No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2006 : Enjeux méthodologiques reliés à la mesure de la santé des populations

2006

Statistics Canada Statistique Canada

Canada

Recueil du Symposion 2006 de Statistique Canada
Enjeux méthodologiques reliés à la mesure de la santé des populations

Une étude de cas du recours à l'estimation assistée par un modèle pour intégrer des données d'enquête et des données administratives

Robert E. Fay[1]

Résumé

Le présent article décrit les travaux de recherche en vue d'intégrer l'estimation assistée par un modèle dans l'American Community Survey (ACS), grande enquête permanente destinée à remplacer le questionnaire de recensement détaillé dans les recensements décennaux des États-Unis. L'application proposée intègre l'information provenant des dossiers administratifs dans l'estimation d'après les données de l'ACS. L'approche de l'estimation assistée par un modèle limite l'utilisation des dossiers administratifs aux ajustements des poids de sondage, tout en retenant les données sur les caractéristiques recueillies directement auprès des participants à l'ASC. Bien que l'ACS soit une enquête générale qui n'est pas spécialement liée à la santé, la présente étude de cas pourrait permettre de proposer des applications éventuelles dans le domaine de la statistique de la santé.

MOTS-CLÉS : American Community Survey; estimation sur petits domaines; estimation par calage.

1. Introduction

L'année 2005 a marqué le début de la phase de production complète de l'American Community Survey (ACS) réalisée aux États-Unis. L'enquête remplace le questionnaire détaillé du recensement décennal des États-Unis qui est l'équivalent du questionnaire détaillé du recensement du Canada. L'ACS est une enquête permanente sur la population des États-Unis, réalisée chaque mois auprès d'un échantillon d'environ 1 sur 480 unités de logement. Le contenu est semblable à celui recueilli au moyen du questionnaire détaillé de 2000. Cumulés sur une année, ces échantillons mensuels donnent un échantillon désigné de 1 sur 40 qui produit environ trois millions d'unités de logement des États-Unis. L'ACS a déjà été utilisée pour produire les estimations de 2005 pour les États, les comtés et les localités dont la population est supérieure à 65 000 habitants. Cumulées sur cinq ans, les données de l'ACS pour la période allant de 2005 à 2009 représenteront un échantillon d'environ 1 sur 8 unités de logement des États-Unis et serviront de base pour le calcul des estimations au même niveau de détail géographique que celui du Recensement de 2000.

L'ACS a, sur le questionnaire détaillé décennal des États-Unis, l'avantage qualitatif de produire des données recueillies par des intervieweurs plus expérimentés et d'aboutir, en général, à des taux de réponse plus élevés. Néanmoins, en raison de contraintes budgétaires, l'échantillon désigné de 1 sur 8 sur une période de cinq ans représente un taux global inférieur à celui de 1 sur 6 du recensement décennal. En outre, dans le cas de l'ACS, un sous-échantillon des personnes n'ayant pas répondu initialement est sélectionné pour une visite sur place, de sorte que l'échantillon réalisé de l'ACS est plus petit que l'échantillon désigné de 1 sur 8. Par conséquent, le Census Bureau offre l'ACS à titre de solution de rechange raisonnée au questionnaire détaillé du recensement décennal, mais reconnaît que les estimations fondées sur les données de cette enquête présenteront une erreur d'échantillonnage plus grande que celle observée dans le cadre des recensements récents.

Les travaux de recherche sur l'ACS comprennent un essai national réalisé de 1999 à 2001 à un taux d'échantillonnage nettement plus faible que celui de la mise en œuvre en production. Toutefois, depuis 1999, une version de l'ACS a été réalisée dans 36 comtés d'essai (sélectionnés parmi les plus de 3 000 comtés des États-Unis) avec approximativement le même taux d'échantillonnage que pour la mise en œuvre complète. Les données provenant de ces comtés servent de fondement à l'évaluation des résultats des méthodes de l'ACS dans le cas des petites régions géographiques, y compris les méthodes d'estimation et les variances résultantes.

[1] U.S. Census Bureau, 4700 Silver Hill Road, Washington, DC U.S.A., 20233 (robert.e.fay.iii@census.gov).

