Techniques d'enquête, décembre 1988
Vol. 14, n° 2, pp. 281–305
Statistique Canada

Redressement des estimations régionales par une repondération des ménages

ALAN M. ZASLAVSKY[1]

RÉSUMÉ

Supposons que des taux de sous-dénombrement ont été estimés pour un recensement et que des estimations du niveau de sous-dénombrement ont été établies pour les îlots. Il peut être alors souhaitable de dresser une nouvelle liste de ménages qui comprendrait les ménages qui auraient été oubliés. Nous proposons dans cet article de dresser une telle liste en pondérant les ménages qui ont été dénombrés. Les poids des ménages sont déterminés par des équations linéaires qui représentent le nombre total voulu de personnes dans chaque classe d'estimation et le nombre total voulu de ménages. On calcule alors des poids qui satisfont les contraintes et qui rapprochent le plus possible le tableau des données ajustées des données brutes. On peut voir dans cette méthode un exemple d'application de la méthode itérative du quotient à des cas où les contraintes ne concernent pas les fréquences marginales d'un tableau de contingence. Des covariables continues ou discrètes peuvent être utilisées dans les opérations de redressement et il est possible de vérifier directement si les contraints peuvent être satisfaites. Enfin, nous proposons des méthodes pour l'utilisation de données pondérées à des fins diverses liées au recensement et pour le redressement de données corrélées sur les caractéristiques des ménages oubliés, par exemple le revenu, qui ne sont pas considérées directement dans l'estimation du niveau de sous-dénombrement.

MOTS CLÉS: Sous-dénombrement; méthode itérative du quotient; redressement d'estimations régionales; données manquantes.

1. INTRODUCTION

Beaucoup de recherches ont été faites ces dernières années afin d'élaborer des méthodes pour estimer le taux de sous-dénombrement dans le recensement de 1990 aux États-Unis (National Academy of Sciences 1985). Parmi les principales voies envisagées, notons l'exécution d'une enquête post-censitaire (EP) dans un échantillon d'îlots peu de temps après le recensement. La proportion des personnes enregistrées dans l'EP, qui ne figureraient pas dans les listes du recensement serait une estimation du taux de sous-dénombrement pour le recensement. Des estimations du niveau de sous-dénombrement seraient établies pour un niveau géographique inférieur quelconque (probablement la plus petite unité géographique utilisée pour le recensement, c'est-à-dire l'îlot). Ces estimations se rapporteraient à des classes formées en fonction des caractéristiques des personnes et peut-être de certaines caractéristiques des ménages ou des îlots. Par «classe» nous entendrons ici les classes ou cellules d'estimation ou de redressement; par «îlot» nous entendrons la plus petite unité géographique pour laquelle il existe des estimations du niveau de sous-dénombrement. Le recensement de 1980 a permis de dénombrer environ 100 millions de ménages dans deux à quatre millions d'îlots, selon les définitions utilisées.

Pour chaque îlot, le résultat des méthodes décrites ci-dessus serait un vecteur d'estimations du sous-dénombrement, formé de S éléments rattachés à S classes de redressement et dont la somme correspondrait au nombre estimé de personnes non recensées pour l'îlot en question. Il n'est pas de notre ressort ici d'exposer les méthodes qui servent à établir ces estimations. Toutefois, dans les exemples que nous utiliserons, nous supposerons qu'il existe un taux de sous-dénombrement pour chaque classe dans chaque îlot, lequel taux exprimera le nombre

[1] Alan M. Zaslavsky, Statistics Center, Massachusetts Institute of Technology, Pièce E40-111, Cambridge, MA 02139, É.-U. et Département de statistique, Harvard University, Cambridge, MA 02138, É.-U.

estimé de personnes oubliées en proportion du nombre de personnes recensées dans la classe en question. Dans cet article, le mot «redressement» désigne toute opération par laquelle on ajoute le chiffre estimé du sous-dénombrement aux chiffres du recensement. Les classes de redressement peuvent ou non correspondre aux strates formées a posteriori dans l'analyse d'un programme post-censitaire. Ce genre de données pourraient bien se prêter au calcul de fréquences marginales par caractéristique pour les personnes. On pourrait, en particulier, se servir des totaux d'îlots pour calculer les données régionales destinées à divers usages officiels ou commerciaux.

Toutefois, il serait préférable dans certains cas d'inclure les personnes additionnelles dans des ménages. Pour cela, nous supposons qu'il est possible d'estimer le nombre de ménages oubliés dans chaque îlot. Il pourrait aussi y avoir des données qui permettraient de distinguer les personnes qui ont été oubliées à l'intérieur des ménages recensés et les personnes qui font partie des ménages oubliés.

Pour que les chiffres redressés prennent toute leur signification, la composition des ménages ajoutés et les liens entre les membres de chaque ménage doivent être cohérents et représentatifs des caractéristiques des ménages de la région. Le terme «composition» désignera ici le nombre de membres du ménage qui appartiennent à chaque classe de redressement. Par exemple, un ménage formé d'un chef de ménage de sexe féminin et de race blanche âgé de 20 ans, d'une personne de sexe masculin et d'origine chinoise âgée de 75 ans et d'une fille de race noire âgée de 10 ans ne serait pas très vraisemblable même si tous ses membres appartenaient à des classes qui sont bien représentées dans l'îlot. Néanmoins sur le plan abstrait, le fait de définir ces profils et de former des ménages qui leur correspondent est une tâche gigantesque.

Exemple 1: *Constitution d'une liste de ménages.*

Le tableau 1 illustre une portion d'une liste de recensement comme celles pouvant figurer sur une bande de microdonnées.

Le tableau 2 représente la même liste mais cette fois sous forme de résumé de la composition des ménages dans le cas où il existe seulement trois classes d'estimation: (1) hommes de plus de 20 ans, (2) femmes de plus de 20 ans et (3) enfants de 20 ans ou moins.

Essentiellement, le même problème se représente chaque fois qu'il faut repondérer les résultats d'une enquête sur les ménages de manière à les rendre compatibles avec les totaux marginaux connus pour diverses catégories de personnes.

L'objet essentiel de la méthode proposée dans cet article est de pondérer les ménages qui figurent sur la liste de recensement de l'îlot de manière que les totaux pondérés de personnes dans chaque classe de redressement et le total pondéré des ménages égalent parfaitement les totaux redressés correspondants. Ainsi, bien que la pondération modifie la composition proportionnelle de l'îlot, tous les ménages sont réels et présentent des caractéristiques et des relations qui sont cohérentes et raisonnables pour l'îlot en question. Cette méthode de pondération ressemble à la méthode itérative du quotient, où le poids appliqué à la fréquence d'une case d'un tableau de contingence est le quotient de la fréquence redressée par la fréquence initiale. Les poids des ménages sont calculés *après* que les totaux d'îlots ont été redressés et sont conformes à ces totaux. Pour la plupart des usages du recensement, les enregistrements pondérés peuvent servir de base à la création de tableaux et de listes.

On peut opposer les méthodes ci-dessus aux méthodes d'imputation, qui consistent à représenter les unités non recensées par des unités entières ajoutées à la liste. Les unités imputées peuvent être des personnes ou des ménages. Bien qu'il soit possible d'imputer des personnes dans l'îlot, on n'a toujours pas trouvé le moyen d'intégrer ces personnes dans des ménages plausibles. L'intégration de ces personnes à des foyers collectifs fictifs, comme cela s'est fait dans certains tests d'opérations de redressement, a pour effet d'éluder le problème en créant une fausse image des relations dans l'îlot. La repondération ou l'imputation de personnes seraient des solutions convenables pour les pensionnaires d'institution ou de foyer collectif car dans ce cas, la configuration du ménage n'a pas d'importance.

Tableau 1

Extrait d'un fichier de microdonnées d'un échantillon

Nom	Adresse	Sexe	Âge
John Smith	328 rue Principale	H	34
Mary Smith	328 rue Principale	F	32
Louise Smith	328 rue Principale	F	7
Nancy Chen	330 rue Principale	F	62
Jorge Ramirez	332 rue Principale	H	21
Juan Ramirez	332 rue Principale	H	24

Tableau 2

Fichier de microdonnées reconstruit selon les ménages et montrant la composition des ménages.

Adresse	Nombre de personnes par classe		
	Classe 1	Classe 2	Classe 3
328 rue Principale	1	1	1
330 rue Principale	0	1	0
332 rue Principale	2	0	0

Une autre méthode d'imputation consiste à utiliser des modèles probabilistes pour le nombre de personnes non recensées à l'intérieur de ménages et le nombre de ménages non recensés et à calculer un nombre de ménages imputés à partir de la distribution a posteriori des ménages oubliés, étant donné les ménages recensés. Cette méthode est assimilée à l'imputation multiple (Rubin 1987), où tout le processus d'imputation est répété plusieurs fois afin de tenir compte de la variabilité attribuable au sous-dénombrement. Toutefois, dans chaque liste d'îlot ainsi créée, les totaux fondés sur les ménages recensés et imputés ne correspondront pas nécessairement parfaitement aux totaux redressés voulus. Dans cet article, nous nous intéressons surtout à des méthodes qui permettent un ajustement *exact* à des estimations démographiques calculées à une étape antérieure.

Dans les sections qui suivent, nous allons élaborer des méthodes qui permettront d'effectuer le redressement par pondération proposé. La section 2 contient une formulation mathématique des objectifs du plan de pondération tandis que la section 3 expose la façon d'ajuster les poids. Dans la section 4, nous voyons comment faire la distinction dans le plan entre les personnes oubliées à l'intérieur de ménages et les personnes oubliées à cause du sous-dénombrement de ménages. La section 5 renferme quelques perfectionnements qui ont pour but d'accroître la robustesse de la méthode à l'égard de la variabilité des petits îlots. La section 6 contient les résultats de simulations. La section 7 porte sur l'utilisation de données pondérées à diverses fins liées au recensement tandis que la section 8 examine les effets du redressement par pondération sur les covariables qui ne font pas partie du plan utilisé pour la formation des classes de redressement. Enfin, dans la section 9, nous exposons brièvement quelques problèmes non résolus et proposons des sujets de recherche.

2. OBJECTIFS D'UN PLAN DE PONDÉRATION ET FORMULATION MATHÉMATIQUE

Un des buts essentiels du plan proposé est de répartir la population de l'îlot dans des ménages valides de sorte que les statistiques portant sur les ménages soient clairement définies. C'est

pourquoi on attribue des poids aux *ménages*, les *membres* d'un ménage ayant tous le même poids.

Pour que les chiffres de la liste pondérée correspondent à ceux obtenus par le redressement préalable, les contraintes suivantes doivent être respectées:

(A1) Dans chaque îlot, la somme des poids des ménages égale le nombre redressé de ménages.

(A2) Dans chaque classe de redressement et chaque îlot, la somme des poids des personnes égale le nombre redressé de personnes.

Pour que la liste d'îlot pondérée ressemble le plus possible à la liste initiale, il faut aussi respecter la contrainte suivante:

(B) Les poids doivent être, pour ainsi dire, aussi près que possible les uns des autres.

