Recueil du Symposium 2002 de Statistique Canada
Modélisation de données d'enquête pour la recherche économique et sociale

PONDÉRATION POUR LE RECENSEMENT DU CANADA DE 2001

Michael Bankier[1]

RÉSUMÉ

Le présent article décrit deux modifications apportées à la méthode de détermination des poids de sondage pour le Recensement du Canada de 2001. En premier lieu, il a été décidé d'utiliser un estimateur par régression pseudo-optimal plutôt qu'un estimateur GREG par projection. En deuxième lieu, le traitement des données a été effectué sur ordinateur personnel plutôt que sur l'ordinateur central, ce qui a permis d'exécuter plusieurs passages de production, en utilisant divers paramètres. Puis, on a sélectionné et retenu le * meilleur + passage de production (en ce qui concerne l'objectif a) de la section 2.2) pour chaque petit domaine. Ces deux changements ont permis de garder un plus grand nombre de variables auxiliaires dans les estimateurs par régression tout en permettant que les poids de recensement soient tous au moins égaux à l'unité.

Mots clés : Élimination de variables auxiliaires; modèles de superpopulation; estimation par calage.

1. PLAN DE L'ARTICLE

À la Section 2, nous décrivons le plan d'échantillonnage du recensement ainsi que les objectifs de la pondération des données de recensement. À la section 3, nous discutons de l'estimateur par régression généralisée (GREG), de l'estimateur par régression optimal et de la situation où ces deux estimateurs sont les mêmes. On utilise les estimateurs par régression dans le recensement parce qu'ils permettent d'uniformiser les estimations/populations pour plusieurs variables auxiliaires simultanément tout en réduisant la variance. En outre, à la section 3, nous montrons pourquoi il pourrait être avantageux d'utiliser un estimateur par régression optimal plutôt que l'estimateur GREG par projection. La section 4 fournit des précisions sur l'estimateur par régression en deux étapes pseudo-optimal utilisé pour le Recensement de 2001, y compris la méthode d'élimination des variables auxiliaires pour assurer que les poids soient au moins égaux à un. La section 5 décrit le traitement des poids du Recensement de 2001, y compris l'analyse effectuée pour déterminer quelles combinaisons de paramètres devraient être utilisées pour les dix passages de production. En outre, elle donne une description de la méthode utilisée, pour chaque petit domaine géographique, pour sélectionner le meilleur des dix passages de production. Enfin, la section 6 donne les conclusions.

2. CONTEXTE

2.1 Plan d'échantillonnage du recensement

Dans le cadre du recensement du Canada, on recueille les renseignements de base sur les personnes et les logements auprès de tous les membres de la population. Les renseignements ainsi recueillis sont appelés données intégrales ou données 2A d'après le questionnaire abrégé 2A du recensement. On pose aussi des questions supplémentaires à un échantillon au 1/5 de ménages privés (2,2 millions de ménages échantillonnés en 2001) stratifié en 35 895 secteurs de dénombrement (SD). Les données recueillies ainsi sont appelées données-échantillon ou données 2B d'après le questionnaire détaillé 2B du recensement. On applique uniformément une fraction d'échantillonnage de 1/5 pour chaque province (à part quelques SD spéciaux où l'échantillonnage est intégral) de sorte que les estimations infraprovinciales fondées sur des échantillons de taille égale aient la même fiabilité dans toutes les régions du pays.

[1]Statistique Canada, Immeuble R.-H.-Coats 15e étage, Ottawa (Ontario) K1A 0T6, Canada, bankier@statcan.ca

2.2 Objectifs de la pondération et données du Recensement de 2001

Pour chaque ménage échantillonné, on calcule un poids unique qui est utilisé pour produire toutes les estimations publiées des caractéristiques des ménages et des personnes. On utilise un poids unique par souci de simplicité et d'uniformité. Les dénombrements intégraux (données 2A) publiés devraient concorder étroitement avec les estimations publiées des dénombrements intégraux (données 2A) basées sur les données-échantillon (20 %), puisque tout écart important entre les données intégrales et les estimations basées sur les données-échantillon (20 %) préoccupe les utilisateurs des données du recensement. Par conséquent, la méthode d'estimation appliquée pour le recensement vise à réduire ou à éliminer ces écarts entre les estimations et les chiffres de population pour les petits domaines géographiques. Parallèlement, cet effort réduit les erreurs-types des estimateurs des chiffres de recensement. La méthode d'estimation a été conçue pour donner de bons résultats pour les milliers d'estimations publiées qui sont produites avec un minimum d'interventions manuelles durant le traitement des données de recensement. Nous donnons aux caractéristiques pour lesquelles il est nécessaire que l'estimation-échantillon et le chiffre de population concordent le nom de variables auxiliaires ou, autrement, de contraintes sur les poids. Suivent les objectifs de la pondération des données du Recensement de 2001.

Pour les tranches d'âge de cinq ans, l'état matrimonial, l'union libre, le sexe et la taille du ménage (32 variables auxiliaires), les objectifs sont les suivants :
a) Obtenir une concordance exacte entre les estimations-échantillon et les chiffres de population au niveau du secteur de pondération (SP) pour un aussi grand nombre que possible des 32 variables auxiliaires. Il existe 6 142 SP susceptibles d'être échantillonnés qui correspondent souvent à de petites municipalités ou à des secteurs de recensement. Un SP comprend, en moyenne, huit aires de diffusion (AD) complètes. Le Canada est subdivisé en 47 933 AD échantillonnées, contenant, en moyenne, 239 logements privés occupés chacune.
b) Obtenir une concordance approximative entre les estimations-échantillon et les chiffres de population pour les grandes AD pour les 32 variables auxiliaires.

En outre, il faut que les critères suivants soient satisfaits :
c) La concordance entres les estimations-échantillon et les chiffres de population devrait être exacte pour le nombre total de ménages et le nombre total de personnes pour autant d'AD que possible.
d) Les poids finals de recensement devraient se situer dans la fourchette de 1 à 25 inclusivement. En 1996, les poids finals de recensement pouvaient varier de 0,01 à 25 inclusivement.
e) La méthode de calcul des poids devrait être hautement automatisée puisqu'on doit traiter les données pour les 6 142 SP en très peu de temps. Cette méthode doit aussi procéder à l'ajustement automatique pour la variation du profil de réponse selon le SP à l'échelle du pays.
g) Pour 2001, on souhaitait trouver un moyen de mieux satisfaire l'objectif (a). Malheureusement, on n'a pu apporter que fort peu de changements au logiciel utilisé pour les recensements de 1991 et 1996, à cause de contraintes budgétaires et de dotation en personnel.

3. ESTIMATEURS PAR RÉGRESSION SOUS ÉCHANTILLONNAGE STRATIFIÉ

Fuller (2002) donne une excellente revue de l'estimation par régression sur échantillon. À la présente section, nous comparons l'estimateur par régression généralisée (GREG) découlant de l'estimateur par projection à l'estimateur par régression optimal.