Il y a quelques années, Paul Voss et ses collègues ont découvert que les erreurs d'échantillonnage dans les estimations de niveau infra-comté d'après les données de l'ACS dans les comtés d'essai étaient plus grandes que prévu, même s'il l'on tenait compte de l'effet de la taille d'échantillon. Leurs observations ont été reproduites par la suite (Starsinic, 2005; Fay, 2005a). Contrairement au questionnaire détaillé du Recensement de 2000, pour lequel on avait appliqué l'estimation par ratissage croisé dans des régions de pondération de niveau infra-comté relativement petites, l'estimation d'après les données de l'ACS était fondée sur des chiffres de population de contrôle à l'échelle du comté ou à un niveau géographique plus agrégé seulement. Cette observation, conjuguée au taux d'échantillonnage un peu plus faible pour l'ACS que pour le recensement, a souligné l'importance d'améliorer l'estimation d'après les données de l'ACS pour les régions de niveau infra-comté dans la mesure du possible. Le présent article est le quatrième d'une série décrivant les travaux de recherche en vue d'améliorer les estimations d'après les données de l'ACS (Fay 2005a, 2005b, 2006). La stratégie de base comprend l'utilisation de dossiers administratifs combinée à l'estimation assistée par un modèle.

Quelques raisons peuvent être avancées pour justifier l'inclusion de cette étude de cas dans une conférence sur la mesure de la santé des populations. Bien qu'il ne s'agisse pas principalement d'une enquête sur la santé (elle contient un petit ensemble de questions sur l'incapacité), l'ACS fournira des données démographiques et économiques géographiquement détaillées qui pourraient être utiles dans le cadre de nombreuses analyses épidémiologiques et autres de la santé de la population. Deuxièmement, les méthodes décrites ici pour combiner les données administratives et les données d'enquête grâce à l'estimation assistée par un modèle pourraient avoir d'autres applications, particulièrement dans les situations où des données administratives auxiliaires sur la santé sont disponibles.

En plus de son lien avec la recherche sur l'utilisation de données administratives, l'article se rattache à deux autres grands thèmes : l'importante contribution des chercheurs canadiens au développement d'estimateurs assistés par un modèle et à leur application à des problèmes à grande échelle, et le lien entre les présents travaux et l'estimation sur petits domaines.

La section qui suit donne une description plus détaillée de l'ACS et du problème d'estimation. La troisième section décrit l'utilisation d'une forme d'estimation assistée par un modèle, l'estimation par la régression généralisée (GREG), en particulier dans le cadre des recensements canadiens récents. La quatrième section résume les résultats encourageants des études empiriques réalisées jusqu'à présent. La discussion présentée à la dernière section laisse entendre comment les méthodes pourraient être appliquées à d'autres problèmes, y compris les enquêtes sur la santé.

2. Le problème d'estimation

2.1 L'American Community Survey

Comme nous l'avons déjà souligné, l'échantillon désigné mensuel de 1 sur 480 est cumulé en un échantillon désigné de 1 sur 40 sur une année, un échantillon de 3 sur 40 sur trois années, et un échantillon de 1 sur 8 sur cinq années. À l'heure actuelle, il est prévu de publier les estimations sur une période d'un an pour les États, les comtés et les localités dont la population est supérieure à 65 000, des estimations sur une période de trois ans pour les comtés et localités dont la population est supérieure à 20 000, et des estimations sur une période de cinq ans pour le même niveau de détail géographique que celui du Recensement de 2000. Pour ce dernier, des estimations ont été publiées pour les secteurs de recensement (région d'environ 4 000 habitants) et pour les groupes d'îlots (environ 1 500 habitants). Des estimations ont également été publiées pour les localités, dont un grand nombre étaient assez petites, et d'autres unités d'administration locale.

La publication des estimations sur un an d'après l'ACS pour 2005 a déjà débuté, mais les premières estimations sur trois ans pour la période de 2005 à 2007 paraîtront en 2008 et les premières estimations pour 2005 à 2009 paraîtront en 2010. Cependant, le plan d'échantillonnage complexe de l'ACS continue de poser des défis. Les travaux en cours comprennent l'élaboration de méthodes pour aider les utilisateurs à interpréter les estimations sur un, trois et

cinq ans. Les utilisateurs devront également établir des approches en vue de relier ces estimations par période à des statistiques provenant d'autres sources, comme les données sur la santé.

