Techniques d'enquête, décembre 1987
Vol. 13, n° 2, pp. 211–220
Statistique Canada

Une méthode intégrée de pondération des personnes et des familles

G. LEMAÎTRE et J. DUFOUR[1]

RÉSUMÉ

Les enquêtes sur les ménages utilisent habituellement des méthodes d'estimation distinctes pour les caractéristiques des personnes et celles des familles. Les auteurs proposent dans cet article une méthode intégrée de pondération et définissent à cette fin un estimateur par les moindres carrés. Ils montrent que cet estimateur est sans biais à certaines conditions générales. Au moyen de données de l'enquête sur la population active du Canada, ils calculent les variances de cet estimateur et montrent qu'elles se comparent avantageusement aux variances calculées pour les méthodes usuelles.

MOTS CLÉS: Estimation pour les familles; pondération des familles; pondération par les moindres carrés.

1. INTRODUCTION

De nombreuses enquêtes sur les ménages comportent souvent une étape de stratification a posteriori par laquelle on contraint les estimations de population de l'enquête, normalement classées selon l'âge et le sexe, à respecter des totaux supplémentaires tirés de sources démographiques. Pour faciliter la totalisation, on attribue habituellement à chaque répondant un poids qui est égal au produit de l'inverse du taux de sondage par un facteur de compensation de la non-réponse par un facteur de correction âge et le sexe sous forme de quotient. On calcule ensuite l'estimation pour une caractéristique donnée en additionnant les poids attribués à tous les répondants de l'échantillon qui présentent cette caractéristique. En règle générale, les membres d'un même ménage n'ont pas le même poids. Lorsqu'il s'agit d'estimer les caractéristiques de personnes, cela ne pose pas de problème particulier; en revanche, lorsqu'il s'agit de produire des estimations relatives aux ménages ou aux familles, il n'est pas évident lequel de ces poids convient d'utiliser dans les circonstances, s'il y a lieu effectivement d'en choisir un.

Pour estimer les caractéristiques de familles, on pourrait recourir à l'estimation par le quotient en se servant de données supplémentaires sur les familles et les personnes. Cependant, on dispose rarement de données supplémentaires fiables et récentes sur les familles. À cause d'événements tels qu'une naissance, un décès, un mariage, un divorce ou le départ d'un membre du ménage ou l'arrivée d'un nouveau membre, des caractéristiques comme la taille de la famille varient d'un recensement à l'autre de façon moins prévisible qu'une caractéristique telle que l'âge. Les dossiers administratifs qui nous renseignent sur les variations postcensitaires de la population (c'est-à-dire, registres des naissances, des décès et des migrations) ne disent rien sur les variations touchant les ménages. Le registre des naissances, par exemple, n'indique pas la taille des familles où un enfant est né. Les dossiers fiscaux peuvent combler en partie cette lacune (voir Auger 1987) mais ils ne s'étendent pas à toute la population et ne renferment pas des données suffisamment récentes pour produire des estimations courantes. En l'absence de données supplémentaires sur les familles, on s'est mis à utiliser les poids attribués aux personnes pour estimer les caractéristiques de familles. Pour diverses raisons, cette méthode n'est pas la solution idéale. Dans cet article, nous proposons une méthode d'estimation qui permet de calculer un poids unique pour le ménage, lequel poids peut servir aussi bien à l'estimation pour les personnes qu'à l'estimation pour les familles.

[1] G. Lemaître et J. Dufour, Division des méthodes d'enquêtes sociales, Statistique Canada, 4-ième étage, Immeuble Jean Talon, Parc Tunney, Ottawa (Ontario) K1A 0T6.