Par souci de simplicité, nous examinons le cas des estimateurs pour un seul secteur de pondération (SP) comprenant H secteurs de dénombrement (SD). Nous supposons qu'on a sélectionné un échantillon aléatoire simple sans remise (e.a.s.s.r.) de taille n_h à partir de la population de N_h ménages dans le h$^{\text{ième}}$ SD, h = 1 à H et que $n = \sum_h n_h$ et $N = \sum_h N_h$. Nous supposons aussi à la présente section (bien que ce ne soit pas le cas dans le recensement) que la fraction d'échantillonnage n_h/N_h peut varier considérablement selon le SD. La raison pour laquelle nous formulons cette hypothèse dans le contexte du recensement sera plus claire à la section 4.

L'estimateur le plus simple possible est l'estimateur d'Horvitz-Thompson $\hat{Y}^{(0)} = \sum_i W_i^{(0)} y_i$ où $W_i^{(0)} = N_h/n_h$ si le i$^{\text{ième}}$ ménage échantillonné est compris dans le h$^{\text{ième}}$ SD. Cependant, en général, il n'est nullement garanti que l'on réalise

l'objectif a) susmentionné pour n'importe laquelle des 32 variables auxiliaires au moyen de l'estimateur d'Horvitz-Thompson. Par conséquent, nous allons examiner divers types d'estimateurs par régression.

3.1 GREG

Les estimateurs par calage prennent la forme $\hat{Y} = \hat{\underset{\sim}{Y}}^{(0)}\underset{\sim}{g} = \sum_i g_i w_i^{(0)} y_i$ où le vecteur $\underset{\sim}{g} = [g_i]$ de dimensions n x 1 des facteurs de correction de la pondération (appelés aussi poids g) est choisi de sorte qu'une fonction de perte L soit minimisée sous les contraintes $\hat{\underset{\sim}{X}}^{(0)}\underset{\sim}{g} = \underset{\sim}{X}\underset{\sim}{1}_N$ où $\hat{\underset{\sim}{Y}}^{(0)} = [w_i^{(0)} y_i]$ est une matrice de dimensions 1 x n, $X = [x_{pi}]$ est une matrice de dimensions P x N, x_{pi} représente la valeur de la p[ième] variable auxiliaire pour le i[ième] ménage dans le SP, $\hat{\underset{\sim}{X}}^{(0)} = \underset{\sim}{x}\, diag(\underset{\sim}{w}^{(0)}) = [w_i^{(0)} x_{pi}]$, $\underset{\sim}{x}$ est une matrice de dimensions P x n qui contient les n colonnes provenant de $\underset{\sim}{X}$ qui correspondent aux ménages échantillonnés, $\underset{\sim}{w}^{(0)} = [w_i^{(0)}]$ est un vecteur de dimensions n x 1 des poids initiaux et $diag(\underset{\sim}{w}^{(0)})$ est une matrice de dimensions n x n dont les valeurs sont $\underset{\sim}{w}^{(0)}$ sur la diagonale et nulle ailleurs.

Sous la forme la plus générale de l'estimateur GREG, la fonction de perte prend la forme $L = (\underset{\sim}{g} - \underset{\sim}{1}_n)'\hat{\underset{\sim}{V}}(\underset{\sim}{g} - \underset{\sim}{1}_n)$ et le vecteur $\underset{\sim}{g}$ qui minimise L est

$$\underset{\sim}{g} = \underset{\sim}{1}_n + \hat{\underset{\sim}{V}}^{-1}\hat{\underset{\sim}{X}}^{(0)'}(\hat{\underset{\sim}{X}}^{(0)}\hat{\underset{\sim}{V}}^{-1}\hat{\underset{\sim}{X}}^{(0)'})^{-1}(\underset{\sim}{X}\underset{\sim}{1}_N - \hat{\underset{\sim}{X}}^{(0)}\underset{\sim}{1}_n) \qquad (1)$$

où on suppose que $\hat{\underset{\sim}{V}}$ est une matrice symétrique de dimensions n x n qui doit être définie positive (ce qui implique qu'elle est non singulière) pour assurer que la fonction de perte L ne soit pas négative.

Il est possible d'écrire $\hat{Y}$ sous la forme type d'un estimateur par régression

$$\begin{aligned}\hat{Y} &= \hat{Y}^{(0)} + \hat{\underset{\sim}{B}}'(\underset{\sim}{X}\underset{\sim}{1}_N - \hat{\underset{\sim}{X}}^{(0)}\underset{\sim}{1}_n) \\ &= \hat{\underset{\sim}{B}}'\underset{\sim}{X}\underset{\sim}{1}_N + \hat{\underset{\sim}{e}}^{(0)}\underset{\sim}{1}_n \qquad (2)\end{aligned}$$

où $\hat{\underset{\sim}{B}} = [\hat{B}_p] = (\hat{\underset{\sim}{X}}^{(0)}\hat{\underset{\sim}{V}}^{-1}\hat{\underset{\sim}{X}}^{(0)'})^{-1}\hat{\underset{\sim}{X}}^{(0)}\hat{\underset{\sim}{V}}^{-1}\hat{\underset{\sim}{Y}}^{(0)'}$ est un vecteur de dimensions P x 1, $\hat{\underset{\sim}{e}}^{(0)} = [w_i^{(0)} e_i] = \hat{\underset{\sim}{Y}}^{(0)} - \hat{\underset{\sim}{B}}'\hat{\underset{\sim}{X}}^{(0)}$ est un vecteur des résidus de dimensions 1 x n et $e_i = y_i - \hat{\underset{\sim}{B}}'\underset{\sim}{x}_i$. On peut montrer que $\hat{\underset{\sim}{B}}$ minimise la fonction de perte $L = \hat{\underset{\sim}{e}}^{(0)}\hat{\underset{\sim}{V}}^{-1}\hat{\underset{\sim}{e}}^{(0)'}$.

Särndal, Swensson et Wretman (1992) utilisent des modèles pour justifier le choix de $\hat{\underset{\sim}{V}}$. Ils supposent que les quantités de population $\underset{\sim}{Y}$ ont été générées par un modèle comme suit. Supposons, sous un modèle ξ, que

$$\underset{\sim}{Y} = \underset{\sim}{\beta}'\underset{\sim}{X} + \underset{\sim}{e} \qquad (3)$$

où $E_\xi(\underset{\sim}{e}) = \underset{\sim}{0}$ et $Cov_\xi(\underset{\sim}{e}, \underset{\sim}{e}') = \underset{\sim}{V}$ tandis que $\underset{\sim}{Y}$ et $\underset{\sim}{X}$ sont, respectivement, des matrices de dimensions 1 x N et P x N au niveau de la population. E_ξ, Cov_ξ et V_ξ représentent la valeur prévue, la covariance et la variance par rapport au modèle ξ tandis que $\underset{\sim}{\beta}$ (un vecteur de dimensions P x 1) et $\underset{\sim}{V}$ (une matrice symétrique de dimensions N x N) sont les paramètres du modèle. Nous souhaitons trouver un estimateur $\underset{\sim}{B}$ tel que, pour un vecteur $\underset{\sim}{t}$ arbitraire de dimensions P x 1, $\underset{\sim}{t}'\underset{\sim}{B}$ soit le meilleur estimateur sans biais (m.e.s.s.b) de $\underset{\sim}{t}'\underset{\sim}{\beta}$ où

