Assemblée annuelle de la SSC, mai 2010
Recueil de la Section des méthodes d'enquête

CALAGE AUX MARGES DES POIDS D'ENQUÊTES À PLAN COMPLEXE POUR LE REFUS À L'APPARIEMENT

François Verret[1] et Christina Kevins[2]

RÉSUMÉ

Les agences statistiques recueillent de l'information à partir de plusieurs enquêtes et sources de données administratives. Apparier ces sources augmente le pouvoir analytique tout en réduisant les besoins de collecte. Pour les enquêtes, le consentement à l'appariement par le répondant est généralement requis sans quoi on observe une non-réponse supplémentaire. Le calage aux marges des poids est une approche intéressante pour compenser cette non-réponse. Särndal et Lundström (2008) ont développé une technique de sélection des contraintes de calage minimisant une fonction du biais approximatif dû à la non-réponse. Dans cet article, la méthode est raffinée pour éviter la multicollinéarité et le calage sur un trop petit ensemble de répondants. Elle est appliquée aux données de l'Enquête sur la santé dans les collectivités canadiennes et comparée à la méthode des scores.

MOTS CLÉS : Ajustement de non-réponse; biais de non-réponse; calage aux marges; couplage d'enregistrements; multicollinéarité; sélection *Forward*.

ABSTRACT

Statistical agencies collect information via several surveys and administrative sources. Linking these sources augments the analytical potential while reducing collection needs. Survey respondents' consent is generally required for record linkage and disagreement to link is a supplementary form of non-response to the survey. Weight calibration is an interesting avenue to compensate for this non-response. Särndal and Lundström (2008) have developed a technique to select the constraints of the calibration estimator to minimize a function of the approximate non-response bias. In this paper, the method is refined to avoid multicollinearity and calibration on a small group of respondents. It is applied to the Canadian Community Health Survey and compared to the scores method.

KEY WORDS: Calibration, Forward selection, Multicollinearity, Non-response Adjustment, Non-response Bias, Record linkage.

1. INTRODUCTION

Le couplage d'enregistrements augmente l'information statistique en combinant plusieurs fichiers entre eux. Ceci vient accroître le potentiel analytique de ces fichiers tout en réduisant potentiellement les besoins en collecte de données, d'où son intérêt. En effet, les enquêtes à plan complexe lorsque couplées à des bases administratives pertinentes deviennent des sources d'information d'autant plus précieuses qu'elles ne l'étaient à l'origine. Par exemple, dans le cadre du projet des données longitudinales administratives et sur la santé (DLAS) de Statistique Canada, on appariera entre eux au travers de registres de numéros d'assurance-maladie de nombreux fichiers tels que ceux de l'Enquête sur la santé dans les collectivités canadiennes (ESCC), de la Base canadienne des données sur la mortalité, du Registre canadien du cancer et de la Base de données sur la morbidité hospitalière. Dans le cas de l'ESCC, qui est une enquête transversale, l'appariement aux nombreux autres fichiers viendra augmenter l'information disponible pour chaque répondant apparié et apportera une dimension longitudinale au fichier d'analyse résultant.

[1] François Verret, Statistique Canada, 100, prom. du pré Tunney, Ottawa, Canada, K1A 0T6, Francois.Verret@statcan.gc.ca.
[2] Christina Kevins, Université Carleton, 1125, Colonel By Drive, Ottawa, Canada, K1S 5B6, ckevins@connect.carleton.ca.

Dans cet article, on se penchera sur les fichiers d'enquête en vue d'un appariement comme ceux de l'ESCC avec comme objectif de faire de l'inférence basée sur le plan de sondage. Typiquement, de tels fichiers sont sous la forme de fichiers maîtres, c'est-à-dire que l'information fournie par chaque répondant à l'enquête y est présente en plus de données reliées à l'échantillonnage et à l'estimation telle que le poids d'enquête. C'est à partir de cette information qu'on calcule et publie des estimations. Afin de faire le couplage, un filtrage des répondants du fichier maître est souvent effectué, ce qui crée une certaine forme de non-réponse. Premièrement, le couplage d'enregistrements peut nécessiter le consentement du répondant à coupler ses données avec une autre source. Deuxièmement, pour faire un appariement efficace on doit disposer de données de bonne qualité et discriminantes. On peut alors se restreindre à ne tenter d'apparier que les enregistrements des individus ayant fourni de telles données. Dans ce qui suit, le sous-ensemble de répondants satisfaisant ces deux conditions sera appelé fichier pour couplage. Cette non-réponse crée un biais si la probabilité de faire partie du fichier pour couplage est liée à la variable étudiée. On peut alors essayer de compenser en appliquant un ajustement au poids d'enquête du fichier maître afin de produire un poids pour le fichier pour couplage. Le poids d'enquête du fichier maître sera appelé dans la suite « poids maître » et le poids du fichier pour couplage sera appelé « poids couplage. »