On a déjà proposé des méthodes qui visaient à déterminer un poids unique pour chaque ménage; on insistait alors sur l'utilisation de données supplémentaires sur les personnes pour améliorer les estimations relatives aux familles. Oh et Scheuren (1978) ont proposé une technique d'estimation itérative "multidimensionnelle" fondée sur la méthode du quotient, qui consiste à corriger successivement par la méthode du quotient les estimations de population des strates formées a posteriori en utilisant les facteurs de correction calculés pour chaque strate, puis à répéter l'opération jusqu'à ce qu'il y ait convergence. Les corrections calculées à chaque étape sont appliquées aux ménages qui comptent des membres dans la strate (formée a posteriori) qui fait l'objet de la correction. Zieschang (1986) a proposé une méthode des moindres carrés généralisés (MCG) selon laquelle on minimise la somme des carrés pondérés des corrections apportées aux poids du plan de sondage (ou poids initiaux) en respectant une série de contraintes linéaires. Alexander (1987) analyse plusieurs méthodes de pondération fondées sur la minimisation d'une fonction de distance, sujette à certaines contraintes, y compris la méthode MCG, et les évalue par rapport au sous-dénombrement dans les enquêtes. Bien que les méthodes ci-dessus aient été proposées initialement dans le but d'améliorer les estimations relatives aux familles, les poids découlant des divers estimateurs peuvent très bien servir à estimer les caractéristiques de personnes. Cet article préconise l'application d'une méthode d'estimation adaptée aussi bien aux personnes qu'aux familles. Dans la section 2, nous analysons les faiblesses des méthodes d'estimation actuellement en usage pour ce qui a trait aux caractéristiques des personnes et des familles. La section suivante sert à définir un estimateur fondé sur un modèle et inspiré d'une méthode de pondération généralisée mise de l'avant par Bethléhem et Keller (1987). La section 4 présente des résultats empiriques tirés de l'enquête sur la population active du Canada. Enfin, la section 5 expose des projets de recherche supplémentaires.

2. MÉTHODES D'ESTIMATION COURANTES

La plupart des enquêtes sur les ménages ont pour objectif principal de produire des estimations pour les caractéristiques de personnes, notamment les caractéristiques de l'activité. Si l'unité d'échantillonnage ultime de ces enquêtes est le ménage, c'est surtout pour des raisons d'économie et de commodité. Bien que les premières étapes de la pondération (compensation de la non-réponse, correction en fonction des régions rurales et urbaines, etc.) tiennent compte du ménage en tant qu'unité, ce fait n'est pas reconnu dans l'étape finale (c'est-à-dire que l'on ne tient pas compte du fait que les membre d'un ménage sont échantillonnés intégralement). De façon plus particulière, les biais de couverture qui pourraient être liés à l'unité échantillonnée ne sont pas directement pris en considération ou compensés dans l'estimation. On suppose donc que le sous-dénombrement est "ignorable" au sens de Rubin (1976); autrement dit, l'étape de l'estimation traite de façon identique toutes les personnes comprises dans une strate formée a posteriori selon les critères d'âge et de sexe, peu importe que ces personnes vivent seules ou qu'elles appartiennent à un ménage formé de plusieurs personnes. Toutefois, une étude sur la non-réponse dans l'enquête sur la population active (Paul et Lawes 1983) a démontré que les ménages de faible taille, particulièrement les ménages sans enfant, tendent à être sous-représentés dans l'échantillon. Bien qu'il n'existe pas d'étude comparable pour les ménages oubliés dans l'enquête sur la population active, des études portant sur le sous-dénombrement des ménages privés dans le recensement ont montré que les ménages non-dénombrés sont effectivement de plus petite taille en moyenne que les ménages dénombrés (Gosselin et Théroux 1980). Une méthode qui suppose que les ménages sont absents de façon aléatoire peut introduire un biais dans les estimations de la population active pour les personnes, surtout si la répartition des membres des ménages de faible taille par rapport aux caractéristiques de l'activité est différente de celle des membres des ménages de taille plus élevée, toutes choses étant égales par ailleurs. Intuitivement, une méthode d'estimation qui tiendrait compte (ne serait-ce qu'indirectement) du fait que les ménages de faible taille sont

moins exposés à la non-réponse et au sous-dénombrement que les ménages de taille plus élevée pourrait corriger en partie cette faiblesse de l'échantillon.