- par * meilleur + nous entendons que $V_\xi(\underset{\sim}{t}'\underset{\sim}{B})$ est minimisé;
- par linéaire nous entendons que $\underset{\sim}{t}'\underset{\sim}{B} = \underset{\sim}{\gamma}'\underset{\sim}{Y}$ pour certains $\underset{\sim}{\gamma}$ (c.-à-d. une fonction linéaire pour y_i, i = 1 à N);
- par sans biais nous entendons que $E_\xi(\underset{\sim}{t}'\underset{\sim}{B}) = \underset{\sim}{t}'\underset{\sim}{\beta}$

Si le modèle supposé ξ est correct, on réalise ces objectifs en choisissant

$$\underset{\sim}{B} = (\underset{\sim}{X}'\underset{\sim}{V}^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{V}^{-1}\underset{\sim}{Y} \qquad (4)$$

Selon Särndal et coll. (1992), le rôle du modèle ξ est de décrire le nuage de points de la population finie. Le modèle sert de véhicule à la recherche d'un $\underset{\sim}{\hat{B}}$ approprié à intégrer dans la formule de l'estimateur par régression. Étant donné que $\underset{\sim}{t}'\underset{\sim}{B}$ est le m.e.s.b. de $\underset{\sim}{t}'\underset{\sim}{\beta}$, il semble raisonnable, si le modèle ξ est correct, d'estimer $\underset{\sim}{B}$ et, donc, $\underset{\sim}{\beta}$ au moyen de l'estimateur approximativement sans biais $\underset{\sim}{\hat{B}}$ défini plus haut.

Typiquement, Särndal et coll. (1992) supposent que $\underset{\sim}{V}$ est une matrice diagonale dont les éléments $v_i = \sigma_i^2$ se situent sur la diagonale et dont les autres valeurs sont nulles. Ils donnent des exemples simples d'une variable auxiliaire unique où, par exemple, il est supposé que $\sigma_i^2 = \sigma^2 x_i$. Cependant, en pratique, on ne connaît pas les valeurs de σ_i^2, i = 1 à N, et il est difficile de les estimer. Les erreurs qui entachent leur estimation appauvrissent la qualité de l'estimateur $\underset{\sim}{B}$ de $\underset{\sim}{\beta}$. En outre, les valeurs de σ_i^2 varient habituellement pour chaque caractéristique y considérée. Par conséquent, il faut produire des poids calés différents pour les diverses caractéristiques y, ce qui n'est pas pratique si l'enquête porte sur de nombreuses caractéristiques. Enfin, Särndal et coll. (1992) supposent généralement en pratique, si le nombre de variables auxiliaires est égal ou supérieur à deux, que $\sigma_i^2 = \sigma^2$ si bien que σ_i^2 disparaît de la formule de la variance pour $\underset{\sim}{\hat{B}}$. Lorsqu'on utilise l'estimateur par régression optimal décrit plus bas, on ne recourt pas à un modèle de superpopulation. Donc, au lieu d'utiliser σ_i^2, nous utiliserons ici la notation plus générale v_i lorsqu'il est supposé que $\underset{\sim}{\hat{V}}$ est une matrice diagonale.

3.2 Estimateur par régression optimal

Cochran (1942) et Rao (1994) recommandent d'utiliser l'estimateur par régression optimal. Il est dénommé ainsi parce que la variance de

$$\hat{Y}_{opt} = \hat{Y}^{(0)} + \underset{\sim}{B}'_{opt}(\underset{\sim}{X}\underset{\sim}{1}_N - \underset{\sim}{\hat{X}}^{(0)}\underset{\sim}{1}_n) \qquad (5)$$

est minimisée si $\underset{\sim}{B}_{opt} = \underset{\sim}{\Sigma}_{xx}^{-1}\underset{\sim}{\sigma}_{yx}$ où $\underset{\sim}{\Sigma}_{xx}$ et $\underset{\sim}{\sigma}_{yx}$ représentent, respectivement, la matrice des covariances des $\hat{X}^{(0)}\underset{\sim}{1}_n$ de dimensions P x P et le vecteur des covariances $Cov(\hat{Y}^{(0)}, \hat{X}_p^{(0)})$ de dimensions P x 1 où $\underset{\sim}{\hat{X}}^{(0)}\underset{\sim}{1}_n = [\hat{X}_p^{(0)}]$. L'estimateur standard $\underset{\sim}{B}_{opt}$ (qui est approximativement sans biais) est $\underset{\sim}{\hat{B}}_{opt} = \underset{\sim}{\hat{\Sigma}}_{xx}^{-1}\underset{\sim}{\hat{\sigma}}_{yx}$ où $\underset{\sim}{\hat{\Sigma}}_{xx}$ et $\underset{\sim}{\hat{\sigma}}_{yx}$ sont des estimateurs sans biais de $\underset{\sim}{\Sigma}_{xx}$ et $\underset{\sim}{\sigma}_{yx}$.

3.3 Situation où les estimateurs GREG et optimal sont les mêmes

Supposons que l'ensemble de contraintes $\underset{\sim}{\hat{X}}^{(0)}g = \underset{\sim}{X}\underset{\sim}{1}_N$ inclut les contraintes $\hat{X}_h = X_h$, h = 1 à H, où $X_h = N_h$ est le nombre de ménages dans le h^{ième} SD et $\hat{X}_h$ est l'estimation correspondante après calage. Il est alors possible d'écrire (à condition de choisir de façon appropriée $\underset{\sim}{\hat{V}}$ pour $\hat{Y}$) :

$$\hat{Y} = \hat{Y}^{(0)} + \hat{\underset{\sim}{B}}' (\underset{\sim}{X} \underset{\sim}{1}_N - \hat{\underset{\sim}{X}}^{(0)} \underset{\sim}{1}_n)$$
$$= \hat{Y}_{opt} = \hat{Y}^{(0)} + \hat{\underset{\sim}{B}}'_{opt} (\underset{\sim}{X}_{\&} \underset{\sim}{1}_N - \hat{\underset{\sim}{X}}^{(0)}_{\&} \underset{\sim}{1}_n) \qquad (6)$$

où les H premières lignes de $\hat{\underset{\sim}{X}}^{(0)}$ et $\underset{\sim}{X}$ sont en relation avec les contraintes $\hat{X}_h = X_h$ de niveau SD et peuvent être représentées par les matrices $\hat{\underset{\sim}{X}}^{(0)}_H$ et $\underset{\sim}{X}_H$, respectivement, tandis que les P - H dernières lignes de $\hat{\underset{\sim}{X}}^{(0)}$ et $\underset{\sim}{X}$ sont en relation avec les autres contraintes et peuvent être représentées par les matrices $\hat{\underset{\sim}{X}}^{(0)}_{\&}$ et $\underset{\sim}{X}_{\&}$, respectivement. $\hat{\underset{\sim}{B}}_{opt} = \hat{\underset{\sim}{\Sigma}}^{-1}_{xx\&} \hat{\underset{\sim}{\sigma}}_{yx\&}$ est un vecteur de dimensions (P - H) x 1 où les covariances utilisées pour calculer $\hat{\underset{\sim}{B}}_{opt}$ sont limitées aux variables auxiliaires représentées par les lignes de $\hat{\underset{\sim}{X}}^{(0)}_{\&}$. $\hat{\underset{\sim}{B}} = [\hat{B}_p] = (\hat{\underset{\sim}{X}}^{(0)} \hat{\underset{\sim}{V}}^{-1} \hat{\underset{\sim}{X}}^{(0)})^{-1} \hat{\underset{\sim}{X}}^{(0)} \hat{\underset{\sim}{V}}^{-1} \hat{\underset{\sim}{Y}}^{(0)}$ est un vecteur de dimensions P x 1 où, pour réaliser $\hat{Y} = \hat{Y}_{opt}$, il est nécessaire que $\hat{\underset{\sim}{V}}$ soit une matrice diagonale de dimensions n x n contenant les éléments $\hat{V}_{hi} = W_h^{(0)} v_{hi} > 0$ sur la diagonale et des valeurs nulles ailleurs et pour laquelle $v_{hi} = (n_h - 1)/(n_h(W_h^{(0)} - 1))$ où $W_h^{(0)} = N_h/n_h$. $\hat{V}_{hi}$ correspond au i[ième] ménage échantillonné à partir du h[ième] SD.