2.2 Variances des estimations au niveau infra-comté

Parmi les 36 comtés d'essai sélectionnés pour 1999 à 2001, 34 ont été échantillonnés au taux de 3 % (cinq grands comtés, y compris San Francisco, en Californie et le comté du Bronx, à New York) ou de 5 % par année (29 comtés). Les deux taux étaient supérieurs à celui de 2,5 % par année pour l'ACS complète. Sur trois années, l'échantillon cumulé de 9 % dans les grands comtés s'approche du taux de production de l'ACS de 12,5 % sur cinq ans et l'échantillon de 15 % dans les 29 autres comtés d'essai le surpasse. Donc, les données pour 1999 à 2001 recueillies pour les 34 comtés d'essai fournissent un modèle utile de la performance des estimations sur une période de cinq ans d'après les données de l'ACS. (Les deux comtés restants ont été échantillonnés au taux d'environ 1 % et sont exclus de l'étude.)

Comme nous le mentionnons dans l'introduction, les données pour la période de 1999 à 2001 recueillies auprès des comtés d'essai ont fourni le première avertissement que la variance des estimations ayant trait au secteur de recensement d'après l'ACS était considérablement plus grande que prévu, même après avoir tenu compte de la taille de l'échantillon. Au départ, on s'attendait à ce que les estimations infra-comté d'après l'ACS, comme les estimations touchant le secteur de recensement, aient une variance comparable à celle des estimations produites d'après le recensement décennal. Toutefois, en 2000, les données de dénombrement complet du recensement, recueillies au moyen des questionnaires abrégé et détaillé, ont fourni des totaux de contrôle pour l'estimation par le ratissage croisé dans des secteurs de pondération qui étaient habituellement contigus à des secteurs de recensement. Il n'existe aucune source directement comparable de totaux de contrôle pour l'ACS : le programme des estimations démographiques intercensitaires du Census Bureau est incapable de fournir des estimations satisfaisantes à ce niveau de finesse géographique. De 1999 à 2001, ainsi que pour la production de l'ACS en 2005, des chiffres de population de contrôle intercensitaires n'ont été utilisés que sur le plan du comté ou à un niveau d'agrégation plus élevé.

On pourrait soutenir que l'estimation au cours des recensements des États-Unis réalisés dans le passé est le problème d'estimation le plus semblable à celui qui se pose dans le cas de l'ACS. En s'efforçant d'établir une analogie entre l'estimation d'après les données de recensement et les objectifs d'estimation sur petits domaines dans le cas des estimations sur cinq ans d'après les données de l'ACS, une façon de s'attaquer au problème consisterait à essayer de trouver une nouvelle approche pour produire des totaux de contrôle comparables à ceux du recensement pour ce qui est du secteur de recensement en vue de simuler la pondération du questionnaire détaillé du recensement. Mais ce genre d'approche pose vraisemblablement des problèmes étant donné l'expérience antérieure de recherche portant sur des estimations démographiques intercensitaires.

L'approche adoptée ici s'apparente donc plutôt à ce que l'on pourrait considérer comme le deuxième problème d'estimation le plus semblable, c'est-à-dire l'estimation d'après les données du questionnaire détaillé canadien. Depuis 1991, Statistique Canada utilise l'estimation par la régression généralisée (GREG) dans la pondération de son recensement pour lequel, comme pour celui des États-Unis, des dénombrements complets sont utilisés dans les estimations. Afin d'adapter l'approche à l'ACS, les données de recensement correspondant à des dénombrements complets peuvent être remplacées par des données administratives dans l'estimateur GREG, comme nous le décrirons ici.

3. Estimation assistée par un modèle

3.1 Théorie générale

L'estimation par la régression généralisée fait partie d'une classe plus importante d'estimateurs assistés par un modèle. Dans les applications, il existe des chevauchements fréquents entre cette dernière et la classe apparentée des *estimateurs par calage* (Deville et Särndal, 1992). La littérature sur l'estimation assistée par un modèle est abondante; parmi les références clés, mentionnons les contributions et revues de Särndal (1984), Särndal, Swensson et Wretman (1992), Rao (1994; 2003, ch. 2) et Fuller (2002).