Dans un tel contexte, l'information pour faire l'ajustement de non-réponse a les caractéristiques suivantes :

- Elle est nombreuse car toutes les variables du fichier maître sont disponibles;
- Elle devrait être reliée aux variables à l'étude car elle comprend les variables d'intérêt de l'enquête;
- Elle pourrait être ultimement les variables réponses à l'étude dans le fichier couplé;
- Elle est connue à la fois pour les individus du fichier pour couplage et pour les individus éliminés lors du filtrage duquel découlent la non-réponse et le besoin de faire un ajustement;
- Elle peut être représentée sous la forme de totaux estimés qui pourraient avoir été publiés.

Dans ces conditions, le calage aux marges est une option intéressante. Il permet de caler sur ces derniers totaux et sur des totaux qui sont connus sans erreur et ainsi de garder une certaine concordance avec les données publiées.

Quand les totaux disponibles pour le calage sont nombreux comme dans la production du poids couplage, il peut être nécessaire de choisir les contraintes à retenir. À cette fin, Särndal et Lundström (2008) ont développé une méthode de sélection des contraintes de l'estimateur par calage qui minimise le biais approximatif dû à la non-réponse. Cette méthode novatrice est expliquée à la section suivante. À la section 3, on présente quelques adaptations et raffinements à la méthode qui sont utiles pour certains ensembles de contraintes plus complexes et quand le nombre de contraintes considérées est grand. À la section 4, un exemple d'application utilisant les données de l'ESCC est présenté. De plus, on y compare sommairement le calage à la méthode des scores suivie d'une poststratification qui est un compétiteur proche de ce qui est fait dans la pratique dans l'ESCC. Une conclusion termine l'article.

2. LA MÉTHODE DE SÉLECTION DES CONTRAINTES DE CALAGE DE SÄRNDAL ET LUNDSTRÖM

Cette section résume la théorie énoncée dans Särndal et Lundström (2008).

2.1 Information auxiliaire et estimateur par calage du total de la population

Soit U, la population finie d'intérêt, notée $U = \{1, 2, ..., k, ..., N\}$. On tire un échantillon s avec probabilités de sélection connues $\pi_k > 0$ et poids de sondage $d_k = 1/\pi_k$. Le sous-ensemble de répondants est noté r et la probabilité de réponse est définie comme $\theta_k = \mathrm{P}\left(k \in r \middle| k \in s\right) > 0$. L'objectif est d'estimer le total de la population $Y = \sum_U y_k$ en utilisant un estimateur par le calage $\hat{Y}_W$. L'information de calage disponible peut être de deux types. D'une part, certains totaux de contrôle peuvent être connus au niveau de la population. On note ces totaux $\mathbf{T}^* = \sum_U \mathbf{x}_k^*$, où $\mathbf{x}_k^*$ est l'information auxiliaire correspondante de l'individu k. Ces totaux peuvent être par exemple des comptes de population pour certains groupes donnés comme des groupes d'âge et de sexe. D'autre part, des totaux inconnus au niveau de la population U peuvent être estimés au niveau de l'échantillon complet s. On les note $\mathbf{T}^\circ = \sum_s d_k \mathbf{x}_k^\circ$, où $\mathbf{x}_k^\circ$ est l'information auxiliaire

correspondante de l'individu *k*. L'information de calage complète est donc donnée par $\mathbf{x}_k = \left(\mathbf{x}_k^{*\prime}, \mathbf{x}_k^{\circ\prime}\right)'$, avec totaux correspondants $\mathbf{T} = \left(\mathbf{T}^{*\prime}, \mathbf{T}^{\circ\prime}\right)' = \left(\sum_U \mathbf{x}_k^{*\prime}, \sum_s d_k \mathbf{x}_k^{\circ\prime}\right)'$. De plus, on suppose qu'il existe un vecteur constant $\boldsymbol{\mu}$ tel que $\boldsymbol{\mu}'\mathbf{x}_k = 1, \forall k \in U$. Dans la suite, cette condition sera satisfaite pour tous les ensembles de contraintes considérés car chacun comprendra le calage au total de la population entière. Sous cette hypothèse, l'estimateur par le calage est donné par $\hat{Y}_W = \sum_r w_k y_k = \sum_r d_k g_{kr} y_k$ où $g_{kr} = \mathbf{T}'\left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k$ est l'ajustement de calage apporté à d_k.