Avec des poids d'unités (ou poids égaux), la composition de l'îlot demeure inchangée. Si les poids ne sont pas très différents les uns des autres, la composition de l'îlot pour le recensement demeure à peu près la même grâce au plan de pondération. Conformément aux règles d'usage sur l'utilisation des poids dans les enquêtes, nous ne devons pas modifier en profondeur notre représentation de la composition de l'îlot, sauf si les données sur le sous-dénombrement le justifient.

Nous allons maintenant passer à la formulation mathématique de ces critères. Supposons que dans l'îlot considéré, il y a S classes de redressement et I ménages recensés et que le ménage i compte C_{is} membres de la classe s. Supposons que H est le nombre total de ménages que l'on souhaite voir dans la liste redressée pour l'îlot et D_s est le nombre total de personnes que l'on souhaite voir dans la classe s. Soient W_i, $i = 1,2, \ldots I$, les poids correspondant aux ménages. (A1) exige que

$$\sum_{i=1}^{I} W_i = H \tag{1}$$

et (A2) exige que

$$\sum_{i=1}^{I} W_i C_{is} = D_s, \; s = 1,2 \ldots S. \tag{2}$$

Ces contraintes peuvent être représentées par une équation matricielle de la forme $AW = B$, où

$$A = \begin{bmatrix} \mathbf{1} \\ C' \end{bmatrix}, \; B = \begin{bmatrix} H \\ D \end{bmatrix}, \; W' = [W_1 \; W_2 \ldots \; W_I] \text{ et } D' = [D_1 \; D_2 \ldots D_s] \tag{3}$$

et **1** représente une ligne de uns.

Pour réaliser l'objectif (B), nous choisissons une fonction objectif qui représente la distance entre les poids W et les poids uniformes et nous la minimisons. Nous utiliserons à cette fin la fonction objectif $T = \sum W_i \log (W_i)$. Cette mesure est proportionnelle à l'information discriminante (information de Kullback-Liebler) de la distribution de probabilité discrète (selon les ménages) avec poids relatifs W_i par rapport à la distribution de probabilité avec poids égaux, et elle est la fonction objectif qui est à la base de la méthode itérative du quotient classique (ajustement proportionnel itératif), utilisée pour le redressement des tableaux de contingence (Deming et Stephan 1940; Ireland et Kullback 1968; Oh et Scheuren 1978; ce dernier ouvrage renferme une bibliographie plus complète). Notre méthode peut donc être considérée comme

un prolongement de la méthode itérative du quotient. Scheuren (1973) utilise la méthode itérative du quotient pour la repondération des ménages; Cilke et Wyscarver (1988) font une repondération en fonction de contraintes linéaires mais utilisent pour cela une fonction objectif différente de celles considérées ici. Alexander (1987) a élaboré de son côté des méthodes semblables à celles exposées ici.

Pour ce qui a trait à la méthode itérative du quotient, des fréquences initiales X figurent dans les cases d'un tableau de contingence et on calcule de nouvelles fréquences de case Y pour minimiser la fonction objectif $\sum Y_i \log (Y_i/X_i)$. Par conséquent, les poids des observations originales sont les rapports $W_i = Y_i/X_i$. Pour ce qui a trait à notre méthode, si X_i ménages se trouvaient avoir exactement la même composition, nous pourrions les imaginer comme une seule entité dans la liste, avec une fréquence initiale X_i et une fréquence redressée Y_i. Cependant, s'il y a un grand nombre de classes de redressement, il est peu probable que plusieurs ménages du même îlot aient exactement la même composition. C'est pourquoi nous ne tenterons pas de grouper les ménages; il sera plus facile, au point de vue de la notation et du calcul, de lister les ménages séparément de sorte que pour la composition de chaque ménage recensé, la fréquence initiale $X_l = 1$ et $Y_l = W_l$. Abstraction faite de la différence de notation, la seule chose qui distingue la formulation mathématique d'un plan de pondération de celle d'un redressement par la méthode itérative du quotient est que les contraintes linéaires n'ont pas la structure particulière des fréquences marginales d'un tableau de contingence. Pour abréger la présentation des exemples, nous indiquerons parfois sur une ligne un nombre qui représentera le nombre de lignes identiques à cette ligne dans la liste des ménages.

Pour ce qui a trait au tableau de contingence, la méthode itérative du quotient préserve les rapports de produits croisés des cases de même que l'indépendance des variables, lorsque celle-ci existe dans le tableau initial. C'est pourquoi, dans le calcul d'estimations régionales, la méthode itérative du quotient est appelée «estimation avec préservation des structures» (Purcell 1979; Purcell et Kish 1979). Voir section 10.1 pour une analyse approfondie des fonctions objectif.

Notre méthode diffère de la méthode itérative du quotient en ce que les contraintes linéraires ne se rapportent pas nécessairement aux fréquences marginales d'un tableau de contingence. La méthode itérative du quotient est un cas particulier d'application de notre méthode; il en est de même de la version généralisée de la méthode itérative du quotient de Oh et Scheuren (1978), qui utilise des tableaux différents pour ajuster chaque fréquence marginale. De fait, les covariables continues aussi bien que discrètes peuvent faire l'objet de contraintes; des applications de ce genre sont proposées dans la section 8.3. Par ailleurs, les algorithmes que nous proposons permettent de déterminer directement s'il y a des poids qui sont compatibles avec toutes les contraintes données. Nous pouvons ensuite choisir des contraintes qui doivent être adoucies pour ajuster les poids. Grâce à ces caractéristiques, les possibilités d'application des méthodes proposées vont au-delà de la représentation du sous-dénombrement.

3. AJUSTEMENT DES POIDS

Nous allons maintenant chercher à calculer des poids qui satisfont les contraintes $AW = B$, $W \geq 0$, en minimisant la fonction objectif $T = \sum W_i \log (W_i)$. Pour que T soit une fonction continue de W, nous adoptons la règle usuelle $0 \log 0 = 0$.

Nous désignerons tout vecteur de poids qui satisfait les contraintes linéaires (équations et inéquations) comme une *solution réalisable*. Tant que le poids total des ménages est soumis à une contrainte, l'ensemble des solutions réalisables est borné et T suppose par conséquent une valeur minimum pour l'ensemble; de plus, comme T est strictement convexe, la solution est unique.

Le calcul des poids se divise donc naturellement en trois étapes: (1) déterminer si les contraintes $AW = B$ sont cohérentes; (2) déterminer s'il y a des solutions réalisables; et (3) trouver la solution réalisable qui minimise T. Nous supposons qu'il y a I ménages et p contraintes de sorte que A est une matrice $p \times I$.

Tableau 3

Liste de ménage

Ligne n°	Nombre de membres du ménage par classe			Nombre de ménages
	Classe 1 (hommes)	Class 2 (femmes)	Class 3 (enfants)	
1	0	1	0	50
2	0	1	1	40
3	1	0	0	40
4	1	0	2	15
5	1	1	0	50
6	1	1	1	60
7	1	1	2	40

Tableau 4

Totaux redressées

	Chiffre brut	Taux de redressement	Chiffre redressé
Classe 1	205	.05	215
Classe 2	240	.03	247
Classe 3	210	.04	218
Ménages	295	.02	301

Exemple 2: *Ajustement de poids.*

Le tableau 3 est un exemple d'une liste des ménages d'un îlot, où trois classes sont représentées comme dans l'exemple 1; nous pouvons penser à des classes comme «hommes», «femmes» et «enfants». Ce tableau peut être considéré comme la version condensée d'un tableau de 295 lignes, où chaque ligne correspond à un ménage.

Le tableau 4 contient le nombre non redressé et le nombre redressé de ménages et de personnes dans chaque classe. On obtient le nombre redressé en appliquant le taux de redressement indiqué et en arrondissant. La méthode utilisée pour obtenir les chiffres redressés n'a toutefois rien à voir avec le reste du processus.

3.1 Cohérence des contraintes linéaires

Tant que les lignes de A sont indépendantes, les contraintes $AW = B$ seront cohérentes. Si une ligne quelconque de A est dépendante des autres, la contrainte correspondante est soit incohérente ou superflue, selon les valeurs de B. On peut repérer les lignes dépendantes en appliquant la décomposition Q-R à A'. Si les contraintes correspondantes sont superflues, on peut les supprimer sans perte d'information; si elles sont incohérentes, elles doivent être reformulées de quelconque façon.

Exemple 2: (*suite*).

Dans cet exemple, les lignes de la matrice A sont indépendantes et les contraintes sont donc cohérentes.

Dans la section 5, nous examinons des situations où il est susceptible d'exister des contraintes incohérentes et analysons des méthodes qui permettent de les résoudre.

Tableau 5

Poids optimaux pour l'exemple 2

Ligne n°	Poids	Chiffres pondérés
1	0.9554	47.77
2	0.9557	38.23
3	0.9816	39.27
4	0.9823	14.73
5	1.0730	53.65
6	1.0734	64.40
7	1.0737	42.95

3.2 Existence de solutions réalisables

Déterminer l'existence de solutions réalisables équivaut à déterminer une solution réalisable initiale dans un problème de programmation linéaire où les algorithmes réguliers peuvent être utilisés. Supposons que nous devions trouver une solution positive W à l'équation $AW = B$, où $B \geq 0$. (Si cette dernière condition n'est pas respectée, nous pouvons faire en sorte qu'elle le soit en modifiant le signe des éléments négatifs de B et des lignes correspondantes de A) Nous pouvons élargir le problème en posant $[A \mid I]\ [W' \mid Z']' = B,\ W, Z \geq 0$, où I est une matrice unité $p \times p$ et Z est une variable vectorielle à p éléments. Ce problème a automatiquement une solution initiale $W = 0, Z = B$. Nous appliquons ensuite la méthode du simplexe (comme dans Gass (1964) ou tout autre ouvrage sur la programmation linéaire) pour minimiser $\sum Z_i$. Si cette somme peut être réduite à 0, les valeurs W correspondantes sont une solution au problème original, tandis que si elle ne peut l'être, le problème original n'a pas de solution.

Exemple 2: (*suite*).

Une solution réalisable (mais non optimale) donne des totaux pondérés de 86, 54, 29 et 132 pour les ménages des lignes 2, 3, 5 et 6 respectivement dans le tableau 3. Il est possible de vérifier que ces chiffres donnent les totaux redressés prévus pour les ménages et les personnes de chaque classe.

Le problème de l'inapplicabilité se rapproche de celui de l'incohérence et est aussi analysé dans la section 5.

3.3 Optimalisation de la fonction objectif

Selon la méthode des multiplicateurs de Lagrange, la solution de minimisation doit satisfaire les équations $\partial T/\partial W_i = \log W_i + 1 = a_i'\lambda$, où a_i représente la i-ème colonne de A et $\lambda' = (\lambda_1, \lambda_2, \ldots \lambda_p)$. Alors, $W_i = \exp(a_i'\lambda - 1)$; le modèle pour les poids est donc de forme logarithmique linéaire, comme celui utilisé pour les redressements par la méthode itérative du quotient. λ_s représente l'accroissement de poids logarithmique lié à un accroissement unitaire du coefficient de liaison correspondant a_{is}, par exemple, l'inclusion dans le ménage d'une personne faisant partie de la classe de redressements.