Faute de pouvoir incorporer des données supplémentaires sur les ménages ou les familles dans une méthode de pondération appropriée afin d'attribuer un poids bien défini aux familles, beaucoup de méthodes actuellement en usage attribuent à la famille le poids de la "personne principale" de cette famille. Dans l'enquête sur la population active du Canada, la personne principale est le conjoint de sexe féminin, si elle est présente, sinon c'est le chef de ménage. Comme ces méthodes supposent que les ménages non-dénombrés sont absents de façon aléatoire, les estimations produites à l'aide du poids de la personne principale tendent à surestimer les familles de taille plus élevée et à sous-estimer les personnes seules. En outre, on peut estimer de nombreuses caractéristiques (par exemple, population, revenu) à l'aide du poids attribué aux personnes ou du poids attribué aux familles et en règle générale, les deux séries d'estimations diffèreront entre elles, parfois de façon substantielle. Il est vrai que même dans des conditions d'échantillonnage et d'interview idéales, avec un taux de non-réponse ou de sous-dénombrement uniforme, il y aura toujours des écarts entre les estimations fondées sur les familles et celles fondées sur les personnes pour une même caractéristique. Toutefois, si l'échantillon est suffisamment grand, les différences devraient être faibles. Dans les conditions réelles d'enquête, qui ne sont pas des conditions idéales, les différences sont parfois trop fortes pour que l'on puisse les expliquer simplement par la variabilité d'échantillonnage. En appliquant une méthode d'estimation qui attribue un poids unique au ménage, lequel poids concorde avec les chiffres de population supplémentaires lorsqu'il est utilisé comme poids de personne, nous nous trouvons à résoudre le problème des deux systèmes d'estimation parallèles. C'est ce à quoi nous nous attachons dans la section suivante en définissant un estimateur particulier.

3. ESTIMATEUR PROPOSÉ

Nous allons commencer par définir une méthode de pondération généralisée fondée sur des modèles linéaires de Bethlehem et Keller (1987) et, comme eux, nous allons l'appliquer tout d'abord à l'estimation fondée sur les personnes. Nous modifions ensuite la méthode de manière qu'elle produise des poids de ménages qui conviennent à l'estimation des caractéristiques de personnes. Nous allons ici nous inspirer largement de l'article de Bethlehem et Keller.

Supposons une population cible d'enquête formée de N unités, un N-vecteur Y contenant les valeurs d'une variable cible et une matrice X de dimension $N \times p$ contenant les variables auxiliaires définies pour chaque unité de la population cible. On suppose que les chiffres de population pour chaque varaible auxiliaire sont connus et on les désigne collectivement par le p-vecteur x. Dans notre modèle, x sera constitué des totaux de groupes d'âge-sexe. S'il y a corélation entre les variables auxiliaires et la variable cible, les valeurs de $E = Y - XB$ varieront moins que les valeurs de la variable cible Y pour un p-vecteur B approprié. En appliquant les moindres carrés ordinaires à toutes les unités de la population cible, on obtient

$$B = (X'X)^{-1}X'Y, \tag{3.1}$$

pourvu que X soit une matrice à rang complet. Une estimation de B pour l'échantillon est définie par

$$\hat{B} = (X'\Pi^{-1}TX)^{-1}X'\Pi^{-1}TY, \tag{3.2}$$

où T est une matrice diagonale dont le i-ième élément est égal à 1 si la i-ième unité de la population est incluse dans l'échantillon, et est égal à 0 dans le cas contraire, et $E(T) = \Pi$.

Il est possible de montrer que $\hat{B}$ sera approximativement non biaisé pour de grands échantillons. Or, le paramètre d'intérêt n'est pas B mais le chiffre de population y. Si nous posons $\hat{y} = \hat{B}'x$, $\hat{y}$ sera un estimateur approximativement non biaisé de y à la condition que $B'x = y$, ou que la somme des résidus pour le modèle de population $Y = XB + E$ soit égale à 0. Cette condition sera respectée si le N-vecteur formé de chiffres 1 est dans l'espace délimité par les colonnes de X et, en particulier, si les variables auxiliaires X renferment un ensemble exhaustif de variables indicatrices qui s'excluent mutuellement (pour les groupes d'âge-sexe par exemple).