Les caractéristiques générales de l'estimation assistée par un modèle sont des éléments essentiels de la logique de la présente application. Contrairement à l'estimation basée sur un modèle, l'estimation assistée par un modèle est essentiellement sans biais par rapport au plan lorsqu'elle est appliquée sous les conditions que nous allons élaborer ici. Selon ces conditions, l'amélioration attribuable à l'estimation assistée par un modèle peut être mesurée simplement par la réduction réalisée de la variance. En général, le modèle utilisé pour l'estimation assistée par un modèle ne doit pas être parfaitement exact, mais une réduction importante de la variance n'a lieu que si le modèle est fortement prédictif.

Dans son ouvrage récent traitant de l'estimation sur petits domaines, Rao (2003, ch. 2) passe en revue l'estimation assistée par un modèle à titre de forme possible d'estimation sur petits domaines. Dans sa notation, considérons un échantillon, *s,* d'éléments, *j,* provenant d'une population, et un ensemble de poids initiaux, w_j, éventuellement égaux à l'inverse de la probabilité de sélection ($=\pi_j^{-1}$) ou à un poids similaire fondé sur le plan de sondage. Au départ, les estimations d'un total de population pour une caractéristique, *Y,* sont estimées par $\hat{Y} = \sum_s w_j y_j$. Les totaux de population pour les données auxiliaires $X = (X_1,..., X_p)^T$ sont connus, mais ils sont également estimés par $\hat{X} = \sum_s w_j x_j$. L'estimateur GREG prend la forme

$$\hat{Y}_{GR} = \hat{Y} + (X - \hat{X})^T \hat{B} \tag{1}$$

où

$$\hat{B} = (\hat{B}_1,..., \hat{B}_P)^T = \left(\sum_s w_j x_j x_j^T / c_j \right)^{-1} \sum_s w_j x_j y_j / c_j \tag{2}$$

pour des constantes $c_j > 0$.

Les formules qui précèdent semblent dépendre du choix de la caractéristique, *Y,* mais l'estimateur GREG peut être exprimé sous la forme d'un ajustement, $g_j(s)$, du poids initial, w_j, donnant $w_j^*(s) = g_j(s) w_j$, où

$$g_j(s) = 1 + (X - \hat{X})^T \left(\sum w_j x_j x_j^T / c_j \right)^{-1} x_j / c_j. \tag{3}$$

Grâce à cet ajustement, les nouveaux poids sont *calés* sur les données auxiliaires, en ce sens que l'estimation d'après les données d'enquête pondérées concorde avec le total connu de population,

$$\sum_s w_j^* x_j = X. \tag{4}$$

3.2 Estimation pour le questionnaire détaillé canadien

En 1991, Statistique Canada a remplacé l'estimation par le ratissage croisé utilisée pour le Recensement de 1986 par l'estimation GREG, qui a été progressivement perfectionné en 1996 et en 2001. Michael Bankier et ses collègues (Bankier, Rathwell et Majkowski, 1992; Bankier, Houle et Luc, 1997; Bankier et Janes, 2003) ont élaboré les détails de la mise en œuvre.

Les dénombrements complets obtenus d'après les questionnaires abrégé et détaillé constituaient les variables *x*. L'unité de travail de base pour l'estimation d'après le recensement était le secteur de pondération, qui a été subdivisé en aires de diffusion représentant le niveau d'agrégation le plus faible considéré pour la publication. En 2001, les secteurs de pondération contenaient, en moyenne, 1 865 logements privés occupés et l'aire de diffusion moyenne en contenait 239 (Bankier et Janes, 2003). Pour chaque recensement, on a utilisé une approche d'estimation en deux étapes, combinant une première étape destinée à réaliser le calage approximatif de certaines composantes de *X*

pour chaque aire de diffusion, suivie d'une deuxième étape consistant à caler exactement un aussi grand nombre que possible de composantes de X au pour ce qui est du secteur de pondération.