Pour satisfaire l'équation $AW = B$, nous pouvons résoudre en fonction de λ à l'aide de la méthode de Newton. Le modèle d'itération utilisé est

$$\lambda^{(t+1)} = \lambda^{(t)} - (AW^* A')^{-1} (AW - B), \tag{4}$$

où W est la matrice dont la diagonale est formée des éléments de $W = W(\lambda^{(t)})$. Une bonne valeur initiale pour λ est $\lambda^{(0)} = (AA')^{-1}B$, laquelle peut être déduite d'une approximation linéaire fondée sur des poids initiaux égaux. (Dans la section 10.2, nous décrivons une méthode de la descente cyclique pour résoudre ces équations; cette méthode est une généralisation de l'ajustement proportionnel itératif).

Exemple 2: (*suite*).

Le tableau 5 donne le poids par ménage et le total pondéré des ménages (poids multiplié par le chiffre brut) pour chaque ligne du tableau 3. Aucun ménage n'est pondéré de plus de 8% à la hausse et de plus de 5% à la baisse.

4. REDRESSEMENT DU NOMBRE DE MÉNAGES ET DU NOMBRE DE PERSONNES À L'INTÉRIEUR DES MÉNAGES

Nous voyons maintenant la distinction à faire entre le redressement du nombre de personnes à l'intérieur des ménages (c'est-à-dire le redressement en fonction du sous-dénombrement à l'intérieur des ménages) et le redressement du nombre de ménages (c'est-à-dire le redressement en fonction du sous-dénombrement des ménages). Cette distinction avait déjà été faite en vue d'analyser les causes du sous-dénombrement (Fay 1986). Nous voulons ici nous en servir pour représenter plus fidèlement le sous-dénombrement dans une opération de redressement.

Le redressement du nombre de personnes à l'intérieur des ménages ne consiste pas à ajouter des ménages à la liste mais uniquement à redistribuer les poids des ménages de manière à accroître les totaux pondérés de personnes dans les diverses classes. En d'autres termes, les ménages qui n'ont personne ou qui comptent très peu de personnes dans une classe donnée sont pondérés à la baisse tandis que ceux qui en comptent beaucoup sont pondérés à la hausse de manière que le poids total des ménages demeure constant. Ainsi, ce mode de redressement aura pour conséquence inéluctable de réduire le poids de certains ménages. En revanche, le redressement du nombre de ménages touche les ménages qui n'ont pas été du tout recensés. Comme ce mode de redressement ne doit pas modifier les données relatives aux ménages recensés, il convient d'ajouter des ménages à la liste sans réduire le poids des ménages qui ont été recensés.

Nous proposons de distinguer ces deux modes de redressement. Une série de contraintes représente le redressement du nombre de personnes à l'intérieur des ménages. Dans ce cas, le poids total des ménages doit être égal au nombre de ménages recensés tandis que le poids total des personnes dans chaque classe doit être égal à la somme du nombre de personnes recensées et du nombre redressé de personnes pour cette classe. $AW_1 = B_1$, où B_1 est constituée du nombre de ménages *recensés* et des effectifs de classes redressés en fonction du sous-dénombrement à l'intérieur des ménages.

Une seconde série de contraintes représente le redressement du nombre de ménages. Dans ce cas, le poids total des ménages doit être égal au nombre estimé de ménages oubliés tandis que le poids total des personnes dans chaque classe doit être égal au nombre estimé de personnes dans les ménages oubliés. $AW_2 = B_2$, où B_2 est constituée du nombre de ménages additionnels et du nombre de personnes additionnelles par classe, redressés en fonction du sous-dénombrement *des ménages*.

Après avoir ajusté deux séries de poids correspondant aux deux séries de contraintes, nous additionnons les deux poids pour chaque ménage afin d'obtenir un poids qui intègre les deux modes de redressement ($W = W_1 + W_2$). La différenciation des deux modes de redressement peut produire une série de poids redressés différente de celle qui serait obtenue si on ne faisait pas une telle distinction, comme le montre l'exemple 3. Toutefois, si on ne fait pas cette distinction dans l'estimation du taux de sous-dénombrement, il est quand même possible d'opérer un redressement en une seule étape.

Exemple 3: *redressement en fonction du sous-dénombrement des ménages.*

Supposons qu'il n'y a que deux classes de redressement et qu'un îlot hypothétique est composé de la façon décrite par les trois premières colonnes du tableau 6.

Tableau 6

Nombre hypothétique de ménages pour l'exemple 3 (chiffres bruts et chiffres redressés)

Composition du ménage		Chiffres bruts (nombre de ménages)	(1) Les personnes oubliées font partie de ménages recensés ou non	(2) Les personnes oubliées font partie de ménages non recensés	
Nombre de personnes de la classe 1	Nombre de personnes de la classe 2		Chiffres redressés	Nombre de ménages non recensés	Totaux redressés, ménages recensés et non recensés
1	1	10,000	9904.54	.01	10,000.01
1	2	10,000	10106.46	10.99	10,010.99
2	1	10,000	10106.46	10.99	10,010.99
2	2	10	13.54	99.01	109.01

Supposons que nous devons ajouter 231 personnes de chaque classe et 121 ménages aux 30,010 ménages recensés. Les trois dernières colonnes du tableau 6 contiennent les chiffres redressés selon deux hypothèses: (1) les personnes oubliées peuvent faire partie de ménages qui ont été recensés ou non et (2) toutes les personnes oubliées font partie des ménages qui n'ont pas été recensés.

Dans le cas de la première hypothèse, l'algorithme pondère à la baisse les ménages qui ne comptent qu'une personne de chaque classe (1,1) et pondère à la hausse les ménages qui comptent deux personnes d'une classe et une de l'autre (1,2 et 2,1). Même si les ménages qui comptent deux personnes de chaque classe sont fortement pondérés à la hausse (par un facteur de 1,354), ils ne reçoivent qu'une faible proportion des personnes ajoutées à cause de leur très petit nombre à l'origine.

Dans le cas de la seconde hypothèse, on calcule tout d'abord des poids de manière à intégrer $231 \times 2 = 462$ personnes dans 121 ménages additionnels, puis on additionne ces poids au poids unitaire contenu dans les chiffres bruts. Tandis qu'aucune catégorie de ménages n'est pondérée à la baisse, les ménages qui comptent deux personnes de chaque classe (2,2) sont très fortement pondérés à la hausse (par un facteur de 10,901). De fait, il est mathématiquement impossible d'intégrer 462 personnes dans 121 ménages formés de deux, trois ou quatre personnes chacun sans avoir au moins 99 ménages de quatre personnes. Par conséquent, le fait de savoir que les personnes ajoutées (ou une proportion connue de celle-ci) font partie des ménages qui ont été oubliés modifie sensiblement notre représentation du genre de redressement à effectuer.

5. APPLICABILITÉ DES CONTRAINTES

Dans les sections précédentes, nous avons supposé qu'il y avait des solutions réalisables pour le problème d'optimisation conditionnelle. Nous allons maintenant considérer des cas où les solutions sont inexistantes ou insatisfaisantes et allons examiner des méthodes pour résoudre ces cas.

5.1 Dans quels cas les contraintes sont-elles inapplicables?

Il y a trois cas où les contraintes peuvent éliminer toute possibilité de solution satisfaisante: (1) lorsqu'elles sont effectivement incohérentes, (2) lorsqu'elles sont cohérentes mais qu'il n'y a aucun poids positif pour les satisfaire, et (3) lorsqu'il y a une solution réalisable mais qu'il

faille pour cela redresser à outrance le poids de certains ménages. Les problèmes liés à ces trois situations sont relativement comparables.

Il est possible de définir des contraintes qui sont incohérentes en soi, par exemple que toutes les classes d'hommes sont redressées de 2% à la hausse tandis que le nombre total d'hommes est redressé de 4% à la hausse. Selon la méthode que nous utilisons, chaque contrainte porte uniquement sur le nombre de personnes contenues dans une classe de redressement de sorte qu'il n'y a pas d'incohérence du genre de celle mentionnée ci-dessus. Toutefois, il y a encore la possibilité d'une incohérence conditionnelle, c'est-à-dire d'une incohérence qui dépend du nombre de ménages correspondant à chaque mode de composition dans un îlot. Nous donnons ci-dessous des exemples d'incohérence conditionnelle, d'inapplicabilité ou de poids insatisfaisants:

(1) Selon les méthodes proposées pour estimer le taux de sous-dénombrement, il est question de définir plus de 100 classes de redressement. Dans un îlot restreint mais hétérogène, le nombre de classes représentées pourrait être supérieur au nombre de ménages; par conséquent, il y aurait plus de contraintes que de poids à ajuster. Une incohérence est alors presque inévitable.

(2) Si tous les ménages d'un îlot comptent exactement le même nombre de personnes d'une classe de redressement en particulier (par ex., chaque ménage compte une jeune fille d'origine hispanique), le nombre de membres de cette classe n'est pas modifié par la redistribution des poids.

(3) Le redressement du nombre de ménages peut aboutir à des résultats incompatibles avec ceux qui découleraient du redressement du nombre de personnes dans une classe quelconque (cela pouvant être interprété comme une défaillance du modèle pour le redressement du nombre de ménages). Prenons par exemple le cas où il faudrait ajouter plus d'hommes que de ménages mais où aucun ménage de l'îlot ne compte plus d'un homme. Les contraintes seraient alors cohérentes mais inapplicables puisqu'on ne pourrait les satisfaire qu'en attribuant des poids négatifs aux ménages qui ne comptent pas d'hommes.

(4) Un îlot peut avoir des taux de sous-dénombrement qui diffèrent totalement de ceux qui ont été estimés pour des îlots de l'EP. Supposons, par exemple, qu'il y a un fort taux de sous-dénombrement dans la plupart des îlots (y compris la plupart des îlots de l'échantillon de l'EP) pour les hommes qui présentent certaines caractéristiques mais que dans l'îlot faisant l'objet d'un redressement, les hommes de cette catégorie sont bien dénombrés et se retrouvent dans la plupart des ménages. L'estimation du niveau de sous-dénombrement pour cette classe pourrait entraîner un tel redressement à la hausse qu'il ne serait pas possible de concilier les résultats de ce redressement avec les chiffres des ménages existants.

(5) On pourrait devoir hausser sensiblement le poids des ménages qui comptent des personnes issues d'une combinaison de classes de redressement propre à un ménage, de sorte que les ménages en question recevraient des poids extrêmes. Dans ce cas, le problème peut être résolu mais la solution n'est pas très satisfaisante.