Toutefois, pour les recensements de 1991 et de 1996, $\hat{\underset{\sim}{V}} = diag(\hat{\underset{\sim}{X}}^{(0)} \underset{\sim}{1}_P)$, où $\underset{\sim}{1}_P$ est un vecteur de 1 de dimensions P x 1. Par conséquent, la valeur de v_{hi} est habituellement plus grande pour les ménages comptant un plus grand nombre de personnes, ce qui signifie que le facteur de correction de la pondération g_{hi} a tendance à être plus faible, toutes choses étant égales par ailleurs. Ce choix de $\hat{\underset{\sim}{V}}$ est conforme à la suggestion de Särndal et coll. (1992) voulant que $\hat{\underset{\sim}{V}} = diag(\hat{\underset{\sim}{V}}^{\gamma})$ où $\hat{\underset{\sim}{V}}^{\gamma} = \hat{\underset{\sim}{X}}^{(0)} \underset{\sim}{\gamma} = [W_h^{(0)} v_{hi}]$ soit un vecteur de dimensions n x 1 et $\underset{\sim}{\gamma}$, un vecteur de dimensions P x 1 qui n'aboutit pas à ce que tout élément de $\hat{\underset{\sim}{X}}^{(0)} \underset{\sim}{\gamma}$ devienne nul. Ils proposent des matrices $\tilde{\underset{\sim}{V}}$ de cette forme, car, alors, $\hat{Y}$ prend la forme de projection simple $\tilde{Y} = \tilde{\underset{\sim}{\beta}}' \underset{\sim}{X} \underset{\sim}{1}_N$. Il convient de souligner que, sauf dans des cas spéciaux, l'estimateur GREG par projection n'est pas l'estimateur par régression optimal.

Särndal et coll. (1992) supposent souvent que $\sigma_i^2 = \sigma^2$, ce qui revient à supposer que $v_{hi} = 1$ et $\hat{V}_{hi} = W_h^{(0)} = N_h/n_h$. Cette valeur de v_{hi} peut être obtenue pour l'estimateur GREG par projection en supposant que les H premières lignes de $\hat{X}^{(0)}$ représentent les contraintes $\hat{N}_h = N_h$ au niveau du SD, tandis que $\underset{\sim}{\gamma}$ a la valeur $\gamma_p = 1$ pour p = 1 à H et $\gamma_p = 0$ autrement.

Särndal (1996) montre que la variance de l'approximation par série de Taylor de l'estimateur GREG est minimisée sous e.a.s.s.r. stratifié (en supposant que $\hat{X}_h = X_h$, h = 1 à H fait partie des contraintes utilisées) si $v_{hi} = n_h(N_h - 1)/(N_h(N_h - n_h))$. Cette expression de la variance est approximativement égale à l'expression $v_{hi} = (n_h - 1)/(n_h(W_h^{(0)} - 1))$ établie plus haut, où on rend l'estimateur GREG identique à l'estimateur par régression optimal. Särndal (1996) déclare à la remarque 3.2 * qu'il pourrait y avoir certains avantages (quoiqu'habituellement modestes) à prendre + $v_{hi} = n_h(N_h - 1)/(N_h(N_h - n_h))$ plutôt que $v_{hi} = 1$. Dans la suite de la section, nous montrons que, si la fraction d'échantillonnage varie considérablement selon la strate, faire varier v_{hi} selon la strate pourrait offrir des avantages importants.

Supposons que la fraction d'échantillonnage varie considérablement selon le SD dans un SP. Le tableau 1 qui suit donne, pour l'estimateur optimal, les valeurs de $\hat{V}_{hi} = (n_h - 1) W_h^{(0)}/(n_h(W_h^{(0)} - 1))$ et, pour l'estimateur GREG par projection, les valeurs de $\hat{V}_{hi} = W_h^{(0)}$ pour $N_h = 400$ ménages privés et pour diverses valeurs de n_h. La fonction de perte minimisée à la fois pour l'estimateur optimal et pour l'estimateur GREG par projection (avec des valeurs différentes de $\hat{V}_{hi}$) est $L = \sum_h \sum_i \hat{V}_{hi}(g_{hi} - 1)^2$ où g_{hi} est le facteur de correction de la pondération pour la i[ième] unité échantillonnée à partir du h[ième] SD.

Pour l'estimateur optimal, quand la fraction d'échantillonnage passe de 5 % à 94 %, $\hat{V}_{hi}$ passe de 1 à 16 dans le tableau 1. Autrement dit, les SD pour lesquels les fractions d'échantillonnage sont grandes ont tendance à avoir un g_{hi} proche de 1, toutes choses étant égales par ailleurs, à cause de la minimisation de la fonction de perte. Ce résultat est sensé, car les estimations calculées pour les SD pour lesquels la fraction d'échantillonnage est grande devraient être plus fiables et, donc, devraient nécessiter une correction moins importante que celle calculée pour les SD dont la fraction d'échantillonnage est plus petite.

Pour l'estimateur GREG par projection, à mesure que la fraction d'échantillonnage augmente pour passer de 5 % à 94 %, $\hat{V}_{hi}$ diminue, passant de 20 à 1,1 dans le tableau 1. Autrement dit, les SD pour lesquels la fraction d'échantillonnage est grande ont tendance à avoir un g_{hi} dont la valeur s'écarte beaucoup de 1, toutes choses étant égales par ailleurs, à cause de la minimisation de la fonction de perte. Ce résultat est contraire à l'intuition. Il est également indésirable, car si la valeur de $w_h^{(0)}$ est proche de 1 et que $g_{hi} < 1$, il est fort possible que le poids corrigé $g_{hi} w_h^{(0)}$ soit inférieur à 1, voire négatif. Aux termes de la méthode courante d'estimation du recensement (qui sera décrite à la section 4), les contraintes sont relâchées lorsque cela se produit. Donc, utiliser v_{hi} ' $(n_h$&1)/$(n_h(w_h^{(0)}$&1)), comme l'exige l'estimateur par régression optimal, semble préférable, puisque, outre la minimisation de la variance, cette expression permet de maintenir un plus grand nombre de contraintes sous l'exigence que les poids corrigés ne soient pas inférieurs à 1.