Dans le cadre d'une étape précoce de la recherche, Fay (2005a) a examiné plus en profondeur les méthodes appliquées au Canada et leurs résultats afin d'en dégager les principes généraux éventuellement utiles pour la résolution du problème de l'ACS.

3.3 Transfert des méthodes à l'ACS

Brièvement, la concordance entre l'estimation pour le recensement du Canada et le problème de l'ACS peut être établie grâce aux liens suivants :

1. Remplacer le dénombrement complet X par des données administratives pour la même année que celle de la collecte de l'ACS, y compris déterminer les variables x pour les cas échantillonnés de l'ACS.
2. Caler les poids sur X, mais ne pas publier les statistiques obtenues directement d'après les données administratives.
3. Insérer l'estimation GREG tôt dans le processus d'estimation, en n'altérant aucune des étapes de pondération de l'ACS subséquentes en ce qui concerne le comté.
4. S'attendre à ce que l'estimation GREG améliore la variance d'un grand nombre d'estimations d'après l'ACS $\hat{Y}$ au niveau du secteur de recensement.

Dans cette application, une caractéristique cruciale est l'importance de la base de sondage de l'ACS, c'est-à-dire le fichier maître d'adresses (MAF pour *Master adress file*). Le MAF est un inventaire permanent de toutes les unités de logement aux États-Unis. Il est mis à jour périodiquement pour tenir compte des changements dans la liste d'unités de logement. Le MAF a été mis à jour plusieurs fois en prévision du Recensement de 2000, puis après le Recensement de 2000 afin de tenir compte des résultats du recensement. Un numéro d'identification permanent MAFID qui reste le même dans toutes les versions est attribué à chaque unité de logement. Lors de l'échantillonnage pour l'ACS, deux versions différentes du MAF sont utilisées comme base de sondage chaque année afin que l'échantillon de l'ACS reflète fidèlement l'univers courant des logements.

Les données administratives utilisées dans le présent projet sont l'un des produits d'un effort pluriannuel du Census Bureau pour systématiser le traitement et l'utilisation de données administratives provenant de sources multiples, y compris la Social Security Administration et l'Internal Revenue Service. Avec un taux de réussite de l'ordre de 80 % à 90 %, les données administratives sont couplées aux MAF. Les données brossent un tableau de type recensement de la composition de chaque ménage, avec des valeurs connues ou imputées pour l'âge, le sexe, la race et l'ethnicité de chaque membre. Les fichiers utilisés dans ce projet ne contiennent pas de noms, de numéros de sécurité sociale, de données relatives au revenu, ni même une indication que les données proviennent d'une déclaration de revenus ou d'une autre source.

Dans ce projet, les données administratives qui sont couplées avec succès à la base de sondage MAF/ACS sont utilisées comme source du total non pondéré, X, dans l'équation (1) et, similairement, les cas échantillonnés pondérés pour les cas de l'ACS interviewés servent de base pour $\hat{X}$. L'étape GREG est insérée juste après la correction pour la non-réponse dans la série d'étapes de l'estimation. Comme les taux de réponse à l'ACS sont assez élevés, on peut soutenir que $\hat{X}$ est effectivement une estimation sans biais de X. Les autres étapes de l'estimation de l'ACS suivent alors, de sorte que l'intégration de l'étape GREG n'empêche pas les poids finals de l'ACS de concorder avec les contrôles antérieurs à des niveaux d'agrégation géographique plus élevés, comme les comtés.

Le principe fondamental qui sous-tend les stratégies de réduction de la variance est que les données administratives fournissent une prédiction statistique, et non une détermination exacte, de la composition démographique des ménages couverts par l'ACS. Le couplage a lieu pour ce qui est du numéro d'identification MAFID et aucun effort n'est fait en vue de déterminer si les caractéristiques de personnes particulières dans le fichier administratif concordent avec les caractéristiques déclarées dans les ménages participant à l'ACS. La réduction de la variance ne dépend pas d'une concordance cohérente, mais elle nécessite que les données administratives fournissent une bonne prédiction statistique de la composition géographique du ménage.