Si nous écrivons $\hat{y} = \hat{B}'x = Y'\Pi^{-1}TX(X'\Pi^{-1}TX)^{-1}x$, nous constatons que l'estimateur définit implicitement un N-vecteur de poids donné par

$$W = \Pi^{-1}TX(X'\Pi^{-1}TX)^{-1}x,$$

qui ne dépend pas de la variable cible faisant l'objet de l'estimation. Si nous utilisons les poids pour produire des estimations pour les variables auxiliaires, nous obtenons $X'W = x$. Les poids donnent en effet les totaux de population espérés. De plus, si X est constituée entièrement d'un ensemble exhaustif de variables indicatrices qui s'excluent mutuellement, l'estimateur de régression y équivaudra à l'estimateur régulier de la stratification a posteriori. Pour plus de renseignements, voir Bethlehem et Keller (1987).

Il peut être utile de souligner que selon cette méthode, le poids d'une personne quelconque échantillonnée i peut être défini en règle générale de la façon suivante:

$$W_i = \sum_j \frac{x_{ij}b_j}{\pi_i}, \tag{3.3}$$

où $(b_1, \ldots, b_p) = (X'\Pi^{-1}TX)^{-1}x$ et π_i est la probabilité de sélection de la personne i. Cela donne à penser que l'on peut modifier la méthode d'estimation décrite ci-dessus de manière à obtenir les poids voulus en définissant les variables auxiliaires de la même façon pour tous les membres du ménage. Une manière simple de le faire est de définir des variables auxiliaires pour le ménage en remplaçant, par exemple, les variables correspondantes pour les personnes par la moyenne du ménage. De façon plus formelle, soit Z une matrice $N \times p$ définie pour une personne i $(i = 1, \ldots, N)$ appartenant au ménage h $(h = 1, \ldots, H)$ par

$$Z_{ij} = \frac{U_{hj}}{n_h},$$

où U_{hj} est le total pour la caractéristique j dans le ménage h, c'est-à-dire $U_{hj} = \Sigma_k X_{kj}$, où la sommation porte sur tous les membres k du ménage h, n_h = taille du ménage h, et $\Sigma_h n_h = N$. Définissons Y comme un N-vecteur de valeurs d'une variable cible arbitraire définie pour les *personnes*. Comme pour l'estimation relative aux personnes, nous utilisons le modèle de population $Y = ZC + E$ et appliquons les moindres carrés aux données de l'échantillon pour obtenir une estimation

$$\hat{C} = (Z'\Pi^{-1}TZ)^{-1}Z'\Pi^{-1}TY. \tag{3.4}$$

Nous posons $\hat{y} = \hat{C}'x$, où x est encore le vecteur des chiffres de population pour les variables auxiliaires. $\hat{y}$ sera un estimateur approximativement non biaisé de y pourvu que le N-vecteur formé de chiffres 1 se trouve dans l'espace délimité par les colonnes de Z. Comme pour (3.3), le poids d'une personne quelconque échantillonnée dans le ménage h sera défini par

$$W_h = \sum_j \frac{U_{hj}c_j}{\pi_h n_h}. \tag{3.5}$$

Comme tous les membres du ménage se rattachent au même vecteur ligne de Z et que la probabilité de sélection du premier ordre est la même pour tous, ils auront tous le même poids. De plus, lorsqu'on se sert du oids du ménage pour els personnes, on obtient des résultats compatibles avec les chiffres de population supplémentaires. Bien que cette méthode puisse produire des poids négatifs (si la valeur de quelques-uns des c_j's est inférieure à zéro), les ménage dont le poids est modifié le plus par cette méthode sont en règle générale des ménages dont la composition est peu commune et que l'on retrouve rarement dans l'échantillon et dans la population en général, du moins en ce qui concerne les échantillons qui ne sont pas exposés à un taux de non-réponse ou de sous-dénombrement élevé. La méthode proposée a servi récemment à pondérer des données de l'enquête sur la population active portant sur une période de 24 mois; à cette occasion, un seul ménage a reçu un poids négatif, faible par surcroît. Les poids négatifs font problème parce qu'il est difficile de leur attacher le sens que l'on attribue normalement aux poids, c'est-à-dire nombre de personnes ou de ménages dans la population en général représentés par une personne ou un ménage particuliers échantillonnés. Cependant, selon la formule décrite ci-dessus, les poids finals ne sont définis qu'implicitement et on pourrait de fait considérer que ces poids ne sont qu'un moyen commode de produire des estimations. En pratique, il est peu probable qu'une estimation significative de la valeur d'une caractéristique d'intérêt devienne négative sous l'effet de quelques poids négatifs. En revanche, c'est tout autre chose de vouloir faire comprendre à un utilisateur perplexe la notion de poids négatif.