Tableau 1 : $\hat{V}_{hi}$ pour diverses fractions d'échantillonnage en supposant que N_h ' 400

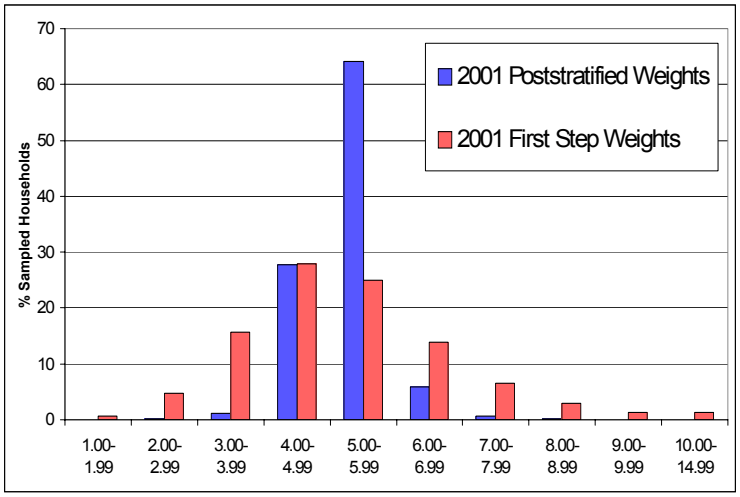
n_h	$100n_h/N_h$	$\hat{V}_{hi}$ Optimal	$\hat{V}_{hi}$ GREG
20	5	1	20
40	10	1,1	10
80	20	1,2	5
120	30	1,4	3,3
200	50	2	2
300	75	4	1,3
375	94	16	1,1

4. ESTIMATEUR PAR RÉGRESSION UTILISÉ DANS LE RECENSEMENT DE 2001

Le lecteur trouvera une description plus détaillée de la méthodologie dans Bankier, Rathwell et Majkowski (1992).

4.1 Estimateur par régression en deux étapes

À la présente section, par souci de simplicité, nous ignorons le fait que les contraintes sont relâchées pour diverses raisons (voir la section 4.2). Nous calculons séparément les poids pour chaque SP. Nous appliquons aux poids initiaux au niveau du SD du Recensement de 2001 $w_i^{(0)}$ ' N_h/n_h deux ou bien trois facteurs de correction de la pondération. Pour commencer, il arrive que les ménages soient stratifiés a posteriori au niveau du SP sur la taille du ménage, car les petits et les très grands ménages ont tendance à être sous-représentés dans l'échantillon. Puis, une deuxième correction des poids est faite pour essayer de réaliser une concordance approximative entre les estimations-échantillon et les



chiffres de population au niveau de l'AD comme il l'est précisé à l'objectif (b) de la section 2.2. Enfin, une troisième correction des poids est faite pour réaliser une concordance exacte entre les estimations-échantillon et les chiffres de population au niveau du SP et de l'AD, comme il l'est précisé aux objectifs (a) et (c) de la section 2.2. Ces trois corrections sont décrites de façon plus approfondie dans les paragraphes qui suivent.

Premièrement, nous procédons parfois à la **stratification a posteriori** des ménages selon la taille du ménage (1,2,3,4,5,6+ personnes) au niveau du SP, puis nous calculons les poids stratifiés a posteriori $w_i^{(} ' g_i^{(} w_i^{(0)}$. Très occasionnellement, nous tronquons $w_i^{(}$ pour nous assurer que la valeur soit comprise dans la fourchette de 1 à 20 inclusivement. Nous choisissons un seuil supérieur de 20 au lieu de 25 pour avoir une certaine * liberté + pour d'autres corrections.

Ensuite, nous calculons un facteur de correction de la pondération par une **première étape** de régression au niveau de l'AD. Les 32 variables auxiliaires (âge, sexe, état matrimonial, taille du ménage) qu'il faut appliquer au niveau du SP à la deuxième étape sont classées par ordre décroissant d'après le nombre de ménages auxquels elles s'appliquent dans la population au niveau de l'AD. Les première, troisième, et ainsi de suite, contraintes qui figurent sur cette liste ordonnée vont dans un groupe, tandis que les 16 autres vont dans un autre groupe. Nous calculons la moyenne des facteurs de correction de la pondération résultants pour chaque groupe de contraintes (annotés $g_i^{(A1)}$ et $g_i^{(A2)}$) pour créer $g_i^{(A)} ' (g_i^{(A1)} % g_i^{(A2)}) / 2$ qui, alors, génère les poids de première étape $w_i^{(A)} ' g_i^{(A)} w_i^{(}$. Habituellement, les écarts entre les estimations-échantillon et les chiffres de population au niveau de l'AD observés pour les 32 contraintes sont réduits, mais non éliminés, au moyen des poids de première étape.

Enfin, nous calculons le facteur de correction de la pondération par une **deuxième étape** de régression au niveau du SP. Nous appliquons les 32 contraintes au niveau du SP, ainsi que deux contraintes (nombre de ménages et nombre de personnes) pour chaque AD comprise dans le SP pour déterminer le facteur final de correction de la pondération g_i. Celui-ci génère alors les poids de deuxième étape $w_i ' g_i w_i^{(A)}$.

À la première étape, $\hat{V}_i ' w_i^{(} / (w_i^{(} \& 1)$, tandis qu'à la deuxième, $\hat{V}_i ' w_i^{(A)} / (w_i^{(A)} \& 1)$. Ces choix de $\hat{V}_i$ à la première et à la deuxième étapes font ressembler la fonction de perte minimisée à celle utilisée pour l'estimateur par régression optimal. Ils favorisent également la génération de facteurs de correction de la pondération de première et de deuxième étapes dont la valeur est proche de 1 pour les poids stratifiés a posteriori et ceux de première étape les plus petits (voir le graphique ci-contre pour la distribution de ces poids pour le Recensement de 2001) et, donc, découragent la création de poids corrigés de valeur inférieure à 1. Étant donné le choix de $\hat{V}_i$, nous appellerons l'estimateur utilisé pour le recensement **estimateur pseudo-optimal en deux étapes**. Nous pouvons estimer la variance de cet estimateur par régression en deux étapes en utilisant le développement en série de Taylor pour linéariser numériquement les données deux à trois fois de façon semblable à celle proposée pour l'estimation par la méthode itérative du quotient décrite dans Bankier (1986).