4. Résultats préliminaires

4.1 Données et méthodes

Les résultats présentés ici résument les observations publiées antérieurement (Fay 2006). Comme nous l'avons souligné plus haut, les données pour la période de 1999 à 2001 recueillies dans 34 comtés d'essai de l'ACS ont été traitées comme étant l'équivalent approximatif de données de production de l'ACS couvrant une période de cinq ans. Dans les secteurs de recensement les plus petits (< 300 unités de logement), les poids des non-répondants n'ont pas été ajustés (autrement dit, $g_j(s) = 1$). L'estimation GREG a été exécutée séparément dans chacun des 2 250 autres secteurs de recensement.

Dans la future application prévue, on utilisera les données administratives pour les années correspondantes. Idéalement, les données administratives pour 1999 à 2001 auraient dû être intégrées à l'essai, mais seules les données administratives pour 2000 étaient disponibles pour cette étude préliminaire. (En fait, un fichier d'enregistrements administratifs comparable n'existe pas pour 1999.) Comme nous l'avons souligné à la section précédente, la réduction de la variance dépend de la qualité de la prédiction statistique des données administratives; ici, la prédiction statistique requise voulait que les données administratives puissent prédire une année dans le passé ainsi qu'une année dans le futur pour la même année.

Pour les besoins de la présente étude, nous avons choisi $c_j = 1$ pour les équations (2) et (3) sans examiner d'autres possibilités. Dans chacun des secteurs de recensement où l'estimation GREG a été appliquée, les variables x comprenaient un vecteur de valeurs 1 qui assurait que les poids ajustés, $w_j^*(s) = g_j(s)w_j$, soient calés sur le total pour ce qui est du secteur de recensement des unités comprises dans la base de sondage. Jusqu'à sept variables âge-sexe ont été incluses (personnes de 0 à 17 ans, de 18 à 29 ans, de 65 ans et plus; hommes de 30 à 44 ans et de 45 à 64 ans; femmes de 30 à 44 ans et de 45 à 64 ans). Comme il est décrit en détail dans Fay (2006), des groupes âge-sexe réduits ont été utilisés lorsque le regroupement était nécessaire. De même, des variables de race/ethnicité pour les Hispaniques, les Noirs non-hispaniques et les autres races (y compris les Asiatiques et les Amérindiens) étaient disponibles et ont été incluses lorsque la diversité était suffisante pour justifier leur utilisation. Les règles de regroupement étaient fondées sur les exigences que l'inversion dans les équations (2) et (3) ne soit pas singulière et qu'aucun poids négatif ne soit calculé. Des tailles minimales ont également été spécifiées pour permettre l'utilisation de variables distinctes de race ou d'ethnicité.

Deux variantes plus simples ont été envisagées, l'une fondée uniquement sur un terme constant et l'autre excluant les variables de race ou d'ethnicité. La régression avec le terme constant est équivalente à une estimation par le quotient sur le total de la base de sondage pour le nombre d'unités de logement dans le secteur de recensement, qui s'approche d'une variante étudiée antérieurement par Starsinic (2005). La régression intermédiaire évalue la contribution distincte des variables de race et d'ethnicité.

4.2 Résultats

Les réductions de la variance ont été évaluées en comparant la somme des estimations de la variance sur l'ensemble des secteurs de recensement, *t,* autrement dit $\sum_t Var(\hat{Y})$ comparativement à $\sum_t Var(\hat{Y}_{RG})$. Le tableau 1 donne les résultats séparément pour les cinq comtés échantillonnés à 3 % et les 29 comtés échantillonnés à 5 %. Les variances ont été estimées par la méthode de rééchantillonnage. Pour l'étape GREG, le choix des variables x dans un secteur de recensement donné a été maintenu fixe, mais les autres aspects de l'ajustement GREG ont été répétés.

Tableau 1 Réduction en pourcentage provisoire de la variance estimée au niveau du secteur de recensement d'après trois stratégies d'estimation GREG possibles dans 34 comtés d'essai de l'ACS, 1999-2001. Les réductions sont présentées séparément pour cinq grands comtés échantillonnés au taux d'environ 3 % par année. Les 29 autres comtés ont été échantillonnés au taux de 5 % par année. Toutes les variances sont estimées pour les totaux estimés de chaque caractéristique.