La variance de l'estimateur $\hat{y} = \hat{C}'x$ défini dans le présent article peut être calculée au moyen des méthodes décrites dans Fuller (1975). De plus, il est possible de montrer que cet estimateur équivaut aux estimateurs de la méthode MCG proposée par Zieschang (1986) et Alexander (1987) lorsque l'espace délimité par les variables auxiliaires Z renferme un vecteur de chiffres 1. Wright (1983) décrit d'autres propriétés de ce genre d'estimateur.

4. RÉSULTATS EMPIRIQUES

L'enquête sur la population active du Canada est une enquête mensuelle à échantillon avec renouvellement qui s'adresse à quelque 48,000 ménages répartis dans tout le pays (voir Platek et Singh 1976 et Singh, Drew et Choudhry 1984). Une fois échantillonnés, les ménages demeurent dans l'échantillon six mois consécutifs avant d'être remplacés. Les strates géographiques primaires sont les dix provinces. La taille des échantillons peut varier de 1,500 ménages, à l'Île-du-Prince-Édouard (la plus petite province), à environ 9,000 en Ontario (province la plus populeuse). L'enquête permet de recueillir des données sur la situation des répondants vis-à-vis de l'activité pendant une semaine de référence donnée à chaque mois et aboutit à la publication de toute une série d'estimations relatives au marché du travail dans le pays.

Les données d'une des enquêtes mensuelles ont servi à faire une évaluation préliminaire de l'estimateur proposé. Nous avons choisi à cette fin l'enquête de mai 1981 pour pouvoir comparer les résultats à ceux du recensement de 1981 tenu à peu près à cette date. Bien que nous ayons utilisé jusqu'ici les termes "ménage" et "famille" indistinctement, l'utilisateur s'intéresse plus souvent aux estimations touchant la "famille économique", laquelle est constituée de tous les membres d'un ménage qui sont apparentés par le sang, par alliance ou par adoption. Or, l'unité échantillonnée (en l'occurrence, le ménage) se prête mieux théoriquement à la pondération. Néanmoins, les résultats empiriques que nous exposons ici reposent sur des estimations qui ont trait aux familles économiques. Dans cette évaluation, nous nous sommes intéressés aussi bien aux caractéristiques des personnes (situation vis-à-vis de l'activité) qu'à celles des familles (nombre de familles économiques et de personnes seules). La pondération par les moindres carrés a porté sur deux séries de groupes d'âge-sexe formés par intervalle de cinq ans, les personnes de 70 ans et plus étant groupées selon le sexe. Les enfants de 0 à 14 ans ont été exclus de la première série de groupes (24 au total) pour que l'on puisse comparer l'estimateur proposé à un estimateur régulier de stratification a posteriori

fondé sur les personnes en utilisant la même information supplémentaire. La seconde série comprenait aussi les enfants répartis en six groupes d'âge-sexe et a servi uniquement à la pondération par les moindres carrés puisque la pondération des enfants par la stratification a posteriori régulière n'a aucun effet sur la pondération des personnes de 15 ans et plus.

Bien que tous les estimateurs étudiés soient approximativement non biaisés pour les estimations des caractéristiques de personnes, les hypothèses concernant la nature du sous-dénombrement et de la non-réponse varient de l'un à l'autre (la méthode de compensation de la non-réponse utilisée dans l'enquête sur la population active suppose que les ménages non répondants sont absents de façon aléatoire à l'intérieur d'une région géographique). L'estimateur par stratification a posteriori suppose implicitement que les différences de taux de non-réponse ou de sous-dénombrement dépendent uniquement de l'âge et du sexe et que, par conséquent, elles peuvent être éliminées au moyen d'une estimation fondée sur les personnes et appuyée de l'information supplémentaire portant sur ces caractéristiques. Dans la pondération par les moindres carrés, le poids d'une personne dépendra de la composition âge-sexe du ménage (sans enfant dans un cas et avec des enfants dans l'autre). Ainsi, toutes choses étant égales par ailleurs, la correction appliquée au poids initial d'une personne appartenant à un groupe d'âge-sexe exosé à un fort taux de sous-dénombrement devrait être plus élevée si cette personne demeure seule que si elle demeure avec des personnes qui appartiennent à des groupes d'âge-sexe bien représentés par l'échantillon.