2.2 Ratio des biais approximatifs

L'approximation par linéarisation de Taylor du biais de l'estimateur par calage sur l'information auxiliaire $\mathbf{x}_k$ est $\text{biaisapprox}(\hat{Y}_W) = \sum_U \left(\theta_k M_k - 1\right) y_k$, où $M_k = \left(\sum_U \mathbf{x}_k\right)' \left(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k$. Särndal et Lundström (2008) comparent cette quantité au biais approximatif obtenu avec le calage le plus simple, soit le calage de base sur le total de la population uniquement, ce qui correspond à $\mathbf{x}_k = 1, \ \forall k \in U$. Dans ce cas, l'estimateur par calage est simplifié à $\hat{Y}_W = N\,\bar{y}_{r;d} = N\left(\sum_r d_k y_k\right) \Big/ \left(\sum_r d_k\right)$ (l'indice *r;d* signifie que les sommes sont faites sur l'ensemble *r* et qu'elles sont pondérées par un facteur d_k) et $\text{biaisapprox}(N\,\bar{y}_{r;d}) = N\left(\bar{y}_{U;\theta} - \bar{y}_U\right)$. La quantité théorique d'intérêt à minimiser dans le processus de sélection des contraintes de calage est le ratio de ces deux biais $P = \sum_U \left(\theta_k M_k - 1\right) y_k \Big/ \left[N\left(\bar{y}_{U;\theta} - \bar{y}_U\right)\right]$. Ce dernier est nul si y_k est une fonction linéaire de $\mathbf{x}_k, \forall k \in U$. Il l'est aussi si l'inverse de la probabilité de réponse peut s'exprimer comme une combinaison linéaire de ce même vecteur $\forall k \in U$. Särndal et Lundström (2008) définissent cette quantité par $\phi_k = 1/\theta_k$ et la nomment « influence ». Bien qu'elle soit pratiquement impossible à atteindre, la seconde condition a l'avantage par rapport à la première de rendre le ratio nul indépendamment de la variable réponse y_k étudiée. La méthode de sélection des auteurs tente donc de prédire l'influence pour minimiser le biais de non-réponse.

2.3 Prédiction des influences et indicateur du biais résiduel

La procédure de Särndal et Lundström (2008) prédit les influences théoriquement au niveau de la population en appliquant leur régression sur $\mathbf{x}_k$ tout en pondérant les observations par la probabilité de réponse θ_k. Les influences prédites sont données par $M_k = \mathbf{x}_k'\left(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1}\left(\sum_U \theta_k \mathbf{x}_k \phi_k\right)$ et leur variance pondérée est $Q^2 = \left(\sum_U \theta_k\right)^{-1} \sum_U \theta_k \left(M_k - \bar{M}_{U;\theta}\right)^2$. Cette dernière statistique est positive, atteint la valeur nulle pour le calage de base, croît si on ajoute des composantes au vecteur $\mathbf{x}_k$ et atteint son maximum (Q_{sup}) quand les influences sont parfaitement prédites. De plus, la relation entre le ratio *P* et Q^2 est approximativement linéaire : $P \approx 1 - Q^2 \big/ Q_{\text{sup}}$. Afin de minimiser *P*, on cherche le vecteur $\mathbf{x}_k$ qui maximise Q^2. En pratique, on doit estimer ces quantités théoriques par $m_k = \mathbf{x}_k'\left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1}\left(\sum_s d_k \mathbf{x}_k\right)$ et $q^2 = \left(\sum_r d_k\right)^{-1} \sum_r d_k \left(m_k - \bar{m}_{r;d}\right)^2 = \bar{m}_{r;d}\left(\bar{m}_{s;d} - \bar{m}_{r;d}\right)$. Cette dernière quantité est toujours positive; atteint la valeur nulle pour le calage de base, quand *s*=*r* ou si $\bar{\mathbf{x}}_{s;d} = \bar{\mathbf{x}}_{r;d}$; n'a pas de valeur maximale et approxime bien Q^2 si le nombre d'individus dans *s* et *r* est assez grand (>1 000 selon les auteurs). En somme, on applique une méthode de sélection des contraintes de type *Forward* en débutant la sélection avec le calage de base et on introduit une à une les contraintes de calage de manière à maximiser à chaque inclusion la statistique q^2.

3. ADAPTATION ET RAFFINEMENT DE LA MÉTHODE DE SÉLECTION

3.1 Adaptation à la production du poids couplage

La théorie développée par Särndal et Lundström peut être adaptée au contexte de la production d'un poids couplage à partir du poids maître. En effet, l'ensemble *s* correspond alors aux répondants du fichier maître, l'ensemble *r* à ceux du fichier pour couplage, le poids d_k au poids maître et le poids w_k au poids couplage. Ceci implique que $\mathbf{T}^{\circ} = \sum_s d_k \mathbf{x}_k^{\circ}$ est estimé avec le poids maître et peut correspondre à des totaux publiés.