4.2 Élimination des contraintes

Pour une discussion des raisons qui justifient l'élimination des contraintes, consulter Silva et Skinner (1997) et Fuller (2002). L'étude de Silva et Skinner a été motivée en partie par la méthodologie décrite à la présente section. Les contraintes sont éliminées durant le calcul des poids si elles sont petites, linéairement dépendantes (LD) ou presque linéairement dépendante (PLD), ou qu'elles produisent des poids dont la valeur est aberrante (en dehors de la fourchette de 1 à 25). Au départ, nous recherchons les contraintes qui sont petites, LD ou PLD au niveau du SP comme suit. Nous définissons la taille d'une contrainte comme étant le nombre de ménages de la population auxquels elle s'applique. Au départ, toute contrainte dont la taille est égale ou inférieure à SMALL (c.-à-d. un paramètre dont la valeur est égale à 20, 30 ou 40 pour 2001) est éliminée parce que, pour les petites contraintes, les estimations ont tendance à être fort instables. Puis, comme la matrice $\hat{\underset{\sim}{X}}^{(0)}\hat{\underset{\sim}{V}}^{\&1}\hat{\underset{\sim}{X}}^{(0)}$ doit être inversée pour calculer $\underset{\sim}{g}$ (voir la section 3.1), nous repérons des ensembles de contraintes linéairement dépendantes, qui rendent cette matrice singulière, et nous éliminons la contrainte la plus petite dans chaque ensemble. Ensuite, nous réduisons le nombre de conditions de $\hat{\underset{\sim}{X}}^{(0)}\hat{\underset{\sim}{V}}^{\&1}\hat{\underset{\sim}{X}}^{(0)}$ (qui est généralement assez grand dans le cas du recensement) en éliminant ce que nous appelons les contraintes PLD. Le nombre de conditions est égal au ratio de la valeur propre la plus grande à la valeur propre la plus faible de $\hat{\underset{\sim}{X}}^{(0)}\hat{\underset{\sim}{V}}^{\&1}\hat{\underset{\sim}{X}}^{(0)}$. Un nombre élevé de conditions indique une quasi-colinéarité des contraintes. Pour réduire le nombre de conditions, nous utilisons une méthode de sélection ascendante. Nous recalculons la matrice $\hat{\underset{\sim}{X}}^{(0)}\hat{\underset{\sim}{V}}^{\&1}\hat{\underset{\sim}{X}}^{(0)}$ en nous fondant uniquement sur les deux contraintes les plus grandes. Si le nombre de conditions excède la valeur du paramètre COND (qui, par exemple, pourrait être égale à 1 000), nous écartons la deuxième contrainte la plus grande par ordre décroissant, puis nous ajoutons la contrainte la plus grande qui suit, nous recalculons la matrice $\hat{\underset{\sim}{X}}^{(0)}\hat{\underset{\sim}{V}}^{\&1}\hat{\underset{\sim}{X}}^{(0)}$et nous déterminons son nombre de conditions. Si celui-ci augmente d'une valeur supérieure à celle de COND, nous éliminons la contrainte qui vient d'être ajoutée. Ce processus se poursuit jusqu'à ce qu'on ait vérifié toutes les contraintes de cette façon. Si, après avoir éliminé ces contraintes PLD, le nombre de conditions excède la valeur du paramètre MAXC (qui, par exemple, pourrait être égale à 10 000), nous éliminons d'autres contraintes. Nous procédons à cette élimination par ordre décroissant de quantité par laquelle a été augmenté le nombre de conditions lorsqu'elles ont été incluses dans la matrice au départ. Le nombre de conditions de la matrice $\hat{\underset{\sim}{X}}^{(0)}\hat{\underset{\sim}{V}}^{\&1}\hat{\underset{\sim}{X}}^{(0)}$ est recalculé chaque fois qu'une contrainte est éliminée. Lorsque le nombre de conditions devient inférieur à la valeur de MAXC, plus aucune contrainte n'est éliminée. Toutes les contraintes éliminées jusqu'à ce point ne sont pas utilisées dans le calcul des poids.

Avant de calculer les facteurs de correction de la pondération de première étape $g_i^{(A)}$ pour la c^e AD (c = 1 à C), nous éliminons les autres contraintes, au besoin, si elles sont petites pour la c^e AD puis, nous répartissons les contraintes retenues en deux groupes tel que décrit à la section 4.1. Puis, pour chaque groupe de contraintes, nous repérons les contraintes linéairement dépendantes et nous les éliminons (les contraintes qui sont linéairement dépendantes au niveau de l'AD peuvent ne pas l'être au niveau du SP). En nous fondant sur les contraintes restantes, nous calculons les facteurs de correction de la pondération de première étape $g_i^{(A1)}$ et $g_i^{(A2)}$. Si tout poids corrigé de première étape se situe en dehors de la fourchette de 1 à 25 inclusivement, nous éliminons d'autres contraintes. Nous appliquons ici une méthode comparable à celle utilisée pour éliminer les contraintes PLD, excepté que nous éliminons une contrainte si elle produit des poids aberrants. Cependant, parce souci d'efficacité des calculs, nous utilisons la méthode de bisection pour déterminer quelles contraintes doivent être éliminées.

Ensuite, nous calculons les facteurs de correction de la pondération de deuxième étape $\underset{\sim}{g}_i$ en nous fondant sur les contraintes qui n'ont pas été éliminées parce qu'elles étaient petites, linéairement dépendantes ou presque linéairement dépendantes d'après l'analyse initiale de la matrice $\hat{\underset{\sim}{X}}^{(0)}\hat{\underset{\sim}{V}}^{\&1}\hat{\underset{\sim}{X}}^{(0)}$. Si tout poids corrigé de deuxième étape tombe en dehors de la fourchette de 1 à 25 inclusivement, nous éliminons des contraintes supplémentaires par la méthode décrite pour la correction de première étape.

5. TRAITEMENT DES POIDS DU RECENSEMENT DE 2001

Les poids du recensement sont calculés au moyen du langage matriciel interactif SAS. Pour 1996, un seul traitement des poids pour l'ensemble du pays a pris environ deux semaines sur l'ordinateur central. En 2001, six ordinateurs personnels Pentium IV 1.7 GHz ont traité les données pour l'ensemble du pays en 24 heures.

L'utilisation des ordinateurs personnels a permis de procéder à des essais étendus sur deux échantillons de SP (121 et 616 SP, respectivement), afin de déterminer les dix * meilleures combinaisons + de paramètres de système de pondération. Par * meilleurs + on entendait les paramètres qui minimisaient le critère ABSDIFF3, qui était égal à la somme de la valeur absolue des différences entre les estimations-échantillon et les chiffres de population pour les 32 variables auxiliaires, où les différences étaient calculées d'après les chiffres estimés et les chiffres de population, totalisés sur les SP échantillonnés. Cette somme a servi de mesure sur un grand domaine de la concordance entre les estimations-échantillon et les chiffres de population pour ces variables.

Puis, on a traité dix fois les données pour l'ensemble du pays au moyen des meilleures combinaisons de paramètres de pondération. Les valeurs de ABSDIFF3 au niveau du Canada produites par les premiers passages de production ont permis d'orienter le choix des paramètres à utiliser pour les échantillons de SP et, plus tard, pour les passages de production.

Une fois que les dix passages de production ont été achevés, on a retenu le * meilleur + pour chaque SP. Par définition, le * meilleur + passage de production était celui qui minimisait la **somme** de
- ABSDIFF2 qui était égal à la somme des valeurs absolues des différences entre les estimations-échantillon et les chiffres de population au niveau du SP pour les 32 variables auxiliaires dont on a alors fait la somme sur tous les SP (une mesure sur petit domaine de la concordance entre les estimations-échantillon et les chiffres de population) **et**
- ABSDIFF1 qui était égal à la somme des valeurs absolues des différences entre les estimations-échantillon et les chiffres de population pour les deux variables auxiliaires au niveau de l'AD (nombres de ménages et de personnes) dont on a alors fait la somme sur toutes les AD (une autre mesure sur petit domaine de la concordance entre les estimations-échantillon et les chiffres de population)

Ce * tri minutieux + des passages de production a permis d'obtenir, au niveau national, des écarts entre les estimations-échantillon et les chiffres de population plus faibles qu'il ne l'avait été possible lors des recensements antérieurs où on avait utilisé la même combinaison de paramètres pour tous les SP.