Tableau 1

Nombre de personnes occupés et en chômage, nombre de familles économiques et de personnes seules, enquête sur la population active, mai 1981 (en milliers)

	Estimateur[a]	Occupées	En chômage	Familles économiques	Personnes seules
Canada	A	11,094	850	6,424	2,432
	B	11,090	850	6,446	2,442
	C	11,120	851	6,410	2,495
Région de l'Atlantique	A	819	102	563	156
	B	819	102	570	154
	C	821	102	569	156
Québec	A	2,725	304	1,723	587
	B	2,724	304	1,725	596
	C	2,735	305	1,714	614
Ontario	A	4,198	274	2,325	863
	B	4,200	273	2,325	861
	C	4,211	273	2,310	881
Région des Prairies	A	2,074	83	1,078	506
	B	2,072	84	1,089	510
	C	2,074	83	1,085	517
Colombie-Britannique	A	1,277	88	735	319
	B	1,276	88	738	321
	C	1,280	88	734	327

[a] A = stratification a posteriori/personne principale, B = moindres carrés (enfants exclus de la pondération) et C = moindres carrés (enfants inclus dans la pondération).

Comme il existe des chiffres de population supplémentaires selon l'âge et le sexe pour chaque province, nous avons établi des estimations pour chacune des provinces. Cependant, dans les tableaux qui suivent, les provinces de moindre importance sont fondues en deux groupes.

Tableau 2

Répartition des écarts en pourcentage entre les poids finals et les poids initiaux, estimateur par stratification a posteriori et estimateur par les moindres carrés, enquête sur la population active, mai 1981

Écart en pourcentage	Pourcentage de l'échantillon		
	Stratification a posteriori	Moindres carrés	Moindres carrés (avec enfants)
> −30%	0.0	0.1	0.2
−30 à −20%	0.0	0.5	0.9
−20 à −10%	0.6	3.0	5.3
−10 à 0%	23.9	20.4	27.1
0 à 10%	53.9	44.6	37.3
10 à 20%	20.6	26.3	21.6
20 à 30%	0.6	4.4	6.2
30 à 40%	0.1	0.4	0.9
40 à 50%	0.0	0.0	0.2
< 50%	0.0	0.0	0.2

Note: La taille de l'échantillon est N = 159014.

Tableau 3

Efficacité estimée de l'estimateur par les moindres carrées par rapport à l'estimateur par stratification a posteriori: nombre de personnes occupées et en chômage, nombre de familles économiques et de personnes seules, enquête sur la population active, mai 1981

	Estimateur[a]	Personnes occupées	Personnes en chômage	Familles économiques	Personnes seules
Canada	B	1.044	0.999	1.565	1.038
	C	1.066	0.999	1.616	1.036
Région de l'Atlantique	B	1.110	0.977	1.266	0.998
	C	1.193	0.992	1.567	1.070
Québec	B	1.059	1.005	1.553	1.020
	C	1.063	0.992	1.582	0.992
Ontario	B	1.028	1.011	1.825	1.064
	C	1.059	1.010	1.828	1.037
Région des Prairies	B	1.001	1.009	1.205	1.009
	C	1.072	1.066	1.420	1.134
Colombie-Britannique	B	1.038	0.964	1.248	1.048
	C	1.053	0.978	1.203	1.045

[a] B = moindres carrés (enfants exclus de la pondération) et C = moindres carrés (enfants inclus dans la pondération).