3.2 Évaluation de l'inclusion des contraintes à un niveau désagrégé

Dans leur exemple d'application, Särndal et Lundström (2008) évaluent l'inclusion d'ensembles agrégés d'information auxiliaire tels que le groupe d'âge, le sexe et la région géographique. Par exemple, si l'ensemble correspondant au groupe d'âge est choisi, on effectue le calage sur tous les groupes d'âge sans exception. Dans cet article, pour les variables de type catégorique, on évalue plutôt l'inclusion des niveaux de ces ensembles un à la fois. On peut alors caler sur un groupe d'âge donné sans caler sur tous les autres. Cette différence permet d'avoir un peu plus de souplesse pour prédire les influences car certains niveaux peuvent être plus importants que d'autres dans un ensemble considéré. De plus, cette approche facilite l'application des autres modifications des sections suivantes.

3.3 Contraintes à faible fréquence et imbriquées

Le second raffinement vient de la forme des données disponibles pour faire le calage. D'une part, il peut exister des catégories de réponse pour lesquelles on a observé une faible fréquence (non pondérée). D'autre part, on peut observer de la non-réponse d'item pour certaines questions du fichier maître, possiblement pour un petit nombre de non-répondants seulement. Le calage sur ces cellules à faible fréquence peut avoir comme effet de rendre les estimateurs instables. On évitera dans la suite de faire un tel calage en utilisant la borne que Särndal, Swensson et Wretman (1992, p.267) suggèrent pour l'estimateur poststratifié sous un plan aléatoire simple sans remise, soit une fréquence de 20 répondants (de l'ensemble *r*) dans la cellule. Appliquer cette règle est davantage compliqué quand les contraintes considérées sont imbriquées. Par exemple, on peut être intéressé à comparer à la fois le calage sur des groupes d'âge, le calage sur le sexe et le calage sur ces groupes d'âge croisés avec le sexe. Si une certaine combinaison d'âge et de sexe contient moins de 20 répondants, il faudra s'assurer de ne pas caler directement sur ce groupe. Il faudra également éviter de caler par construction indirectement (appelé calage indirect) sur ce groupe au travers des autres contraintes choisies, ce qui est plus difficile à faire. Dans des situations simples, la résolution manuelle est envisageable. Cependant, dans des cas complexes, comme dans celui de la production du poids couplage, il est plus approprié d'adopter une approche systématique et automatisée. La section suivante présente une telle approche qui utilise la théorie de la régression linéaire.

3.4 Méthode pour éviter le calage sur les contraintes à faible fréquence

On définit d'abord les matrices d'information auxiliaire (sous la forme de vecteurs lignes $\mathbf{x}_k$) et la partition suivante $\mathbf{X}^{\text{total}} = \mathbf{1} \left| \mathbf{X}^{\text{à éviter}} \right| \mathbf{X}^{\text{potentielles}}$. $\mathbf{X}^{\text{total}}$ est la matrice contenant l'ensemble de l'information auxiliaire considérée. Le vecteur unitaire correspond au calage de base et au point de départ de l'approche de sélection *Forward*. $\mathbf{X}^{\text{à éviter}}$ représente l'information auxiliaire des contraintes à éviter qui correspondent à de trop petits groupes de répondants. $\mathbf{X}^{\text{potentielles}}$ est le reste de l'information auxiliaire qu'on envisage inclure dans le calage.

Le calage direct aux contraintes de $\mathbf{X}^{\text{à éviter}}$ peut être évité en ne considérant pas les colonnes de cette matrice dans la sélection. Cependant, le calage indirect à ces contraintes pourrait être effectué par construction par le choix d'une combinaison de contraintes de $\mathbf{1} \left| \mathbf{X}^{\text{potentielles}} \right.$. Soit $\mathbf{X}^{\text{danger}}$, la sous-matrice des colonnes de $\mathbf{X}^{\text{à éviter}}$ qui peuvent être obtenues par une telle combinaison. Cette matrice peut être construite en combinant les colonnes de $\mathbf{X}^{\text{à éviter}}$ contenues dans l'espace généré par les colonnes de $\mathbf{1} \left| \mathbf{X}^{\text{potentielles}} \right.$. Soient $\mathbf{X}^{\text{choisies}}$ et $\mathbf{X}^{\text{rejetées}}$, les matrices des contraintes de

$\mathbf{1}\big|\mathbf{X}^{\text{potentielles}}$ choisies et rejetées à une étape donnée du processus. La sélection *Forward* débute avec le calage de base, donc avec $\mathbf{X}^{\text{choisies}} = \mathbf{1}$ et $\mathbf{X}^{\text{rejetées}}$ vide. On ajoute une à une des contraintes en suivant les étapes suivantes.