Des problèmes d'inapplicabilité peuvent également surgir lorsqu'on ne peut associer aussi facilement la difficulté à une incohérence particulière dans le processus de redressement.

5.2 Méthodes pour rendre les contraintes applicables

Peu importe l'étape de l'ajustement à laquelle on constate l'inapplicabilité, il existe plusieurs méthodes permettant d'assouplir les contraintes et de les rendre applicables. Nous allons maintenant analyser plusieurs de ces méthodes en exposant la logique intuitive qui est à l'origine du choix de chacune d'elles de même que les méthodes de calcul requises.

5.2.1 Méthodes fondées sur l'élimination de lignes (contraintes) de A

Lorsqu'on vérifie la cohérence des contraintes, on peut découvrir que certaines lignes sont linéairement dépendantes des lignes précédentes et que, par conséquent, elles sont superflues

ou incohérentes. En extrayant simplement ces lignes de la matrice A, on obtient un ensemble cohérent de contraintes et là s'arrêtent les calculs requis.

Si les contraintes sont disposées par ordre décroissant d'importance, les moins importantes sont éliminées si elles sont incompatibles avec les plus importantes. Cette disposition est tout à fait logique si les contraintes initiales portant sur les classes de redressement (définies par une classification de la population selon plusieurs critères) sont redéfinies selon les règles de l'analyse de variance comme des contraintes concernant la population totale («moyenne globale»), les classes définies par une variable de classification («effets principaux») et les classes définies par des interactions. Par exemple, si nous avons 10 classes de redressement définies par le sexe et cinq groupes d'âge, les nouvelles contraintes seraient par ordre d'importance: population totale (1 contrainte), population selon le sexe (1 contrainte additionnelle), population selon l'âge (4 contraintes additionnelles), interactions âge-sexe (les 4 dernières contraintes). Les quatre contraintes fondées sur l'âge pourraient être subdivisées entre personnes jeunes et personnes âgées (1 contrainte) et 3 contraintes additionnelles pourraient être définies pour chacun de ces groupes.

On peut appliquer une méthode similaire au moment de vérifier l'applicabilité des contraintes. S'il n'est pas possible de satisfaire l'équation $Z_i = 0$, pour tous les i, on peut modifier la fonction objectif du problème de programmation linéaire de manière à obtenir $\sum c_i Z_i$, où les coefficients $c_i > 0$ correspondant aux contraintes les plus importantes prennent les valeurs les plus grandes. On peut alors définir un ensemble maximal de contraintes applicables et supprimer les autres contraintes.

De cette façon, on obtient des poids qui donnent des totaux d'îlots justes pour les classes de personnes plus générales mais non pour tous les tableaux croisés.

5.2.2 Méthodes fondées sur l'addition de colonnes (ménages) à A

Lorsque l'inapplicabilité des contraintes n'est que conditionnelle (c'est-à-dire qu'elle dépend des modes de composition des ménages de l'îlot), il suffit d'ajouter des ménages qui ont la composition voulue pour rendre les contraintes applicables. Le moyen le plus simple d'appliquer ce principe est de travailler à un niveau d'agrégation géographique supérieur à celui de l'îlot. Lorsque des difficultés surgissent dans l'ajustement, on peut combiner quelques îlots adjacents ou former une liste au niveau du secteur de dénombrement par exemple, avant la pondération. Plus l'unité sera grande, plus les modes de composition des ménages représentés seront nombreux et moins il y aura de chances d'observer des problèmes d'inapplicabilité.

Une méthode plus poussée consisterait à utiliser un échantillon de ménages prélevé par hot-deck dans des îlots «donneurs» adjacents afin d'élargir le groupe de ménages auxquels un poids peut être attribué. Il importe ici de s'en tenir à un calcul simple puisqu'on pourrait devoir consulter une longue liste de ménages afin de trouver le ou les ménages qui rendront les contraintes applicables. Pour ce qui a trait à la vérification de la cohérence des contraintes, si la ligne j de A est dépendante des lignes précédentes et que la colonne correspondant aux ménages ajoutés est indépendante des colonnes de A (en ce qui a trait uniquement aux j premières lignes), la ligne j de la matrice A enrichie sera indépendante. Pour ce qui a trait à la vérification de l'applicabilité, si l'algorithme est interrompu parce qu'aucune réduction n'est possible dans la fonction objectif $\sum Z_i$, on peut chercher des colonnes de base parmi les colonnes correspondant aux ménages de l'échantillon prélevé par hot-deck. Enfin, si le poids ajusté d'un ménage est extrêmement élevé, on peut chercher parmi les ménages de l'échantillon hot-deck des ménages auxquels on attribuerait aussi des poids élevés avec les valeurs courantes de (c'est-à-dire, colonnes a de telle sorte que $a'\lambda$ est elevé). Si ces ménages sont ajoutés aux autres ménages de l'îlot, ils prendront une partie du poids des ménages surpondérés lorsque les poids seront réajustés puisqu'ils sont susceptibles de compter aussi des membres des mêmes classes de redressement.

Le raisonnement intuitif sur lequel repose cette méthode est que les modes de composition qui ont été observés dans un îlot ne sont qu'un échantillon des modes qui auraient pu y être

observés s'il y avait eu dénombrement complet. On fait le lissage de la distribution observée des modes de composition des ménages en la combinant à la distribution observée pour des îlots adjacents, qui renferment des ménages tout aussi caractéristiques pour ce secteur. Cette méthode se rattache donc théoriquement aux méthodes de lissage baysiennes qui améliorent l'estimation d'une quantité pour une unité en tirant parti de la distribution observée dans des unités semblables. Zaslavsky (1989) intègre ce raisonnement bayesien dans un modèle d'effets aléatoires au niveau de l'îlot.

On pourrait choisir les îlots donneurs par une méthode hot-deck récursive; ces îlots pourraient le plus souvent se trouver près de l'îlot de redressement et aucune série d'îlots en particulier n'aurait une influence anormale sur l'ensemble du recensement. Par une stratification détaillée des îlots, on pourrait faire en sorte de choisir des îlots donneurs qui ressembleraient à l'îlot visé par le redressement au point de vue du revenu moyen, des catégories d'unités de logement et de la répartition raciale.

5.2.3 Combinaison de méthodes

Les deux catégories de méthodes exposées ci-dessus peuvent être combinées par une redéfinition appropriée des contraintes. Le principe, en l'occurrence, est de satisfaire *toutes* les contraintes dans les grandes unités géographiques et de ne satisfaire que les plus importantes dans les petites unités. Ce genre de compromis peut permettre d'obtenir un ajustement assez valide de la distribution voulue sans devoir allonger la liste des ménages.

Supposons que les matrices A de plusieurs îlots ont toutes été réécrites comme des suites de lignes représentant les contraintes principales et les contraintes d'interaction. On peut alors former une seule grande matrice A qui représente toutes les contraintes. Les lignes qui représentent les contraintes principales peuvent demeurer telles quelles tandis que les lignes qui représentent les contraintes secondaires peuvent être combinées dans les îlots. Supposons, par exemple, qu'il y a dix classes de redressement, définies selon le sexe (2 niveaux) et l'âge (5 niveaux), et deux îlots. Il y a en tout 11 contraintes (une pour le nombre de ménages et une pour chaque classe de redressement) dans chaque îlot. Si on combine ces contraintes dans une seule matrice en conservant les effets principaux et les interactions à deux critères, on a les contraintes suivantes: nombre de ménages dans l'îlot (2 contraintes), population de l'îlot (2 contraintes), sexe (1 contrainte), âge (4 contraintes), interaction îlot et sexe (1 contrainte), interaction îlot et âge (4 contraintes) et interaction sexe et âge (4 contraintes) dans les deux îlots. Dans cet exemple, on a éliminé quatre contraintes (interaction îlot $\times$ sexe $\times$ âge); dans un exemple plus réaliste, où il y aurait un plus grand nombre d'îlots, de variables de classification et de niveaux, on pourrait en éliminer beaucoup plus.

6. RÉSULTATS DE SIMULATIONS

Des simulations ont été faites pour répondre à deux séries de questions:

(1) La première série de questions vise à évaluer l'efficacité de l'algorithme en fonction de ses contraintes et de ses objectifs. L'algorithme de repondération donne-t-il une réponse? Dans les problèmes réels, y a-t-il une solution aux contraintes de pondération? De combien les poids varient-ils? La quantité de calculs nécessaires est-elle raisonnable?

Pour répondre à ces questions, nous avons procédé à des «simulations d'applicabilité», où nous avons appliqué l'algorithme de pondération à des îlots simulés composés de vrais ménages en utilisant des taux de redressement réels. Cette méthode reproduit l'application de l'algorithme.

(2) La seconde série de questions vise à évaluer dans quelle mesure l'algorithme améliore la qualité des inférences fondées sur une série de microdonnées: la série de microdonnées pondérées décrit-elle mieux la réalité que les données brutes, non pondérées?

Pour répondre à ces questions, nous avons formé des îlots simulés constitués de vrais ménages pour représenter les modes de composition réels (mais inobservés) des ménages dans les îlots. Pour chaque «vrai» îlot, nous avons fixé un niveau de sous-dénombrement à l'aide de taux de sous-dénombrement réels estimés et d'un modèle plausible pour la répartition du sous-dénombrement entre les ménages. Nous avons appliqué l'algorithme de pondération aux îlots «recensés» ainsi obtenus. Des données sommaires sur la composition des ménages ont été calculées pour les îlots simulés «réels» et les îlots simulés observés où il y avait sous-dénombrement; il s'agissait aussi bien de données non pondérées que de données pondérées pour le redressement en fonction du sous-dénombrement. Ces «simulations d'inférence» avaient pour but de déterminer si la repondération rapprochait les données précitées des valeurs correspondantes dans les îlots «réels»; autrement dit, la repondération corrigeait-elle les biais attribuables au sous-dénombrement?

La source des ménages pour toutes les simulations a été l'échantillon «B» de microdonnées à grande diffusion de 1 % (Public Use Microdata Sample -- PUMS) du recensement de 1980 (Bureau of the Census 1985). Les ménages ont été prélevés dans des secteurs du Los Angeles County en Californie, parmi lesquels se trouve la région ayant servi au Test des opérations de redressement (TOR) du recensement d'essai de 1986.

Les taux de sous-dénombrement étaient ceux qui avaient été calculés dans le TOR de 1986 (Diffendal 1988; tableau 7) pour des classes de redressement définies selon le sexe, l'âge (5 niveaux), l'origine ethnique (hispanique, asiatique ou autres) et le mode d'occupation (propriétaire ou locataire). Les facteurs de redressement calculés à partir de ces taux de sous-dénombrement variaient de 0,982 à 1,211.

Chaque ménage était codé comme un vecteur de chiffres représentant le nombre de membres du ménage qui appartenaient à chacune des 60 classes de redressement.

Zaslavsky (1989) donne des renseignements plus détaillés sur les méthodes de simulation et une autre série de simulations plus vaste.