De façon générale, les trois estimateurs, et plus particulièrement A et B, ne produisent pas des estimations très différentes les unes des autres. Le fait d'inclure les enfants dans la pondération semble se traduire par des estimations légèrement plus élevées pour les personnes occupées et les personnes seules et des estimations légèrement moins élevées pour les familles économiques au niveau national et dans les provinces plus importantes (comparer ces résultats à ceux de Scheuren et coll. 1981). Ces chiffres sont conformes aux prévisions bien qu'il subsiste des écarts notables par rapport aux résultats du recensement, qui indiquent (arrondi au millier près) 6,369,000 familles économiques et 2,583,000 personnes seules au pays. Nous pouvons en conclure que même si l'estimateur par les moindres carrés produit des chiffres qui nous rapprochent plus de la réalité (lorsqu'on tient compte des enfants), il nous faudra de l'information supplémentaire précise et récente pour effacer le biais résiduel.

L'efficacité de l'estimateur par les moindres carrés n'est pas tout à fait évidente. Certes, si nous devions faire une prévision en nous fondant sur les résultats ci-dessus, nous serions portés à dire que l'estimateur par les moindres carrés sera aussi efficace que l'estimateur par stratification a posteriori en constatant la similitude des estimations produites par les deux genres d'estimateur. Par ailleurs, on devrait s'attendre à des gains d'efficacité pour les estimations relatives aux familles économiques parce que l'estimateur par les moindres carrés fait intervenir les chiffres de population supplémentaires dans le calcul du poids des ménages. Cependant, on ne peut attribuer un poids unique au ménage sans une redistribution des poids.

Comme l'indique le tableau 2, la dispersion est un peu plus forte pour les poids calculés par les moindres carrés que pour ceux calculés par les méthodes ordinaires de stratification a posteriori. Le fait de tenir compte des enfants crée une dispersion encore plus forte. Le degré de dispersion des poids reflète essentiellement la disparité qui existe entre la composition de l'échantillon et celle de la population en général au point de vue de l'âge, du sexe et de la taille des ménages. Comme le fait de vouloir calculer un poids unique pour tout le ménage ajoute une contrainte à la méthode d'estimation, on pourrait s'attendre que les variances en subissent quelque peu les conséquences, surtout si aucune autre information supplémentaire n'est utilisée pour l'estimation. Toutefois, la réalité est quelque peu différente.

Les variances de l'estimateur par stratification a posteriori ont été estimées au moyen de la méthode de Keyfitz (1957), où les échantillons répétés étaient des UPÉ (unités primaires d'échantillonnage) ou des UPÉ regroupées. Les variances de l'estimateur par les moindres carrés ont été estimées à l'aide de la méthode décrite dans Fuller (1975). Pour obtenir une meilleure comparabilité, on a calculé les variances pour plusieurs caractéristiques estimées par stratification a posteriori au moyen de la méthode de Fuller et on les a comparées aux variances obtenues par la méthode de Keyfitz. Les deux séries d'estimations étaient très comparables dans tous les cas (1 à 2 pour cent d'écart).

Le tableau 3 donne l'efficacité estimée de l'estimateur par les moindres carrés par rapport à l'estimateur par stratification a posteriori pour les caractéristiques étudiées dans le tableau 1. Les gains d'efficacité pour les estimations relatives aux familles économiques sont substantiels. Les estimations relatives aux personnes occupées et aux personnes seules semblent aussi être légèrement plus efficaces; toutefois, les réductions de la variance pour ces naractéristiques sont faibles, sauf pour les personnes occupées dans la région de l'Atlantique, surtout quand les enfants figurent dans la pondération. Il est intéressant de constater que la taille moyenne d'un ménage dans la région Atlantique est supérieure à ce qu'elle est ailleurs au Canada, bien que l'influence que cela pourrait avoir sur les estimations de personnes occupées n'est pas tout à fait claire. Pour ce qui a trait aux personnes en chômage, la méthodes des moindres carrés ne change essentiellement rien aux variances. Il est permis de croire que ces résultats seront observés de façon générale, c'est-à-dire pour des caractéristiques quelconques. Bien que le critère du poids unique par ménage soit restrictif pour les estimations des caractéristiques de personnes, l'estimateur par les moindres carrés semble résoudre cette difficulté grâce aux variables "explicatives" additionnelles du modèle linéaire, c'est-à-dire les moyennes de toutes les variables auxiliaires des ménages. Les résultats préliminaires

rapportés ci-dessus donnent à penser qu'il serait possible d'intégrer les deux genres d'estimation (personnes et familles) sans que cela ne réduise pour autant l'efficacité des estimations relatives aux personnes.