1. Pour chacun des vecteurs de $\mathbf{X}^{\text{potentielles}}$ qui ne font pas déjà partie de $\mathbf{X}^{\text{choisies}}$ ou de $\mathbf{X}^{\text{rejetées}}$:
   a. On vérifie qu'il ne fait pas partie de l'espace généré par les colonnes de $\mathbf{X}^{\text{danger}}\big|\mathbf{X}^{\text{choisies}}$, sans quoi on l'élimine en l'incluant dans $\mathbf{X}^{\text{rejetées}}$. La raison est que d'une part il se peut qu'on ait déjà calé indirectement pour cette variable dans $\mathbf{X}^{\text{choisies}}$ et son inclusion serait alors inutile. D'autre part, son inclusion peut être équivalente à caler indirectement sur une contrainte de $\mathbf{X}^{\text{danger}}$.
   b. S'il n'est pas rejeté, on calcule le q^2 qui correspondrait à son inclusion dans $\mathbf{X}^{\text{choisies}}$.
2. On augmente $\mathbf{X}^{\text{choisies}}$ de la contrainte ayant obtenu le plus grand q^2.

On répète ces étapes jusqu'à un certain critère, par exemple: toutes les contraintes ont été éliminées ou sélectionnées, le q^2 plafonne ou le nombre de contraintes incluses est suffisamment grand.

Cette approche offre des avantages intéressants pour la production du poids couplage étant donné l'étendue de l'information de calage disponible. Premièrement, elle permet d'éviter le calage direct ou indirect sur des groupes trop restreints de répondants qu'on a identifiés au préalable. Deuxièmement, comme elle ne requiert pas d'intervention manuelle, elle permet d'inclure de façon systématique un grand nombre de contraintes, qu'elles soient imbriquées ou non. Troisièmement, elle limite dans une certaine mesure l'instabilité des estimateurs de calage et elle peut être élargie à cette fin pour éviter la multicollinéarité dans les contraintes choisies comme il le sera démontré à la section suivante. Finalement, appliquer cette approche est comparable à un regroupement guidé par la non-réponse des cellules à faible fréquence aux autres cellules.

Cependant, l'approche comporte aussi ses failles. L'étape 1 peut éliminer des contraintes qui ne représentent pas un véritable danger. En effet, l'espace de $\mathbf{X}^{\text{danger}}\big|\mathbf{X}^{\text{choisies}}$ est trop large à l'étape 1a et c'est plutôt les sous-espaces composés de $\mathbf{X}^{\text{choisies}}$ et des colonnes de $\mathbf{X}^{\text{danger}}$ prises une à une qui sont vraiment problématiques. De plus, la méthode n'est pas exhaustive puisqu'elle évite le calage sur les petits groupes pré-identifiés et non sur tous les petits groupes possibles. D'autre part, pour se rassurer, on peut identifier le plus petit groupe que l'on peut former à partir des contraintes finales choisies à l'aide de la méthode du simplexe et vérifier que la fréquence de ce groupe est suffisamment élevée. On note enfin que du point de vue des outils diagnostiques de sélection, les changements à l'approche de Särndal et Lundström pourront produire un graphique de q^2 en fonction du nombre de contraintes choisies en dents de scies, tandis que l'approche pure des auteurs donne un graphique à l'allure lisse dont la pente est positive et décroissante. Effectivement, il y a plus de dépendance entre les accroissements de l'approche désagrégée que dans l'approche pure.

4. EXEMPLE

La méthode de sélection des contraintes de la section précédente est testée ici dans le but de la mettre en œuvre dans le cadre du projet de DLAS afin de produire un poids couplage pour l'ESCC. La prochaine section présente le contexte dans lequel la méthode a été appliquée. La section 4.2 traite des choix informatiques qui ont été faits. En 4.3 on compare le calage à un compétiteur : la méthode des scores pour corriger la non-réponse suivie d'une poststratification.

4.1 Contexte de l'étude

Pour appliquer le calage, le fichier pour couplage de l'ESCC cycle 3.1 est utilisé. L'ESCC est une enquête transversale mesurant les déterminants de la santé et l'état de santé des Canadiens et l'état du système de soins de santé. Un échantillon de près de 130 000 personnes est choisi pour cette enquête. Des estimations sont produites et publiées à partir du fichier maître pour plus de 120 régions sociosanitaires (RSS) croisées avec dix groupes d'âge et de sexe cibles. L'enquête représente environ 98 % de la population canadienne qui était âgée de 12 ans ou plus en 2005. L'estimation ponctuelle et l'estimation de la variance fondées sur le plan de sondage sont faites à l'aide du poids maître et de poids bootstrap de Rao, Wu et Yue (1992). On n'utilise que les données des quatre provinces de l'Atlantique pour simplifier l'étude. Ceci correspond à 21 RSS et à un ensemble *s* de taille 16 302 représentant environ 2 millions de personnes.

