous avons étudié huit types de facteurs de pondération, fournis par cinq modèles de régression différents. Afin d'éliminer les facteurs négatifs indésirables qui peuvent découler de toute procédure d'estimation par régression utilisant les moindres carrés ordinaires, nous avons adapté les estimateurs de régression restreinte au problème présent. La régression restreinte est justement assez souple pour restreindre les écarts possibles de chaque facteur de pondération final par rapport à son facteur de base, tout en respectant les propriétés mentionnées ci-dessus. On peut donc, notamment, imposer la contrainte que les facteurs soient positifs. Les facteurs de régression restreinte se calculent facilement au moyen de logiciels matriciels comme S-Plus™ ou SAS/ IML™.

La régression restreinte et, de façon plus générale, le calage restreint offrent un certain nombre d'attraits pour les enquêtes-ménages, comme celle que nous étudions ici, mais également pour les enquêtes portant sur d'autres types d'unités comme les hôpitaux, les écoles ou les établissements commerciaux pour lesquels on peut disposer d'un grand nombre de données auxiliaires. Compte tenu des données antérieures pour les variables cibles, on peut utiliser des procédures standard de construction de modèles afin de choisir les variables auxiliaires. Les propriétés de l'estimation par régression peuvent servir à choisir de façon optimale les variables prédictives, afin de réduire la redondance de l'information qui se trouve incorporée dans la procédure d'estimation pour l'enquête. C'est là un des plus grands avantages d'utiliser un estimateur ayant déjà fait l'objet de travaux nombreux et attestés. De bonnes variables prédictives peuvent comprendre des variables qualitatives, p. ex., âge, race, type d'hôpital (hôpital général, hôpital psychiatrique, *etc.*), type d'entreprise (usine, vente au détail, *etc.*) et qui peuvent souvent être utilisées dans la stratification ordinaire ou a posteriori. Les variables prédictives peuvent également être des variables quantitatives comme le revenu familial, les ventes annuelles, le nombre d'étudiants de différents niveaux ou le nombre de journées d'hospitalisation, pour n'en nommer que quelques-unes. Dans notre application, le fait d'inclure une coordonnée à l'origine a également permis de réduire de manière notable les erreurs-types des estimations pour l'enquête. La méthode de régression permet également d'incorporer facilement, dans l'estimation, des données à des niveaux différents. Dans l'enquête-ménage étudiée ici, nous avons inclus des variables auxiliaires pour les personnes et les ménages.

La grande souplesse de la régression offre aux statisticiens des options qu'ils ne pourraient avoir autrement. Si de nouvelles variables prédictives pertinentes deviennent disponibles, les logiciels permettant de calculer les estimations par régression peuvent facilement les incorporer; il suffit alors de modifier la matrice des variables auxiliaires et les vecteurs des contrôles de population. Les logiciels écrits pour effectuer seulement la stratification a posteriori ou l'estimation des quotients avec une seule variable auxiliaire, par exemple, pourraient nécessiter des modifications importantes afin que l'on puisse changer l'estimateur. Il va de soi que si l'estimateur est l'un des types moins généraux, de quotient ou de stratification a posteriori, le logiciel de régression le traitera comme un cas spécial. Aux États-Unis, on envisage d'effectuer une très vaste enquête permanente sur les ménages (Love, Alexander et Dalzell 1995), qui fournirait des estimations très précises pour de nombreuses caractéristiques que l'on pourrait utiliser comme totaux de contrôle dans des enquêtes plus petites. La méthode de régression restreinte permettrait à l'enquête CE d'incorporer sans heurt ces nouvelles données dans les estimations, si elles étaient disponibles.

REMERCIEMENTS

Les opinions exprimées dans cet article sont celles des auteurs et ne représentent pas la politique du Bureau of Labour Statistic. Nous remercions les examinateurs et les rédacteurs pour leurs observations utiles.

BIBLIOGRAPHIE

ALEXANDER, C.H. (1987). Une classe de méthodes utilisant des chiffres de population dans la pondération des ménages. *Techniques d'enquête*, 13, 193-209.

BARDSLEY, P., et CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.

BETHLEHEM, J.G., et KELLER, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.

CASADY, R.J., et VALLIANT, R. (1993). Propriétés conditionnelles des estimateurs de stratification a posteriori selon la théorie normale. *Techniques d'enquête*, 19, 193-203.

CLEVELAND, W.S. (1993). *Visualizing Data*. Summit, New Jersey: Hobart Press.

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.

ESTEVAO, V., HIDIROGLOU, M.A., et SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimating system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.

FULLER, W.A., LOUGHIN, M.M., et BAKER, H.D. (1993). Regression weighting for the 1987-1988, Nationwide Food Consumption Survey. Rapport soumis non-publié à United States Department of Agriculture.

LEMAÎTRE, G., et DUFOUR, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 211-220.

LOVE, S., ALEXANDER, C.H., et DALZELL, D. (1995). Constructing a major survey – operational plans and issues of continuous measurement. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. À paraître.

McCARTHY, P.J. (1969). Pseudo-replication: half-samples. *Revue de l'Institut International de Statistique*, 37, 239-264.

RAO, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

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

STUKEL, D.M., et BOYER, R. (1992). Calibration estimation: an application to the Canadian Labor Force Survey. Documents de travail, SSMD-92-009E, Ottawa: Statistique Canada.

ZIESCHANG, K.D. (1990). Sample weighting methods and estimation methods in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 85, 986-1001.