6.1 Simulations concernant l'applicabilité

Pour l'une et l'autre de quatre tailles d'îlot (20, 50, 100 et 200 ménages), on a prélevé 50 îlots simulés dans l'échantillon complet et on en a formé 50 autres avec les ménages qui ne comptaient aucune personne d'origine asiatique. Pour chaque îlot, les simulations ont été effectuées avec deux valeurs du taux de redressement des ménages (facteur par lequel est redressé le nombre de ménages dans l'îlot).

On a appliqué les algorithmes de la section 3. En résumé, on a tout d'abord vérifié la cohérence des contraintes linéaires, puis leur applicabilité (existence d'une solution positive); enfin, on a calculé les poids à l'aide de la méthode de Newton. Comme il n'y avait pas de données qui permettaient de distinguer le sous-dénombrement des ménages du sous-dénombrement à l'intérieur des ménages, on n'a pas tenté de le faire dans ces simulations ou dans d'autres.

Les résultats des simulations sont résumés dans le tableau 7.

Cohérence et applicabilité:

Les colonnes identifiées «incohér», «inap» et «OK» indiquent le nombre d'îlots simulés (sur les 50 essais) dans chaque simulation, qui faisaient partie de l'une ou l'autre des catégories suivantes: (1) les contraintes étaient incohérentes (ne pouvaient être satisfaites par aucun poids), (2) les contraintes étaient cohérentes mais inapplicables (ne pouvaient être satisfaites par aucun poids positif), ou (3) les contraintes étaient à la fois cohérentes et applicables.

Dans les simulations portant sur les îlots où il n'y avait aucune personne d'origine asiatique, il fallait satisfaire 41 contraintes (dont certaines pouvaient être banales, c'est-à-dire lorsque les classes de redressement correspondantes n'étaient pas représentées dans l'îlot). Ainsi, pour ce qui est des îlots de 20 ménages, les contraintes n'étaient jamais cohérentes; dans le cas des îlots de 50 ménages, les contraintes étaient parfois cohérentes et, en règle générale, applicables.

Elles étaient habituellement applicables pour ce qui est des îlots de 100 ménages et toujours applicables dans le cas des îlots de 200 ménages.

Les colonnes numérotées à droite représentent l'ordre de la plus simple contrainte marginale qui n'a pu être satisfaite au sens de la reparamétrisation hiérarchique définie dans la section 5.2.1. Ainsi, la colonne (1) indique le nombre d'îlots simulés pour lesquels une contrainte d'«effet principal» (total marginal des personnes classées selon une variable de stratification) n'a pu être satisfaite, la colonne (2) indique le nombre d'essais pour lesquels une contrainte d'interaction à deux critères n'a pu être satisfaite, etc. Même lorsque les contraintes étaient incohérentes dans le cas d'îlots de 50 ou 100 ménages, les contraintes d'effet principal et, souvent, les contraintes d'interaction à deux ou même à trois critères étaient applicables. Cela donne à penser que le groupement d'îlots pour des interactions d'ordre supérieur, comme nous l'avons vu dans la section 5.2.3, pourrait être une solution efficace aux problèmes d'inapplicabilité. Les résultats ont été moins encourageants dans le cas des simulations qui portaient sur les échantillons complets. Même pour des îlots de 200 ménages, les contraintes étaient rarement cohérentes et applicables. À mesure qu'augmentait la taille de l'îlot, les contraintes d'ordre inférieur étaient plus susceptibles d'être applicables. Cela s'explique par le faible nombre de ménages qui comptent des personnes d'origine asiatique (environ 5 % dans chaque échantillon). Sur 200 ménages, on s'attendrait à en avoir environ 10 qui comptent des personnes d'origine asiatique, ce qui est insuffisant pour satisfaire les 20 contraintes possibles pour les classes de redressement relatives aux personnes d'origine asiatique. Dans la réalité, il n'est sûrement pas rare de constater que des groupes de classes de redressement sont mal représentés dans une région ou des îlots en particulier. Il faudra alors procéder à un groupement d'îlots à grande échelle pour les contraintes correspondantes tout en essayant de satisfaire à une plus petite échelle les contraintes se rapportant aux classes mieux représentées.

Poids:

Les poids maximum et minimum des ménages de même que leur variance ont été calculés pour chaque îlot simulé pour lequel les contraintes étaient cohérentes et applicables. Le

Tableau 7

Résultats de la simulation concernant l'applicabilité

Ménages ne comptant aucune personne d'origine asiatique

Taille	HH	Taux	incohér	inap	OK	Pmax	Pmin	varP	itér	(1)	(2)	(3)	(4)
10		1.00	50	0	0	NA	NA	NA	NA	22	28	0	0
10		1.05	50	0	0	NA	NA	NA	NA	8	42	0	0
20		1.00	50	0	0	NA	NA	NA	NA	0	50	0	0
20		1.05	50	0	0	NA	NA	NA	NA	0	50	0	0
50		1.00	47	1	2	1.921	0.200	0.142	3.00	0	3	37	8
50		1.05	47	0	3	1.550	0.620	0.036	1.33	0	3	36	8
100		1.00	10	0	40	2.068	0.429	0.088	2.03	0	0	8	2
100		1.05	10	0	40	1.573	0.753	0.020	1.90	0	0	8	2
200		1.00	0	0	50	2.434	0.543	0.063	2.18	0	0	0	0
200		1.05	0	0	50	1.749	0.821	0.015	2.00	0	0	0	0

Échantillon complet

Taille	HH	Taux	incohér	inap	OK	Pmax	Pmin	varP	itér	(1)	(2)	(3)	(4)
100		1.00	49	0	1	--	--	--	--	0	34	15	0
200		1.00	49	0	1	--	--	--	--	0	2	43	4

tableau 7 indique, pour chaque condition de simulation, la valeur moyenne de ces quantités (pour l'ensemble des îlots simulés) dans les colonnes «Pmax», «Pmin» et «varP». Les observations ci-dessous décrivent quelques-uns des effets des facteurs du plan de simulation sur les poids ajustés.

(1) L'utilisation d'un facteur de redressement du nombre de ménages de 1,05 (par rapport à un facteur de (1) donne dans chaque cas une variance moyenne des poids inférieure et une moyenne des poids minimum et des poids maximum plus près de 1. Ces résultats sont acceptables intuitivement car presque tous les facteurs de redressement pour les classes sont supérieurs à 1 et qu'il faut procéder à un redressement plus prononcé lorsqu'on ajoute des personnes dans des ménages existants que lorsqu'on ajoute simplement des personnes et des ménages pour les recevoir. Par exemple, si les facteurs de redressement pour les ménages et chaque classe de redressement étaient tous égaux, chaque ménage serait pondéré à la hausse également.

(2) Les autres facteurs étant fixes, la variance des poids diminue à mesure qu'augmente le nombre de ménages par îlot. Ce résultat est aussi acceptable intuitivement car le groupe des ménages est plus diversifié dans un îlot plus grand; la probabilité de trouver exactement les ménages voulus pour représenter les personnes qui n'ont pas été dénombrées est plus élevée. Les tendances des poids extrêmes se dégagent moins clairement que celles des variances; ici, la réduction de la variance est contrebalancée par l'échantillon plus grand par rapport auquel est calculé le poids extrême dans les îlots plus grands.

(3) En ce qui a trait aux simulations concernant les îlots de 200 ménages, la variance moyenne n'a pas dépassé 0,063. Il est donc permis de conclure qu'en règle générale, la repondération ne produit pas de poids extrême.

Coût des calculs:

Le nombre moyen d'itérations de Newton nécessaires pour ajuster les poids (sur la base des valeurs initiales définies dans la section 3.3) est habituellement de 2; le nombre d'itérations figure dans la colonne «itér» du tableau 7. Deux itérations ont suffi pour satisfaire toutes les contraintes avec une marge d'erreur n'excédant pas 0,001. À l'aide de cette information, il est possible d'estimer grossièrement le nombre d'opérations en virgule flottante nécessaires pour appliquer l'algorithme. Le coût d'application de l'algorithme d'itération modifié est analysé dans la section 10.2.

Supposons que les îlots sont de taille suffisamment grande pour ne pas avoir à vérifier la cohérence et l'applicabilité des contraintes dans chaque cas (sauf peut-être lorsque l'ajustement de poids ne peut être accompli en quelques itérations). L'opération essentielle est donc d'ajuster les poids. Pour les phases d'exécution, les programmes et les structures de données devraient être conçus de manière à tirer profit de la faible densité de la matrice A (qui tient au fait que seules quelques classes sont représentées dans chaque ménage). Si S_1 est le nombre total d'éléments non nuls dans A et S_2 est la somme (pour l'îlot) des *carrés* du nombre d'éléments non nuls pour chaque ménage, chaque itération de Newton exige environ $5S_1/2 + S_2/2$ multiplications (plus un terme indépendant du nombre de ménages par îlot). Dans les échantillons étudiés ici, $S_2 \approx 5S_1$; S_1 est borné par la population totale de l'îlot. Par conséquent, le nombre maximum de multiplications est environ de $15 \times$ total de la population (si l'on compte l'étape de départ comme une itération); le nombre maximum d'additions est comparable.

À une époque où même les micro-ordinateurs ont une puissance arithmétique qui se mesure en mégaflops, il ne semble pas déraisonnable d'exécuter 8×10^9 opérations en virgule flottante pour repondérer les résultats d'un recensement tout entier. Le calcul des poids pourrait exiger moins de ressources que le traitement de données «administratives», qui accompagne toute méthode d'analyse du sous-dénombrement. Évidemment, si la méthode était appliquée à un échantillon de base de données, comme une bande-échantillon à grande diffusion, le coût baisserait en conséquence.

6.2 Simulations d'inférence

Pour les simulations d'inférence, on a prélevé des pseudo-îlots de 50 ménages comptant uniquement des personnes d'origine hispanique. Ces pseudo-îlots ont été traités comme s'ils représentaient de vrais îlots. On a ensuite déterminé un niveau de sous-dénombrement simulé pour ces ménages en supposant que la probabilité qu'un membre du ménage soit oublié (indépendamment des autres) était égale au taux de sous-dénombrement calculé par Diffendal (1988), avec deux taux de sous-dénombrement négatifs ramenés à 0.

On a voulu représenter la distribution des compositions de l'îlot «recensé» en incluant dans la pseudo-liste de recensement la composition réelle des ménages et les compositions que l'on pouvait obtenir si un ou plusieurs membres du ménage étaient oubliés, chaque composition étant pondérée par la probabilité prévue dans le modèle.

On a ensuite repondéré la pseudo-liste de recensement avec sous-dénombrement en fonction des totaux initiaux du pseudo-îlot pour le nombre de ménages et le nombre de personnes dans chaque classe de redressement. Enfin, on a comparé la pseudo-liste de recensement et la liste repondérée au pseudo-îlot original.