Pour les fins de la production du poids couplage, on considère comme répondants les répondants du fichier maître ayant à la fois donné leur consentement à l'appariement de leurs données et fourni un numéro d'assurance-maladie valide. Cet ensemble *r* contient 10 750 personnes, soit environ les deux tiers de l'ensemble *s*. Le calage est réalisé indépendamment par province. L'information disponible pour faire les ajustements de calage comprend entre autres : l'ensemble de l'information se trouvant sur le fichier maître,; les totaux utilisés à l'étape de poststratification du fichier maître, soit les projections démographiques basées sur le Recensement de 2001 par RSS par groupe d'âge par sexe (ces totaux seront considérés comme connus et fixes dans la suite), des variables du plan de sondage et des paradonnées. Les variables considérées dans la sélection sont toutes catégoriques et sont les indicateurs de groupes de poststratification identifiant la RSS, le groupe d'âge et le sexe (avec totaux $\mathbf{T}^*$) et 24 variables comprenant les indicateurs de santé principaux de l'ESCC et les variables sociodémographiques les plus importantes (avec totaux $\mathbf{T}^\circ$). La sélection des contraintes est faite jusqu'à ce que toutes les variables considérées soient choisies ou éliminées du processus de sélection.

4.2 Choix informatiques

SAS/IML a été utilisé à cause des capacités de SAS à manipuler de grands jeux de données et parce que IML facilite la manipulation des nombreuses matrices du processus. En effet, les calculs de m_k et de g_{kr} et la sélection des contraintes impliquent le calcul et l'inversion de plusieurs matrices de la forme $\mathbf{A} = \sum_r d_k \mathbf{x}_k \mathbf{x}_k'$. De plus, ces matrices croissent à mesure que des contraintes sont ajoutées et peuvent devenir grandes et difficiles à inverser. Pour atténuer ce problème, certaines optimisations sont effectuées dans le programme. D'abord, les vecteurs lignes de $\mathbf{X}^{\text{Total}}$ sont centrés et réduits pour que $\sum_r d_k \mathbf{x}_k = \mathbf{0}$ et $\sum_r d_k \mathbf{x}_k^2 = \mathbf{1}$, sauf pour la composante provenant du vecteur $\mathbf{1}$ qui ne satisfait que la seconde condition. Les totaux $\mathbf{T}$ sont centrés et réduits en conséquence. Ceci stabilise numériquement $\mathbf{A}$ en rendant tous les éléments de la diagonale égaux à 1, sans avoir d'effet sur le problème de minimisation. Elle lui donne également la forme d'une matrice de corrélation, ce qui facilite l'interprétation et les calculs subséquents. On tire aussi profit de la décomposition $\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{a}_{12} \\ \mathbf{a}_{12}' & a_{22} \end{pmatrix}$ et de l'identité $\mathbf{A}^{-1} = \frac{1}{b}\begin{pmatrix} b\mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{a}_{12}\mathbf{a}_{12}'\mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{a}_{12} \\ -\mathbf{a}_{12}'\mathbf{A}_{11}^{-1} & 1 \end{pmatrix}$ où $b = a_{22} - \mathbf{a}_{12}'\mathbf{A}_{11}^{-1}\mathbf{a}_{12}$.

De plus, comme l'inclusion d'une nouvelle contrainte n'a pas d'impact sur la composante $\mathbf{A}_{11}$, on limite le nombre d'inversions à effectuer en conservant judicieusement les inverses calculées et en bâtissant sur ces inverses. Finalement, on pourrait vérifier si *b* est nul pour appliquer l'étape 1a de la section 3.4. On a plutôt élargi cela en vérifiant que *b* n'était pas trop proche de 0 afin d'éviter la multicollinéarité entre les contraintes de calage comme le décrivent Belsey, Kush et Welsch (2004, p.92-93) dans le cadre d'une régression linéaire. Effectivement, en raison de la standardisation des vecteurs d'information auxiliaire, le terme *b* est égal à 1 moins le *R*-carré correspondant à la régression de la nouvelle contrainte sur celles déjà incluses. Il est aussi équivalent à l'inverse de l'indicateur VIF (*variance inflation factor*), qui donne l'inflation de la variance due à l'ajout de la nouvelle variable dans un modèle de régression linéaire incluant les vecteurs déjà choisis. On s'est assuré que *b* était plus grand que 0,01 et donc que le VIF était plus petit que 100.

Une fois que les contraintes de calage sont choisies, le programme donne les ajustements de calage à appliquer g_{kr}. Cependant, on peut être intéressé à fixer certaines bornes sur ces ajustements ou sur les poids résultants $w_k = d_k g_{kr}$ (comme w_k>0). L'ensemble de macros SAS StatMx de Statistique Canada qui fait suite au Système généralisé d'estimation du même organisme, a été utilisé pour calculer les ajustements de manière à garantir un poids minimum de un. D'autre part, pour estimer la variance par la méthode du Bootstrap on pourrait répéter le processus de sélection des contraintes pour chaque ensemble de poids. Cependant, cette façon aurait eu comme inconvénient de prendre beaucoup de ressources informatiques. On a plutôt gardé le même ensemble de contraintes pour chaque réplique, en prenant soin de recalculer la partie $\mathbf{T}^\circ$ en fonction de la réplique. On estime alors par réplication la variance étant donné l'ensemble de contraintes choisi. Une autre qualité de StatMx est qu'il réduit au besoin l'ensemble de contraintes à appliquer au plus grand ensemble de contraintes qu'il est possible d'appliquer. Cette situation peut survenir dans une réplique si le groupe correspondant à une certaine contrainte devient vide suite au sous-échantillonnage.