Les priorités implicites de cette approche étaient d'obtenir la meilleure concordance possible entre les estimations-échantillon et les chiffres de population pour les 32 variables auxiliaires au niveau du Canada, ainsi qu'une très bonne concordance entre les estimations-échantillon et les chiffres de population pour ces variables au niveau du SP.

Les paramètres utilisés pour les dix passages machine et le nombre de SP pour lesquels ces paramètres ont été utilisés après le * tri minutieux + sont énumérés au tableau 2. Pour nous assurer que certaines contraintes importantes soient retenues systématiquement, nous avons exécuté le passage pour deux SP en utilisant des paramètres * personnalisés +. Ceux-ci sont énumérés à la fin du tableau 2. Les paramètres MAXC et SMALL ont été définis à la section 4.2. POST = 1 indique que l'on a effectué une stratification a posteriori selon la taille du ménage (telle que décrite à la section 4.1) tandis que POST = 0 indique que cette stratification a posteriori n'a pas eu lieu. Le paramètre COND ne figure pas sur le tableau 2, car sa valeur est toujours égale à MAXC/10. Cette valeur du ratio entre les deux paramètres est celle qui a été utilisée en 1996 et nous ne disposions pas de suffisamment de temps pour expérimenter d'autres valeurs.

Pour chacun des dix passages de production, nous avons utilisé l'estimateur pseudo-optimal, parce que, avec un échantillon de 121 SP et pour diverses combinaisons de paramètres, ABSDIFF3 était, en moyenne, 46 % plus faible pour l'estimateur pseudo-optimal que pour l'estimateur GREG par projection lorsque la valeur des poids était restreinte à la fourchette de 1 à 25 dans l'un et l'autre cas. Ceci s'explique par le fait que le nombre de contraintes éliminées par SP parce qu'elles produisaient des poids aberrants était, en moyenne, de 2,2 pour l'estimateur GREG par projection comparativement à 0,9 pour l'estimateur pseudo-optimal. Pour l'estimateur GREG par projection,

$\hat{V}$ ' $diag(\hat{X}^{(0)})1_p$ à la première étape et $\hat{V}$ ' $diag(\hat{X}^{(A)})1_p$ à la deuxième étape, de façon à reproduire les fonctions de perte des recensements précédents.

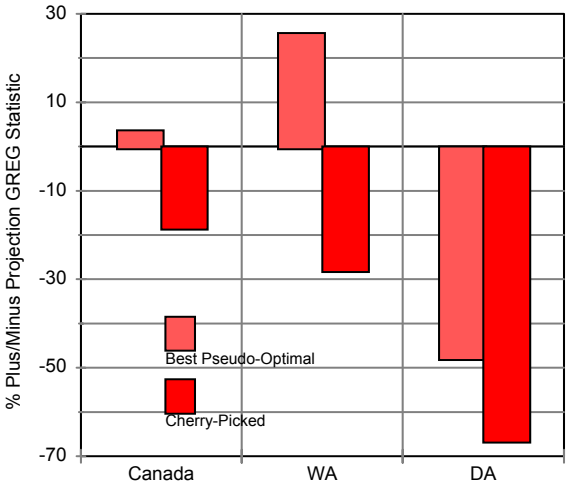
En outre, à titre de valeur de référence, nous avons effectué un passage de production en utilisant l'estimateur GREG par projection avec des poids >0, MAXC = 10 000, SMALL = 20 et POST = 1 pour reproduire les paramètres du Recensement de 1996. À la figure 2, nous comparons le passage de production sélectionné par tri minutieux et le passage de production pour l'estimateur pseudo-optimal (avec MAXC = 80 000, SMALL = 20 et POST = 1 qui a produit la valeur la plus faible du critère ABSDIFF3) à l'estimateur GREG par projection. La figure 2 montre que, comparativement à l'estimateur GREG par projection (avec poids >0), l'estimateur pseudo-optimal (avec des poids égaux ou supérieurs à 1) donne des résultats 4 % moins bons au niveau national (ABSDIFF3), 26 % moins bons au niveau du SP (ABSDIFF2), mais 49 % meilleurs au niveau de l'AD (ABSDIFF1).

La figure 2 montre aussi que les résultats du passage machine sélectionné par tri minutieux sont 19 %, 28 % et 66 % meilleurs que ceux donnés par l'estimateur GREG par projection. Au niveau du Canada, du SP et de l'AD, respectivement. Ces résultats démontrent les avantages importants qu'il y a à sélectionner minutieusement les paramètres au niveau du SP.

Tableau 2 : Paramètres utilisés pour les passages de production du recensement

Nombre de SP	Pourcentage	MAXC	SMALL	POST
1 300	21,2	160 000	20	0
1 135	18,5	80 000	20	1
903	14,7	80 000	30	1
725	11,8	80 000	30	0
539	8,8	40 000	40	1
436	7,1	20 000	40	1
363	5,9	40 000	20	1
255	4,2	40 000	30	1
251	4,1	10 000	40	1
233	3,8	20 000	30	1
1	0,0	10 000	95	0
1	0,0	160 000	50	0
6 142	100,0			

Figure 2 : Comparaison de ABSDIFF1, 2 et 3



Pour expliquer les résultats de la figure 2, il est utile d'étudier le nombre moyen de contraintes éliminées au niveau du SP et de l'AD. La figure 3 montre qu'un même nombre moyen de contraintes ont été éliminées parce qu'elles étaient petites et LD pour le meilleur passage de production au moyen de l'estimateur pseudo-optimal (avec des poids égaux ou supérieurs à 1) et pour l'estimateur GREG (avec des poids supérieurs à 0). Ce résultat n'est pas étonnant étant donné que la valeur de SMALL = 20 pour ces deux passages machine. Le passage machine sélectionné par tri minutieux correspond à l'abandon d'un nombre un peu plus élevé de contraintes parce qu'elles étaient petites (certaines étant LD), mais cette situation est compensée par l'abandon d'un nombre un peu plus faible de contraintes parce qu'elles étaient LD. Le plus grand nombre de contraintes abandonnées parce qu'elles étaient petites est dû au fait que la valeur de SMALL >20 pour certains SP pour le passage machine sélectionné par tri minutieux comme le montre le tableau 2. Le nombre de contraintes abandonnées parce qu'elles étaient PLD est le même pour le passage machine sélectionné par tri minutieux et le meilleur passage de production au moyen de l'estimateur pseudo-optimal. Pour le meilleur passage de production au moyen de l'estimateur pseudo-optimal, MAXC = 80 000, tandis que pour le passage de production sélectionné par tri minutieux, la valeur de MAXC utilisée pour les SP variait, la majorité étant MAXC $ 80 000. Comparativement, un plus grand nombre de contraintes ont été rejetées parce qu'elles étaient PLD pour l'estimateur GREG par projection. Ceci n'est pas étonnant, puisque MAXC = 10 000 pour ce passage machine. Pour le Recensement de 2001, on a utilisé des valeurs plus grandes de MAXC, suivant le conseil de Press (1992, section 2.6) selon lequel les