	Comtés échantillonnés à 3 %			Comtés échantillonnés à 5 %		
	Régres-sion avec terme constant	Régres-sion avec âge/sexe	Régres-sion avec âge/sexe, race/ethn icité	Régres-sion avec terme constant	Régres-sion avec âge/sexe	Régres-ion avec âge/sexe, race/ethn icité
Unités de logement	90	91	91	87	88	88
Unités de logement occupées	69	74	74	62	69	69
Nombre total de personnes	47	67	68	42	66	66
Hommes de 0 à 17 ans	13	39	40	11	40	41
Femmes de 0 à 17 ans	13	39	40	11	41	41
Hommes de 18 à 29 ans	10	26	27	9	26	26
Femmes de 18 à 29 ans	11	28	28	10	28	28
Hommes de 30 à 44 ans	14	40	39	11	42	42
Femmes de 30 à 44 ans	17	46	46	13	47	47
Hommes de 45 à 64 ans	8	43	43	6	47	47
Femmes de 45 à 64 ans	11	45	46	6	49	49
Hommes de 65 et plus	1	25	25	-2	31	31
Femmes de 65 ans et plus	3	29	29	-2	35	35
Hispaniques	21	33	50	23	39	48
Non-Hispaniques noirs	22	33	46	16	32	43
Non-Hispaniques blancs	26	43	51	24	45	51
Autres races	10	30	44	9	19	26

Source : Fay (2006)

Les résultats du tableau 1 pour la régression avec terme contant reproduisent l'observation antérieure de Starsinic (2005) voulant qu'un simple ajustement par le ratio des poids de sondage sur les totaux des unités de logement au niveau du secteur de recensement améliore les estimations, particulièrement celles des caractéristiques de l'unité de logement. Les formes plus complexes de GREG n'apportent que des améliorations modestes pour les variables de logement.

Par contre, pour les variables démographiques, les améliorations de la variance due à l'étape GREG sont nettement supérieures à celles associées au simple ajustement par le ratio.

L'addition sélective des variables de race ou d'ethnicité provenant du fichier administratif dans X améliore la variance globale des estimations de l'ACS pour ces groupes dans le secteur de recensement. Les variances pour les autres caractéristiques étudiées demeurent essentiellement inchangées.

Enfin, les résultats sont assez semblables pour les comtés échantillonnés à 3 % et à 5 %, ce qui donne à penser que les améliorations de la variance sont stables sur une gamme de taux d'échantillonnage qui inclut ceux utilisés pour la version de production de l'ACS.

5. Discussion

5.1 Futurs travaux de recherche

Pour les besoins de l'application de l'ACS, un certain nombre de prolongations de la présente étude viserait à étudier plus en profondeur les questions de l'utilité potentielle de cette approche. L'étude était axée sur 1) une période, 1999 à 2001, 2) les dossiers administratifs pour 2000 uniquement et 3) les secteurs de recensement comme la seule unité géographique infra-comté d'intérêt. Dans les 34 comtés d'essai, des études sont en cours à l'heure actuelle en utilisant les données de l'ACS pour 1999 à 2005, des données administratives pour 2000 à 2005 et, pour les localités, les divisions civiles mineures (comme les petites villes) et d'autres unités géographiques infra-comté en plus des secteurs de recensement. Pour les estimations sur cinq ans, le secteur de recensement demeurera le secteur de pondération de base, mais l'élargissement de la recherche à des estimations sur trois ans ou sur un an se fonde sur des secteurs de pondération formés à partir d'unités géographiques d'un niveau d'agrégation plus élevé, comme les localités dont la population est supérieure à une taille spécifiée.

La mise en œuvre spécifique de l'estimateur GREG, quoique assez prometteuse, pourrait sans aucun doute être perfectionnée sous plusieurs angles, notamment celui de la sélection des variables, le nombre de variables qu'il convient d'inclure dans une situation donnée et la question de savoir si un autre choix pour c_j améliorerait la performance globale. Une variance de la procédure en deux étapes utilisée pour les recensements de la population canadienne, qui consiste à caler approximativement certaines caractéristiques au niveau de l'aire de diffusion et à caler exactement un plus grand nombre au niveau du secteur de pondération pourrait être essayé pour la relation analogue entre les groupes d'îlots et les secteurs de recensement des États-Unis dans les estimations sur cinq ans de l'ACS.