5. PROJETS DE RECHERCHE

On procède actuellement à une étude empirique plus poussée des propriétés de l'estimateur par les moindres carrés, où l'on s'intéresse particulièrement à l'évolution chronologique des estimations et à leur efficacité, pour un plus grand nombre de caractéristiques, par rapport aux estimations produites par la méthode itérative du quotient utilisée actuellement dans l'enquête sur la population active. Les résultats ci-dessus donnent à penser que la composition âge-sexe d'un ménage a un "pouvoir explicatif" au moins aussi grand que celui du groupe d'âge-sexe même, du moins en ce qui concerne certaines caractéristiques de personnes. Il sera intéressant de voir si les efficacités relatives seront aussi satisfaisantes pour des caractéristiques qui ont une plus forte corrélation avec l'âge et le sexe. Par ailleurs, même si en pratique les poids négatifs sont rares, il faudra penser à élaborer une méthode qui permettra de traiter ces poids le cas échéant. On pourrait, par exemple, les considérer comme des valeurs aberrantes ou peut être les prévenir en limitant la valeur des modifications de poids (Zieschang 1987). Enfin, il serait utile de définir explicitement le modèle de sous-dénombrement sur lequel repose l'estimateur par les moindres carrés pour que l'on puisse faire une évaluation du modèle proprement dit.

REMERCIEMENTS

L'auteur tient à remercier F. Scheuren pour ses commentaires et suggestions lors de la rédaction de cet article.

BIBLIOGRAPHIE

ALEXANDER, C.H. (1987). Une classe de méthodes utilisant des chiffres de population dans la pondération des ménages. *Techniques d'enquête,* 13, 193-210.

AUGER, E. (1987). Family data from the canadian personal income tax file. Dans *Statistics of Income and Related Administrative Record Research: 1986-1987,* (éd. W. Alvey et B. Kilss), Washington, D.C.: Internal Revenue Service, 177-184.

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

FULLER, W.A. (1975). Regression analysis for sample surveys. *Sankhyā,* C37, 117-132.

GOSSELIN, J.-F., et THÉROUX, G. (1980). Recensement du Canada de 1976 Qualité des données Série I; Sources d'erreurs - Couverture. N° 99-840 au répertoire, Statistique Canada.

KEYFITZ, N. (1957). Estimates of sampling variance where two units are selected from each stratum, *Journal of the American Statistical Association,* 52, 503-510.

OH, H.L., et SCHEUREN, F. (1978). Multivariate raking ratio estimation in the 1973 exact match study. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 716-722.

PAUL, E.C., et LAWES, M. (1982). Caractéristiques des ménages répondants et non répondants dans l'enquête sur la population active du Canada. *Techniques d'enquête,* 8, 48-85.

PLATEK, R., et SINGH, M.P. Méthodologie de l'enquête sur la population active du Canada, N° 71-526 au répertoire, Statistique Canada.

RUBIN, D.B. (1976). Inference on missing data. *Biometrika,* 63, 581-592.

SINGH, M.P., DREW, J.D., et CHOUDHRY G.H. (1984). Remaniement de l'enquête sur la population active au Canada à partir des résultats du recensement de 1981. *Techniques d'enquête,* 10, 139-154.

SCHEUREN, F., OH, H.L., VOGEL, L., et YUSKAVAGE, R. (1981). studies from interagency data linkages, report No. 10: methods of estimation for the 1973 exact match study. U.S. Department of Health and Human Services, Social Security Administration, SSA Publication No. 13-11750.

WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association,* 78, 879-884.

ZIESCHANG, K.D. (1986). A generalized least quares weighting system for the Consumer Expenditure Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

ZIESCHANG, K.D. (1987). Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. Document non publié, U.S. Bureau of Labour Statistics.