matrices peuvent être inversées avec une précision raisonnable à condition que l'inverse du nombre de conditions ne s'approche pas de la précision en virgule flottante de l'ordinateur. Puisque les calculs sont exécutés en double précision, il semble que les matrices dont le nombre de conditions ne s'approche pas de 10^{12} puissent être inversées avec une certaine confiance. Enfin, l'examen de la figure 3 montre que le nombre de contraintes éliminées parce qu'elles produisaient des poids aberrants est plus élevé pour le meilleur passage de production au moyen de l'estimateur pseudo-optimal que pour l'estimateur GREG par projection. Ce résultat n'est pas surprenant, puisque l'estimateur pseudo-optimal ne tolère pas les poids dont la valeur est inférieure à 1 tandis que l'estimateur GREG par projection les accepte. Il semble, si l'on s'en tient à l'analyse qui précède, que le passage de production sélectionné par tri minutieux donne de meilleurs résultats que l'estimateur GREG par projection au niveau national et des SP, car un moins grand nombre de contraintes sont éliminées parce qu'elles sont PLD (à cause des valeurs plus élevées de MAXC) et donne des résultats comparables à l'estimateur GREG par projection en ce qui concerne le nombre de contraintes éliminées parce qu'elles produisent des poids aberrants. Dans l'ensemble, l'estimateur GREG par projection est celui pour lequel le nombre de contraintes éliminées est le plus élevé et le passage de production sélectionné par tri minutieux, celui pour lequel le nombre est le plus faible.

La figure 4 montre que le nombre moyen de contraintes au niveau de l'AD (nombre de personnes et nombre de ménages) éliminées pour un SP. Elle montre que les meilleurs résultats (comme le montre la figure 2) du passage de production sélectionné par tri minutieux comparativement au passage pour l'estimateur GREG par projection au niveau de l'AD sont dus au fait qu'un nombre nettement plus faible de contraintes sont éliminées parce qu'elles sont PLD. Il en est ainsi à cause des valeurs plus élevées de MAXC utilisées en général pour le passage de production sélectionné par tri minutieux. En outre, toujours pour ce dernier, un nombre un peu plus faible de contraintes sont éliminées parce qu'elles produisent des poids aberrants.

Figure 3: Constraints Dropped WA Level

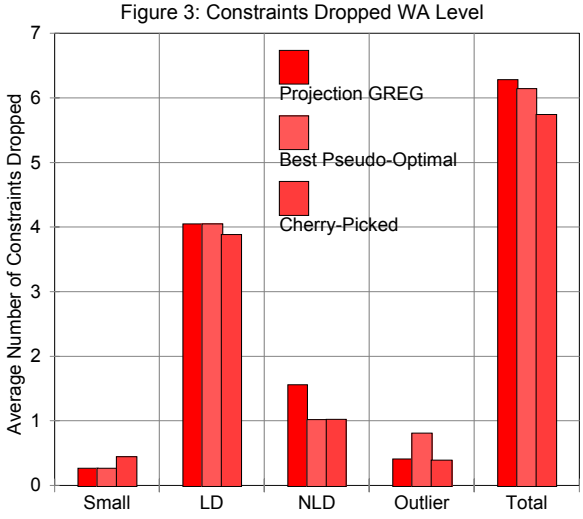
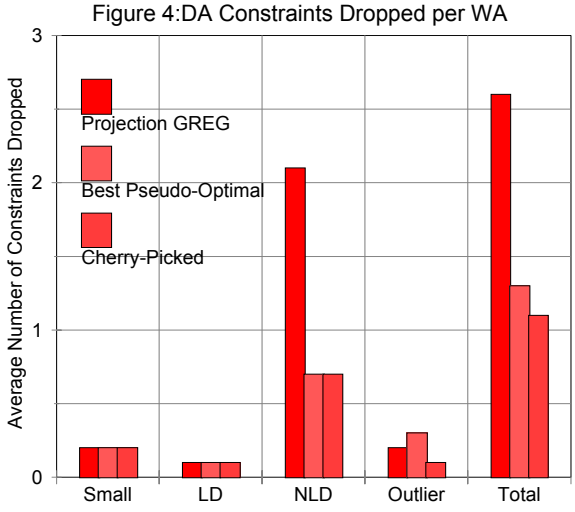


Figure 4:DA Constraints Dropped per WA



6. CONCLUSION

L'analyse réalisée à la section 3 donne à penser que le passage de l'estimateur GREG par projection à l'estimateur par régression pseudo-optimal pour le recensement donnerait lieu à l'élimination d'un moins grand nombre de contraintes parce qu'elles produisent des poids corrigés dont la valeur est inférieure à 1. Cette conclusion est confirmée numériquement à la section 5 en se fondant sur un échantillon de 121 SP. Si l'on compare le meilleur estimateur par régression pseudo-optimal pour des poids égaux ou supérieurs à 1 et l'estimateur GREG par projection pour des poids supérieurs à 0, l'écart entre les estimations-échantillon et les chiffres de population est un peu plus important au niveau national et significativement plus important au niveau du SP pour l'estimateur par régression pseudo-optimal. Cependant, l'exécution de dix passages de production pour l'estimateur par régression pseudo-optimal avec divers paramètres, suivie de la sélection par tri minutieux du meilleur passage de production pour chaque SP a produit des écarts entre les estimations-échantillon et les chiffres de population nettement plus faibles que ceux observés pour

l'estimateur GREG par projection au niveau du Canada, du SP et de l'AD. La transition du traitement des données sur l'ordinateur central à leur traitement sur ordinateur personnel pour le Recensement de 2001 a donné la capacité, du point de vue du temps et des coûts, de réaliser dix passages de production, puis de sélectionner le meilleur passage machine.

RÉFÉRENCES

Bankier, M. D. (1986), "Estimators Based on Several Stratified Samples With Applications to Multiple Frame Surveys", *Journal of the American Statistical Association*, **81**, pp. 1074-1079.

Bankier, M. D., Rathwell, S. et Majkowski, M. (1992), "Two Step Generalized Least Squares Estimation in the 1991 Canadian Census", Methodology Branch Working Paper, August 1992.

Cochran, W.G. (1942), "Sampling Theory When the Sampling Units are of Unequal Sizes", *Journal of the American Statistical Association*, **37**, pp. 199-212.

Fuller, Wayne A. (2002), "Regression Estimation for Survey Samples", *Survey Methodology*, **28**, No. 1, pp. 5-23.

Press, William H. (1992), *Numerical Recipes in C: The Art of Scientific Computer Programming*, Cambridge University Press.

Rao, J.N.K. (1994), "Estimating Totals and Distribution Functions Using Auxiliary Information at the Estimation Stage", *Journal of Official Statistics*, **10**, pp. 153-165.

Särndal, C.E. (1996), "Efficient Estimators with Simple Variance in Unequal Probability Sampling", *Journal of the American Statistics Association*, **91**, pp. 1289-1300.

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

Silva, P.L.D.N. et Skinner, C.J. (1997), "Variable Selection for Regression Estimation in Finite Populations", *Survey Methodology*, **23**, No. 1, pp. 23-32.