5.2 Élargissement à d'autres applications

Les formules classiques de l'estimateur GREG, y compris celles de Rao (2003), précisent qu'il est nécessaire de connaître les totaux de population pour les données auxiliaires, $X=(X_1,...,X_p)^T$ ainsi que pour les valeurs auxiliaires spécifiques, x_j, pour les cas échantillonnés. Mathématiquement, la connaissance des valeurs individuelles des données auxiliaires pour les cas non échantillonnés n'est pas nécessaire. Toutefois, pratiquement, la combinaison de données administratives à des données d'enquête comporte souvent une forme d'appariement. Si l'appariement est imparfait, alors les totaux auxiliaires contiennent des valeurs qui risquent de ne pas être représentées dans l'échantillon, ce qui affaiblit les affirmations d'absence de biais asymptotique.

Dans la présente application, le problème a été résolu en appariant les dossiers administratifs à la base de sondage complète (le MAF) pour valider les hypothèses en construisant les totaux $X=(X_1,...,X_p)^T$ qui pourraient être estimés d'après l'échantillon de l'ACS.

Une approche semblable ou équivalente pourrait être utilisée avec des données auxiliaires relatives à la santé dans des situations où pourraient survenir des ambigüités dans l'appariement de données auxiliaires. Dans de nombreuses applications, les conditions sur lesquelles repose la présente application, à savoir une liste comme base de sondage et la capacité de coupler la plupart de l'information auxiliaire à la base de sondage, pourraient ne pas être satisfaites. Néanmoins, les résultats empiriques donnent à penser ici que, lorsque des circonstances favorables se présentent, l'application de l'estimateur GREG peut produire des améliorations importantes de la variance.

Références

Bankier, M., A.-M. Houle, et M. Luc (1997), "Calibration Estimation in the 1991 and 1996 Canadian Censuses", *Proceedings of the Survey Research Methods Section, American Statistical Association,* pp. 66-75.

Bankier, M. et D. Janes (2003), "Regression Estimation of the 2001 Canadian Census", *Proceedings of the 2003 Joint Statistical Meetings on CD-ROM, American Statistical Association,* pp. 442-449.

Bankier, M. D., S. Rathwell, et M. Majkowski (1992), "Two Step Generalized Least Squares Estimation in the 1991 Canadian Census", *Proceedings of the Survey Research Methods Section, American Statistical Association,* pp. 764-769.

Deville, J.etd C.-E. Särndal (1992), "Calibration Estimators in Survey Sampling", *Journal of the American Statistical Association,* 87, pp. 376-382.

Fay, R. E. (2005a), "Model-Assisted Estimation for the American Community Survey", *Proceedings of the 2005 Joint Statistical Meetings on CD-ROM, American Statistical Association,* pp. 3016-3023.

_____ (2005b), "Potential Applications of Model-Assisted Estimation to Demographic Surveys in the U.S.", article présenté au Federal Committee on Statistical Methodology Research Conference, Arlington, VA, disponible sur www.fcsm.gov/05papers/Fay_IIIC.pdf.

_____ (2006), "Using Administrative Records with Model-Assisted Estimation for the American Community Survey", article présenté au Joint Statistical Meetings, Seattle, WA.

Fuller, W. A. (2002), "Estimation par régression appliquée à l'échantillonage", *Techniques d'enquête,* 28, pp. 5-25.

Rao, J. N. K. (2003), *Small Area Estimation,* John Wiley, New York.

Särndal, C.-E. (1984), "Design-Consistent Versus Model-Dependent Estimation for Small Domains", *Journal of the American Statistical Association,* 79, pp. 624-631.

Särndal, C.-E., B. Swensson, and J. Wretman (1992), *Model Assisted Survey Sampling,* Springer-Verlag, New York, NY.

Starsinic, M. (2005), "American Community Survey: Improving Reliability for Small Area Estimates", *Proceedings of the 2005 Joint Statistical Meetings on CD-ROM, American Statistical Association,* pp. 3592-3599.