La simulation a été effectuée ainsi de manière à éliminer la variabilité attribuable au caractère aléatoire du taux de sous-dénombrement d'un îlot (autour du taux de sous-dénombrement moyen) et de la répartition du nombre de personnes oubliées entre les ménages de l'îlot. En outre, l'applicabilité est garantie car les ménages initiaux sont toujours inclus (avec les poids) dans la pseudo-liste de recensement. L'interprétation que l'on peut donner de cela est que chaque îlot simulé représente une très grande population où les taux de sous-dénombrement observés et la distribution des compositions observées tendent vers leur espérance mathématique.

Plusieurs séries de données ont servi à analyser la méthode de repondération. Elles ont toutes été choisies parce qu'elles résumaient des caractéristiques du ménage qui ne sont pas fonction du nombre de personnes dans les classes de redressement. La première série de données concernait la distribution des tailles (nombre de membres) des ménages. Il convient de souligner que le nombre moyen de personnes par ménage, comme toute autre fonction des totaux de classe et du nombre de ménages, sera redressé automatiquement en fonction des valeurs exactes (avant sous-dénombrement); cependant, la distribution des tailles des ménages n'est pas soumise à ce redressement.

La seconde série de données portait sur la distribution du nombre d'*adultes* (plus de 14 ans) dans les ménages qui comptent un ou plusieurs *enfants* (14 ans et moins). Dans ce cas, la moyenne n'est pas redressée automatiquement en fonction de la valeur exacte puisqu'elle dépend des totaux marginaux et de la distribution conjointe des effectifs de diverses classes à l'intérieur des ménages.

Les deux dernières séries de données portaient sur la distribution du groupe d'âge (codes de 1 à 5 comme pour la formation des classes de redressement) de l'*homme le plus âgé* du ménage (code 0 si le ménage ne compte pas d'homme) et sur la même distribution pour les ménages qui comptaient un ou plusieurs enfants. Là encore, ni la distribution ni sa moyenne ne correspondant directement à leur valeur réelle par contrainte.

Le tableau 8 donne un résumé des résultats de ces simulations. Comme presque tous les écarts indiqués ici sont très significatifs (par rapport aux variances inter-pseudo-îlot des écarts), les erreurs types ne figurent pas dans les tableaux. Dans chaque tableau, il y a une ligne «réel» (pour les pseudo-îlots originaux), une ligne «recens» (pour les îlots recensés simulés, c.-à-d. compte tenu des omissions dues au sous-dénombrement) et une ligne «redress» (îlots recensés, après redressement pour sous-dénombrement). Les chiffres de chaque colonne, sauf celle des moyennes, représentent le pourcentage de ménages dans l'îlot.

La distribution des tailles des ménages était biaisée à la baisse dans les îlots recensés. Le redressement a eu pour effet non seulement de corriger la moyenne mais aussi de rapprocher sensiblement le pourcentage estimé pour chaque taille du pourcentage réel.

La distribution du nombre d'adultes dans les ménages comptant un ou plusieurs enfants était aussi biaisée à la baisse. Comme la majorité de es ménages comptaient deux adultes, ce

groupe de taille était celui qui avait été le plus sous-estimé par les données du recensement. Toutefois, étant donné la structure linéaire logarithmique du redressement, les corrections les plus importantes ont été apportées aux ménages ayant le plus de personnes et le moins de personnes. Par conséquent, les groupes de taille supérieurs ont été sur-redressés légèrement et les groupes intermédiaires ont été sous-redressés; le groupe des ménages de «taille 2» a été redressé légèrement dans la mauvaise direction. Néanmoins, la moyenne de la distribution redressée était beaucoup plus près de la valeur «réelle» que la moyenne redressée.

On observe la même chose dans le cas de la distribution du groupe d'âge de l'homme le plus âgé du ménage. Bien que ces données n'ont qu'un lien indirect avec les effectifs de classes, les distributions et les moyennes redressées reflètent mieux la «réalité» que les distributions et les moyennes non redressées dans presque tous les cas.

En définitive, ces simulations donnent à penser que le redressement par pondération peut améliorer les estimations des mesures de la structure des ménages de même que les données agrégées auxquelles il s'adresse. Toutefois, la repondération produit de moins bons résultats avec certaines formes de données, comme dans le cas des nombreux ménages qui comptent deux adultes; dans ce cas, on pourrait devoir utiliser une méthode d'imputation axée sur un modèle, comme celle décrite par Zaslavsky (1989).

Tableau 8

Résultats de la simulation d'inférence

Répartition des ménages selon la taille

	taille 1	taille 2	taille 3	taille 4	taille 5+	moyenne
réel	7.240	16.200	20.240	22.600	33.720	3.971
recens	10.349	19.631	21.772	20.690	27.558	3.632
redress	7.372	16.421	20.596	21.392	34.219	3.971

Répartition des ménages avec enfants selon la taille (nombre d'adultes)

	taille 0	taille 1	taille 2	taille 3	taille 4	taille 5+	moyenne
réel	0.000	6.925	58.404	17.214	9.125	8.332	2.585
recens	1.736	18.309	49.874	15.965	7.677	6.439	2.323
redress	0.924	13.277	48.557	18.223	9.810	9.209	2.562

Âge de l'homme le plus âgé (selon 5 groupes d'âge)

	aucun	groupe 1	groupe 2	groupe 3	groupe 4	groupe 5	moyenne
réel	7.080	4.000	28.680	33.800	21.960	4.480	2.730
recens	9.981	7.388	26.296	30.972	21.160	4.203	2.585
redress	7.853	5.989	26.307	33.439	21.931	4.480	2.690

Âge de l'homme le plus âgé (selon 5 groupes d'âges) pour les ménages avec enfants

	aucun	groupe 1	groupe 2	groupe 3	groupe 4	groupe 5	moyenne
réel	3.602	6.214	30.744	42.649	15.843	0.949	2.638
recens	5.809	11.723	27.321	39.096	15.158	0.894	2.488
redress	4.272	9.069	27.242	42.038	16.418	0.962	2.601

7. UTILISATIONS DE DONNÉES PONDÉRÉES

Le résultat des méthodes décrites dans les sections précédentes serait une liste de recensement dans laquelle les ménages auraient des poids, les membres de ces ménages auraient des poids identiques à celui du ménage et les pensionnaires d'établissement institutionnel auraient des poids qui leur auraient été attribués individuellement. Dans cette section, nous voyons comment ces listes servent aux fins du recensement.

7.1 Formation de tableaux de chiffres de population

Comme pour n'importe quelle série de données pondérées, la somme de poids remplace le nombre d'observations dans la création de tableaux. La seule difficulté liée à l'utilisation de poids est d'obtenir des nombres entiers dans les tableaux. Cette difficulté surgit même avant le calcul des poids des ménages: en effet, lorsqu'on estime le nombre de personnes ou de ménages oubliés, les effectifs des classes ne sont ordinairement pas exprimés en nombres entiers.

Si les totaux de classe redressés sont arrondis, un tableau qui fait la somme de plusieurs classes (par exemple, un nombre d'adultes de sexe masculin, qui est la somme des adultes de sexe masculin de diverses classes) contiendra aussi des entiers puisqu'il doit être conforme à ces totaux. En ce qui a trait aux tableaux qui ne reposent pas sur les totaux redressés, l'addition de poids dans un groupe particulier ne donnera pas nécessairement des nombres entiers. Par exemple, si une classe comprend des femmes de 20 à 40 ans, la somme des poids se rattachant aux femmes de 20 à 30 ans ne sera pas nécessairement un nombre entier. Quoi qu'il en soit, il serait peu vraisemblable d'arrondir tous les poids se rapportant à des classes puisque l'erreur d'arrondissement pourrait bien être du même ordre que les corrections. Toutefois, on devrait pouvoir se servir des méthodes d'arrondissement actuellement en usage au Census Bureau («arrondissement contrôlé») pour résoudre la difficulté, surtout lorsque les règles de protection du secret statistique exigent que les données destinées à la publication soient arrondies de toutes façons (Cox et coll. 1986; Cox 1987).

7.2 Formation de tableaux de sommes et de moyennes

Comme les sommes (de quantités continues) et les moyennes nc sont habituellement pas des nombres entiers, la question de l'arrondissement n'est pas pertinente ici. En outre, puisque les tableaux fondés sur les données de questionnaires complets proviennent déjà d'un échantillon, une source de poids additionnelle changerait peu de choses au processus. Ce à quoi il faut surtout s'intéresser ici sont les valeurs des covariables non classificatoires qu'il faut attribuer aux ménages qui sont «rajoutés» dans le recensement; cette question est traitée dans la section 8.

7.3 Bandes-échantillon à grande diffusion

Les bandes à grande diffusion sont un échantillon d'enregistrements du recensement qui sont mis à la disposition des utilisateurs à des fins d'analyse.

Pour produire ces bandes-échantillon à partir de listes pondérées du recensement, il suffit de modifier légèrement la méthode d'échantillonnage de manière que les probabilités d'échantillonnage soient proportionnelles aux poids. Même pour la bande à 5% (taux de sondage le plus élevé), les probabilités d'échantillonnage pondérées devraient être inférieures à 1. Une fois que ces bandes sont produites, l'utilisateur n'est pas obligé de connaître les méthodes de redressement et de pondération qui ont servi à la production de ces bandes.

Les bandes à grande diffusion sont une source de données précieuse pour les analyses complexes réalisées par les sociologues, les économistes, les planificateurs, *etc.*, et dans lesquelles la composition des ménages et les chiffres de population ont de l'importance. Il est essentiel que ces bandes puissent être produites facilement et utilisées comme des données brutes du recensement.

Afin de satisfaire les utilisateurs qui aimeraient vérifier la sensibilité de leurs analyses au redressement pour sous-dénombrement, la bande devrait contenir des coefficients (notamment l'inverse des poids de redressement rattachés aux enregistrements de ménages qui figurent dans les listes originales du recensement) qui permettraient à l'utilisateur de reproduire les données non redressées du recensement.

8. REDRESSEMENT DES COVARIABLES QUI NE SERVENT PAS À LA CLASSIFICATION

Les méthodes décrites ci-dessus garantissent que les totaux pondérés d'îlots selon des variables classificatoires comme le sexe, l'origine ethnique et le groupe d'âge, égaleront les totaux redressés d'îlots. Toutefois, ces listes serviront aussi à accumuler des totaux ou des chiffres de population pour des variables qui ne serviraient pas à la classification, par exemple le revenu et le niveau d'instruction. Dans cette section, nous examinons l'effet de ces méthodes de redressement sur des données de ce genre. Pour que l'illustration soit des plus concrètes, nous nous servirons du revenu comme exemple principal. Le revenu est une variable non classificatoire importante; des études permettent de croire que les programmes de répartition du revenu peuvent être très touchés par des erreurs de mesure du revenu. (National Academy of Sciences 1985).

En règle générale, il y a deux sources de biais possibles dans l'estimation d'une covariable non classificatoire: (1) biais dans le redressement de la composition du ménage, et (2) écarts systématiques entre les ménages recensés entièrement et les ménages de même composition qui ont été oubliés (en tout ou en partie). Cependant, si nous disposons d'une estimation du revenu moyen pour l'îlot, nous pouvons établir une relation d'égalité entre la moyenne pondérée pour les ménages de l'îlot et la moyenne estimée (redressée), de la même façon que nous le faisons pour le nombre pondéré de personnes dans l'îlot et leur nombre estimé (redressé).