4.3 Comparaison à un compétiteur : la méthode des scores suivie d'une poststratification

Dans cette section, le calage est comparé à une méthode d'ajustement des poids proche de ce qui a été fait dans la pratique dans l'ESCC pour produire le poids couplage. L'ajustement se fait en deux étapes. À la première, on applique la méthode des scores. On ajuste un modèle de régression logistique pour prédire la probabilité θ_k. On construit le modèle à partir des mêmes variables que celles qui ont été considérées pour le calage et on applique une méthode *Stepwise* de sélection avec comme paramètres des probabilités de 10 % à l'entrée et de 20 % à la sortie. À partir des probabilités de réponse prédites du modèle final, on forme des grappes de manière à ce qu'elles contiennent des répondants avec des probabilités de réponse similaires. On combine ensuite ces grappes en groupes homogènes de réponse (GHR) afin de garantir dans chacun un minimum de 20 répondants et un taux de réponse non pondéré minimal de 40 % (pour imiter la pondération de l'ESCC). On cale ensuite le poids des répondants sur le total du poids de son GHR. La deuxième étape de pondération consiste à faire une poststratification par RSS, par groupe d'âge et par sexe aux projections démographiques.

Le tableau 1 donne quelques statistiques sur les résultats du calage et de son compétiteur par province. Un faible nombre de contraintes à faible fréquence devait être évité dans la sélection, entre deux et quatre par province. Au niveau du calage, un grand nombre de contraintes a été éliminé pour cause de multicollinéarité (près de 40 % des contraintes), la condition sur le VIF étant assez restrictive. Les ajustements de calage extrêmes présentés sont ceux avant l'application de la borne inférieure à l'aide de StatMx et ils varient entre -1 et 5,22, la vérification de la multicollinéarité ayant probablement pour effet de limiter cette étendue. Dans toutes les provinces, la première variable incluse dans le calage, celle qui est le plus fortement associée à la non-réponse, correspond aux individus qui ont refusé de répondre à la question sur le revenu personnel. Cette catégorie est considérée dans le calage car elle est le complément de toutes les classes de revenu personnel possibles. L'augmentation du q^2 correspondante représente plus de la moitié de l'accroissement total du q^2 une fois que toutes les contraintes considérées sont triées. Les dernières contraintes choisies dans le processus contribuent peu à ce total, ce qui laisse supposer que leur addition ne corrigera pas le biais de non-réponse de façon importante. Leur inclusion peut tout de même être intéressante parce qu'elle fait en sorte qu'un plus grand nombre de totaux du fichier pour couplage correspondent aux totaux du fichier maître. Pour la méthode des scores, un bien plus petit nombre de variables a été choisi. La première variable choisie est aussi la non-déclaration du revenu personnel pour toutes les provinces et l'ordre de sélection des autres variables diffère de celui du calage. L'ajustement final au poids maître est le produit des deux ajustements marginaux successifs correspondants à la méthode des scores et à la poststratification. Les minimums sont plus élevés et les maximums sont moins élevés qu'avec la méthode du calage.

Tableau 1 – Statistiques sur l'implémentation du calage et de son compétiteur

	T.-N. et Labrador		Î.-P.-É.		N.-É.		N.-B.	
	Calage	Scores + poststr.	Calage	Scores + poststr.	Calage	Scores + poststr.	Calage	Scores + poststr.
Contraintes considérées	197	197	187	187	234	234	245	245
Contraintes à faible fréquence	2	2	4	4	2	2	4	4
Contraintes choisies	120	38	108	28	141	23	143	38
Ajustement minimal global	-0,33	0,91	-0,03	0,87	-1,00	0,95	0,00	0,93
Ajustement maximal global	4,92	3,62	3,75	3,30	5,22	3,14	4,71	3,40

Afin de comparer la performance des deux approches, des estimations de proportions ont été produites à l'aide des deux ensembles de poids couplage résultants. Elles ont été comparées aux estimations du fichier maître et deux statistiques ont été produites. La première estime la racine carrée de l'erreur quadratique moyenne (reqm) et permet d'évaluer à quel point les estimateurs sont précis : $\text{reqm}\left(\hat{p}_{\text{Couplage}}\right) = \sqrt{\hat{V}_B\left(\hat{p}_{\text{Couplage}}\right) + \left(\hat{p}_{\text{Couplage}} - \hat{p}_{\text{Maître}}\right)^2}$. La seconde statistique mesure la contribution du biais à l'erreur quadratique moyenne. Elle est définie par $\left|\hat{p}_{\text{Couplage}} - \hat{p}_{\text{Maître}}\right| \Big/ \text{reqm}\left(\hat{p}_{\text{Couplage}}\right)$ et évalue à quel point l'estimation du fichier pour couplage est proche de l'estimation du fichier maître. Deux ensembles de variables ont été utilisés séparément pour l'estimation des proportions. Le premier est l'ensemble des variables considérées dans la sélection des contraintes de calage, soient les indicateurs principaux de santé, des variables

sociodémographiques ainsi que les indicateurs de poststrates. On évalue à partir de cet ensemble la capacité des méthodes à conserver intactes les valeurs du fichier maître. Le deuxième ensemble est constitué des mêmes variables, mais les proportions sont calculées à un niveau plus fin, soit au niveau du groupe d'âge et de sexe dans la province. Ceci permet d'évaluer la concordance pour des variables qui ne pouvaient pas être choisies.

Tableau 2 – Résultat des comparaisons

	Statistique	Méthode	Nombre de variables	Q1	Médiane	Q3
Variables pouvant faire partie des modèles	REQM	Calage	428	0,0044	0,0065	0,0087
		Scores + ps	428	0,0058	0,0081	0,0113
	Contribution du biais	Calage	428	0	0	0,0252
		Scores + ps	428	0,1308	0,2720	0,5272
Variables ne pouvant pas faire partie des modèles	REQM	Calage	2 836	0,0227	0,0323	0,0453
		Scores + ps	2 836	0,0227	0,0321	0,0449
	Contribution du biais	Calage	2 836	0,1782	0,4002	0,7023
		Scores + ps	2 836	0,1743	0,3918	0,6926

Le tableau 2 donne le nombre de variables dans chaque ensemble ainsi que les quartiles des deux statistiques considérées pour chaque ensemble, statistique et méthode considérés. Pour le premier ensemble de variables, on remarque que le calage a des reqm plus petits que ceux du compétiteur. Ceci est dû au biais comme le confirme l'autre statistique qui est nulle au moins une fois sur deux. Ceci n'est pas surprenant car au moins la moitié des contraintes de calage considérées sont satisfaites. L'estimateur par calage donne donc des valeurs qui concordent mieux avec celles du fichier maître pour cet ensemble de variables. Pour l'autre ensemble de variables, celles ne pouvant pas faire partie des modèles, les deux méthodes donnent des résultats comparables. Aucune des deux n'a donc l'avantage sur l'autre dans ce cas.

5. CONCLUSION

La méthode de sélection des contraintes de calage de Särndal et Lundström permet de choisir un ensemble de contraintes minimisant le biais approximatif dû à la non-réponse de l'estimateur par calage. Les changements proposés à cette méthode dans cet article amènent quelques avantages pratiques tels qu'éviter les contraintes à faible fréquence et celles présentant de la multicollinéarité. Cet article s'intéressait au cas particulier de la production d'un ensemble de poids pour un fichier destiné à l'appariement à partir de poids d'un fichier maître duquel des estimations avaient été publiées. Dans l'exemple étudié, l'estimateur par calage présenté donne des résultats similaires à l'application de la méthode des scores suivie d'une poststratification, sauf pour les variables sur lesquelles on a calé où le calage a l'avantage de donner des estimateurs concordant mieux à ceux du fichier maître. Il faut cependant noter que toutes les contraintes étudiées dans cet exemple étaient de type catégorique, et qu'on pourrait probablement discriminer les deux approches davantage avec des variables continues. On pourrait vérifier l'hypothèse qu'une relation linéaire entre l'influence et une variable continue avantagerait la méthode du calage et qu'une relation logistique entre la probabilité de réponse et une variable continue avantagerait la méthode des scores. Finalement, les changements à la méthode de Särndal et Lundström présentés ici pourraient avoir d'autres applications intéressantes, comme dans le calage régulier où on ne corrige pas pour une non-réponse et où on doit faire un choix quant aux contraintes à choisir.

RÉFÉRENCES

Belsey, D.A., Kuh, E. et Welsch, R.E. (2004). *Regression diagnostics* (2e éd.). New York: John Wiley & Sons, Inc.

Rao, J.N.K., Wu, C.F.J. et Yue, K. (1992). « Some Recent work on Resampling Methods for Complex Surveys ». *Techniques d'enquête*, Vol. **18**, No. 2, pp.209-217.

Särndal, C.-E. et Lundström, S. (2008). « Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator ». *Journal of Official Statistics*, Vol. **24**, No. 2, 167-191

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag, Inc.