8.1 Biais dû à la composition des ménages

Dans cette section, nous supposons que le niveau de revenu moyen qui se rattache à un certain mode de composition des ménages est le même pour les ménages qui ont été recensés entièrement et ceux qui ont été oubliés en tout ou en partie. Autrement dit, nous considérons ici le cas où le sous-dénombrement n'a aucun lien avec le revenu.

Supposons que le revenu du ménage est la somme des contributions indépendantes des membres du ménage répartis dans les diverses classes (autrement dit, supposons que la contribution des membres d'une classe est indépendante de la situation des autres membres dans le ménage). Le revenu total pondéré des ménages sera alors une estimation sans biais du revenu total réel (lorsque les taux de redressement sont exacts) puisque la somme des revenus est une fonction linéaire des effectifs de classe pour l'îlot. Toutefois, selon l'hypothèse plus réaliste de non-linéarité, une mauvaise répartition des personnes entre les ménages (et par conséquent une mauvaise représentation de la composition des ménages dans le redressement) pourrait introduire un biais dans les estimations du revenu. Ainsi par exemple, le revenu moyen des ménages qui comptent deux enfants pourrait ne pas correspondre à la moyenne des revenus moyens des ménages à un et à trois enfants (étant donné la même composition d'adultes). En conséquence, la pondération pourrait donner le bon nombre d'enfants mais si, en moyenne, on créait un trop grand nombre de ménages à deux enfants (comparativement à la réalité) par rapport au nombre de ménages à un ou à trois enfants, les estimations du revenu des ménages seraient biaisées.

Notre méthode tend à ajuster les poids qui donnent aux ménages «rajoutés» une composition semblable à celle des ménages recensés. Cependant, le redressement n'est décrit que par les totaux des classes de redressement, qui en disent peu sur la composition des ménages oubliés.

Par conséquent, si les ménages ayant un mode de composition donné sont sous-dénombrés dans une plus forte proportion que d'autres, ils peuvent être sous-représentés dans les listes pondérées et s'ils sont associés, par exemple, à un faible niveau de revenu, les estimations du revenu total seront biaisées par excès.

Le problème consiste essentiellement en un manque d'ajustement probable du modèle aux tendances des données. Les biais les plus prononcés pourraient se trouver dans les statistiques qui ont trait directement à la composition des ménages, comme le nombre de familles monoparentales.

Si le biais dû à la composition des ménages devait poser un problème sérieux, on pourrait tenter de le réduire en augmentant les taux de redressement des classes à l'aide de données additionnelles sur le taux de sous-dénombrement combiné des personnes de diverses classes (ou de classes groupées).

8.2 Biais de réponse

Étant donné un groupe de ménages ayant la même composition, il n'est pas déraisonnable de croire que ceux qui ont été oubliés dans le recensement diffèreront systématiquement, à certains points de vue, de ceux qui ont été recensés. Autrement dit, le sous-dénombrement peut être une forme de non-réponse avec biais. Par exemple, les ménages caractérisés par un faible revenu et un niveau d'instruction inférieur peuvent être plus susceptibles d'être oubliés entièrement ou partiellement; le revenu et le niveau d'instruction n'étant pas des variables classificatoires, elles ne font donc pas l'objet d'un redressement direct.

Selon les méthodes proposées, le *redressement de ménages entiers* s'effectue par une pondération à la hausse des ménages, sans modification des valeurs des covariables. On suppose implicitement que les ménages oubliés ne diffèrent pas, au point de vue de ces covariables, des ménages recensés qui ont la même composition. Aucune des données relatives à l'îlot faisant l'objet du redressement ne vient réfuter cette hypothèse. Toutefois, l'EP (enquête postcensitaire) devrait pouvoir nous renseigner sur les différences entre les ménages recensés et les ménages oubliés, dont nous pourrions tenir compte dans le redressement. Par exemple, on pourrait définir le rapport entre le revenu des ménages oubliés et le revenu moyen des ménages recensés ayant la même composition au moyen d'une régression linéaire; on pourrait ensuite imputer aux ménages additionnels (rajoutés) le revenu obtenu en appliquant la fonction de régression linéaire au revenu du ménage donneur recensé. Little et Rubin (1987) examinent des méthodes conçues pour résoudre les problèmes de données manquantes avec non-réponse informative. Dans la section suivante, nous décrivons une autre méthode intégrée à la méthode de redressement par pondération.

Le *redressement du nombre de personnes à l'intérieur de ménages* consiste à pondérer à la baisse un ménage ayant certaines caractéristiques recensées et à pondérer à la hausse un autre ménage qui compte un ou plusieurs membres additionnels. S'il n'y a pas d'autre redressement, ce sont les caractéristiques du ménage pondéré à la hausse, et non celles du ménage recensé qui a fourni le poids, qui s'appliqueront à la composante «rajoutée».

Cette situation pose des problèmes que l'on ne peut résoudre sans recueillir certaines données (dans un sous-échantillon de l'EP). Par exemple, si un *enfant* ne figure pas par erreur sur la liste des membres d'un ménage, il n'y a pas de raison de croire que cet oubli faussera le revenu du ménage. Si les ménages qui comptent relativement plus d'enfants avaient un revenu moyen plus élevé que celui des ménages ayant relativement moins d'enfants, la pondération tendrait à sur-estimer les revenus moyens.

Si un *adulte* était oublié, il serait permis de croire que le revenu de cet adulte (s'il en a un) ne serait pas inclus dans le revenu déclaré du ménage. Il est plausible de penser que le revenu moyen non déclaré dans un tel cas serait positif mais inférieur au revenu moyen d'un ménage de même composition, dont aucun des membres adultes n'aurait été oublié. Prenons l'exemple classique d'une famille qui vit d'aide sociale et ne déclare pas un adulte de sexe masculin, dont le domicile est plutôt instable et dont le revenu serait déduit du montant d'aide sociale si cette

personne était déclarée. Le revenu de cette personne est susceptible d'être moins élevé que celui d'un membre adulte permanent d'une famille qui ne dépend pas de l'aide sociale. Par conséquent, ni le revenu du ménage recensé ni celui du ménage «rajouté» ne constituent une valeur satisfaisante pour l'imputation du revenu du ménage redressé.

Comme aucune correspondance directe n'est établie entre les ménages qui sont pondérés à la baisse et ceux qui reçoivent un poids additionnel, il n'est pas possible de reporter directement sur un ménage «rajouté» le revenu non redressé d'un ménage recensé. Toutefois, si l'on entreprenait des études visant à comparer les revenus des ménages recensés et des ménages oubliés, on pourrait se servir des revenus des ménages pondérés à la baisse pour le redressement des revenus. Par exemple, on pourrait poser la condition que le revenu moyen des ménages de l'îlot repondéré égale le revenu moyen des ménages de l'îlot avant redressement.

8.3 Redressement par pondération des caractéristiques non classificatoires

Supposons que nous disposions de données sommaires redressées (par îlot) sur quelques caractéristiques des ménages, exception faite des effectifs des classes de redressement. Par exemple, un modèle de régression pourrait nous avoir fourni une estimation redressée du revenu moyen ou de la proportion des familles monoparentales. Dans la mesure où l'on peut représenter les données sommaires comme une somme pondérée des valeurs de covariables pour chaque ménage, il est possible d'obtenir la conformité avec la valeur redressée voulue en définissant une liaison linéraire pour les poids, qui peut être intégrée à la méthode de redressement par pondération étudiée dans cet article. Par conséquent, si nous reprenons l'exemple du revenu, nous définirions la contrainte de manière que la somme pondérée des revenus égale le produit du nombre de ménages par le revenu moyen redressé. Pour redresser la proportion de familles monoparentales, nous définirions la contrainte de manière que la somme pondérée des indicateurs 0-1 pour cette caractéristique égale le nombre total voulu.

8.4 Résumé et incidences

La méthode proposée permettra de pondérer à la hausse des ménages et de fonder les analyses uniquement sur les caractéristiques de ces ménages sans que l'on s'interroge sur la possibilité d'un biais. Si le redressement et les biais introduits dans les caractéristiques des ménages sont de faible importance, le biais global contenu dans les caractéristiques estimées de l'îlot sera de second ordre et ne devrait pas poser de problème majeur. On pourrait peut-être réduire davantage les biais en effectuant quelques ajustements par régression simple.

9. PROPOSITIONS POUR ORIENTER LA RECHERCHE ET PERFECTIONNER LA MÉTHODE

Cette section résume un certain nombre de propositions concernant l'application et le perfectionnement de la méthode de redressement étudiée dans cet article.

9.1 Collecte de données et modélisation statistique dans le cadre d'une enquête postcensitaire (EP)

L'EP devrait distinguer le sous-dénombrement des personnes dans les ménages recensés du sous-dénombrement des personnes dans les ménages oubliés et les deux taux de sous-dénombrement devraient faire l'objet de modèles distincts pour chaque classe de redressement. Les taux de sous-dénombrement des ménages devraient aussi faire l'objet de modèles. (Section 4). Une série de mesures (comme dans la Section 6.2) pourraient servir à comparer la composition des ménages «rajoutés» et celle des ménages oubliés qui ont été relevés dans l'EP; si

les études révélaient que le «biais dû à la composition» représente un problème appréciable, il faudrait élaborer des paramètres statistiques plus raffinés (Section 8.1). On devrait demander à un échantillon de ménages de l'EP qui ont été oubliés dans le recensement de répondre au questionnaire complet de manière à pouvoir mettre en modèle pour le redressement (Sections 8.2, 8.3) le rapport entre le sous-dénombrement et des covariables comme le revenu et le niveau d'instruction.

9.2 Applicabilité des redressements

On devrait tester et comparer les méthodes de la Section 5 à l'aide de données de l'EP.

9.3 Imputation multiple

Bien que les méthodes proposées dans cet articles soient de nature déterministe, les statistiques fondées sur des enregistrements pondérés renferment toujours un certain nombre de sources d'incertitude, comme l'incertitude dans l'estimation des taux de sous-dénombrement, la variabilité des taux de sous-dénombrement de classes d'un îlot à l'autre par rapport à la moyenne nationale, la variabilité binomiale du nombre réel de personnes ou de ménages oubliés par rapport au nombre prévu, étant donné le taux de sous-dénombrement, et l'incertitude concernant les écarts entre les valeurs de covariables pour les ménages oubliés et celles pour les ménages recensés qui sont pondérés à la hausse en vue de remplacer les premiers.

À des fins de recherche, on pourrait préparer des fichiers qui représenteraient toutes ces formes d'incertitude par l'imputation multiple (Rubin 1987). On pourrait représenter deux versions (ou plus) de la série de données repondérées en incluant plusieurs séries de poids dans le fichier. Les chercheurs pourraient reprendre leurs analyses en se servant à chaque fois d'une nouvelle série de poids. La variabilité des résultats des diverses versions donnerait une estimation de la variabilité attribuable au redressement pour sous-dénombrement. Zaslavsky (1989) analyse des méthodes d'imputation multiple dans ce contexte.

10. MÉTHODES ADDITIONNELLES

10.1 Choix d'une fonction objectif pour la pondération

On a proposé un certain nombre de fonctions objectif pour déterminer un tableau de données ajustées optimales (il s'agit le plus souvent de tableaux de contingence, voir Fagan et Greenberg 1988). Dans chaque cas, la fonction est de la forme $T = \sum T_1(W_i)$, où T_1 prend une des formes indiquées dans le tableau 9. On peut ramener chacune de ces fonctions à une fonction équivalente T_0 en multipliant par un coefficient constant et en ajoutant une fonction linéaire de W de sorte que $T_0(1) = 0$, $T_0'(1) = 0$, $T_0''(1) = 1$. Comme $\sum W_i$ est limitée à une valeur donnée, les poids optimaux ne sont pas modifiés. Les deux premiers termes des fonctions objectif normalisées développées en une série de Taylor de centre 1 sont les mêmes dans chaque cas et ces fonctions devraient produire des résultats comparables lorsque les poids se rapprocheront de 1. Par contre, elles diffèrent l'une de l'autre pour ce qui a trait au degré d'asymétrie entre les coûts de la pondération à la baisse et ceux de la pondération à la hausse, ce degré d'asymétrie étant défini par l'exposant de W dans la formule de dérivée seconde, $T_0''(W) = W^{-k}$. La méthode des moindres carrés ($k = 0$) traite la pondération à la hausse et à la baisse de façon parfaitement symétrique et, de ce fait, peut produire des poids nuls ou négatifs. À mesure que k augmente, le coût de la pondération à la hausse diminue par rapport à celui de la pondération à la baisse. Toutes les autres fonctions objectif ($k > 0$) produisent un poids positif pour chaque élément des données brutes; cela est évident dans le cas de la méthode itérative du quotient, si l'on en juge par la forme des poids illustrée dans la Section 3.3. Si l'on préfère ici la méthode itérative du quotient à la méthode du maximum de

Tableau 9

Comparaison de fonctions objectif pour l'ajustement de tableaux

Nom de la méthode d'ajustement	Fonction objectif $T_I(W)$, forme usuelle	Fonction objectif $T_0(W)$, uniformisée	Dérivée seconde, $T_0'(W)$
Moindres carrés (variance minimum)	$(W - 1)^2$	$(W - 1)^2/2$	1
Méthode itérative du quotien	$W \log W$	$(W \log W) - W + 1$	$1/W$
Maximum de vraisemblance	$- \log W$	$W - 1 - \log W$	$1/W^2$
χ^2 minimum	$(W - 1)^2/W$	$(W - 1)^2/2W$	$1/W^3$

vraisemblance ou à celle du χ^2 minimum, c'est surtout à cause de la forme élémentaire de sa solution et de l'analogie avec l'itération dans les tableaux de contingence. Cressie et Read (1984) étudient systématiquement les propriétés de cette catégorie de mesures d'ajustement.

10.2 Méthode de la descente cyclique pour l'ajustement de poids

Dans cette section, nous présentons une méthode d'ajustement analogue à l'ajustement proportionnel itératif (API) appliqué dans les tableaux de contingence. Selon l'API, les fréquences de case sont transformées par multiplication de manière à ne pas modifier les produits croisés (condition nécessaire pour minimiser la fonction objectif) et à rendre le tableau compatible avec chaque série de contraintes marginales. L'algorithme converge vers un tableau qui satisfait toutes les contraintes, et préserve nécessairement les produits croisés. (Bishop, Feinberg et Holland 1974; Ireland et Kullback 1968).

Selon la méthode proposée, les poids doivent avoir la forme linéaire logarithmique $W_i = \exp(a_i'\lambda - 1)$ définie dans la Section 3.3, et satisfaire en même temps les contraintes $AW = B$. Dans cet exposé, nous supposerons que la contrainte de poids total $\sum W_i = H$ est exclue de $AW = B$ et que A est de dimension p (contraintes) $\times$ I (nombre de modes de composition des ménages). Nous allons exécuter une série d'étapes dans lesquelles nous multiplierons à chaque fois chacun des poids W_i par $c\rho^{a_{ji}}$ pour obtenir un nouveau poids W_i', ce qui aura pour conséquence de conserver la structure linéaire logarithmique; c et p sont choisies de manière que les contraintes $\sum W_i' = H$ et $\sum W_i' a_{ji} = b_j$ soient satisfaites. Si nous procédons de façon cyclique, de manière que $j = 1,2, \ldots p$ désigne les contraintes l'une après l'autre, lalgorithme convergera éventuellement vers des poids qui satisfont toutes les contraintes.

À l'étape j du cycle t, les nouveaux poids sont définis $W_i^{(t,j)} = cp^{aji}W_i^{(t,j-1)}$ (initialisés pour $j = 1$ par lutilisation des derniers poids du dernier cycle, $W_i^{(t,0)} = W_i^{(t-1,p)}$). Par conséquent c et ρ doivent satisfaire les équations suivantes:

$$\sum_i cp^{a_{ji}}W^{(t,j-1)} = H, \quad \sum_i a_{ji}c\rho^{a_{ji}}W_i^{(t,j-1)} = b_j. \tag{5}$$

Si nous supprimons c de ces équations, ρ est une racine de

$$\sum_i \left(Ha_{ji} - b_j\right) W_i^{(t,j-1)}p^{a_{ji}} = 0. \tag{6}$$

Nous devons avoir $Ha_{j,min} \leq b_j \leq Ha_{j,max}$ où $a_{j,min}$ et $a_{j,max}$ sont respectivement les valeurs minimum et maximum de a_{ji}. Si tel n'était pas le cas, aucun poids ne pourrait satisfaire la contrainte j. Il doit donc y avoir au moins une racine ρ, et si les a_{ji} sont non négatifs, l'expression s'accroît en ρ et cette racine est par conséquent unique. On détermine alors la valeur réelle de ρ par la méthode de Newton ou par une expression en forme analytique pour les racines d'un polynome (puisque dans le cas de la matrice A, a_{ji} désigne le nombre de membres de la classe j dans un ménage, nombre qui dépasse rarement 2).

Bien que nous n'ayons pas encore démontré que cet algorithme converge constamment, nous avons constaté son efficacité dans la pratique. Cet algorithme n'exige pas d'inversion de matrice et si les a_{ji} sont de petits nombres entiers, le nouveau calcul des poids à chaque étape n'exige que le calcul de quelques puissances en nombre entier. De plus, si certaines contraintes ont la forme de fréquences marginales simples, le redressement dans ce cas se fait par l'itération classique.

Si la matrice A originale est utilisée, la méthode peut tirer parti de la faible densité de A (attribuable au fait que seulement quelques classes sont représentées dans chaque ménage). À chaque étape (par exemple, redressement en vue de l'ajustement à la fréquence marginale b_j), seuls les poids correspondant aux valeurs de a_{ij} non nulles doivent être modifiés; par conséquent, chaque cycle n'exige que S_1 multiplications (le nombre d'éléments non nuls de A, qui est moindre que la population de l'îlot) et peut-être $3S_1$ additions, comparativement à $5S_1 + S_2$ opérations par itération pour ce qui est de la méthode de Newton. Par ailleurs, comme les lignes de A tendent à être fortement dépendantes, la convergence peut être lente (en règle générale, 20 cycles dans nos simulations); l'orthogonalisation de A a pour effet de détruire la structure à faible densité des coefficients. Ainsi, à moins que S_2 soit beaucoup plus grand que S_1 (ou que l'on trouve une autre méthode pour activer l'algorithme), la méthode itérative du quotient n'est pas plus rapide que la méthode de Newton.

REMERCIEMENTS

Cette étude a été rendue possible grâce aux conventions 86-8 et 87-7 (Joint Statistical Agreements) entre le U.S. Bureau of the Census et l'Université Harvard sur la statistique. L'auteur a mis à profit les commentaires reçus de Donald B. Rubin et des autres participants au colloque sur le sous-dénombrement dans le recensement, organisé par le département de statistique de l'Université Harvard, de même que les commentaires reçus de Nathaniel Schencker, de Nash Monsour et d'autres membres du Undercount Research Staff au U.S. Bureau of the Census. Les données ayant servi aux simulations ont été fournies par le Consortium interuniversitaire pour la recherche politique et sociale et avaient été recueillies à l'origine par le Bureau of the Census, Département du commerce des États-Unis.

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.

BISHOP, Y.M.M., FIENBERG, S.E., et HOLLAND, P.W. (1974). *Discrete Multivariate Analysis.* Cambridge: M.I.T. Press.

BUREAU OF THE CENSUS (1985). Census of Population and Housing, 1980: Public Use Microdata Samples.

CILKE, J.M., et WYSCARVER, R.A. (1988). The Individual Income Tax Simulation Model, dans *Compendium of Tax Research* 1987, Office of Tax Analysis. Washington: Government Printing Office.

COX, L. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 520-524.

COX, L., FAGAN, J., GREENBURG, B., et HEMMIG, R. (1986). Research at the Census Bureau into disclosure avoidance techniques for tabular data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 388-393.

CRESSIE, N., et READ, T.R.C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society*, Série B, 46, 440-464.

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

DIFFENDAL, G. (1988). Test des opérations de redressement de 1986 dans le Central Los Angeles County. *Techniques d'enquête*, 14, 75-92.

FAY, R.E. (1986). Implications of the 1980 PEP for future census coverage evaluation. U.S. Bureau of the Census, non publié.

FAGAN, J.T., et GREENBERG, B. (1988). Algorithms for making tables additive: Raking, Maximum Likelihood, and Minimum Chi-square. *Proceedings of the Section on Survey Research Methods, American Statistical Association* (à paraître).

GASS, S.I. (1964). *Linear Programming: Methods and Applications*. New York: McGraw-Hill.

IRELAND, C.T., et KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika*, 55, 179-188.

LITTLE, R.A., et RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

NATIONAL ACADEMY OF SCIENCES (1985). *The Bicentennial Census: New Directions for Methodology in* 1990. Washington: National Academy Press.

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

PURCELL, N.J. (1979). Efficient estimation for small domains: a categorical data analysis approach. Thèse de doctorat, University of Michigan.

PURCELL, N.J., et KISH, L. (1979). Estimation for small domains. *Biometrics*, 35:365-384.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

SCHEUREN, F.J. (1981). Methods of estimation for the 1973 exact match study. Dans *Studies from Interagency Data Linkages*, Washington: Social Security Administration.

ZASLAVSKY, A.M. (1989). Representing Census undercount at the household level. Thèse de doctorat, Département de mathématiques, Massachusetts Institute of Technology, Cambridge, Massachusetts.