Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

La méthode de calage dans la théorie et la pratique des enquêtes

par Carl-Erik Särndal

Décembre 2007

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2007
Vol. 33, n° 2, pp. 113-135
Statistique Canada, n° 12-001-X au catalogue

La méthode de calage dans la théorie et la pratique des enquêtes

Carl-Erik Särndal [1]

Résumé

Le calage est le thème central de nombreux articles récents sur l'estimation dans le contexte de l'échantillonnage. Des expressions telles que « méthode de calage » et « estimateur par calage » sont fréquentes. Comme tiennent à le souligner les auteurs de ces articles, le calage offre un moyen systématique d'intégrer des données auxiliaires dans la procédure.

Le calage est devenu un instrument méthodologique important dans la production de statistiques à grande échelle. Plusieurs organismes statistiques nationaux ont conçu des logiciels de calcul des poids, qui sont généralement calés sur les données auxiliaires disponibles dans les registres administratifs et d'autres sources de données fiables.

Le présent article fait le point sur la méthode de calage en mettant l'accent sur les progrès accomplis depuis une dizaine d'années. Le nombre d'études sur le calage augmente rapidement et nous abordons ici certaines des questions soulevées.

L'article débute par une définition de la méthode de calage, suivie d'une revue des caractéristiques importantes de cette méthode. L'estimation par calage est comparée à l'estimation par la régression (généralisée), qui est un autre moyen, conceptuellement différent, de tenir compte de l'information auxiliaire. Vient ensuite une discussion des aspects mathématiques du calage, y compris les méthodes permettant d'éviter les poids extrêmes. Dans les premières sections sont décrites des applications simples de la méthode, c'est-à-dire l'estimation d'un total de population sous échantillonnage direct, à une seule phase. Puis est envisagée la généralisation à des paramètres et à des plans d'échantillonnage plus complexes. Un trait commun de ces plans (à au moins deux phases ou deux degrés) est que l'information auxiliaire disponible peut comporter plusieurs composantes ou couches. L'application du calage dans de tels cas d'information composite est passée en revue. Plus loin, des exemples sont donnés pour illustrer comment les résultats de l'approche du calage peuvent différer de ceux obtenus grâce aux approches établies antérieurement. Enfin sont discutées des applications du calage en présence d'erreurs non dues à l'échantillonnage, en particulier les méthodes de correction du biais de non-réponse.

Mots clés : Conhérence; estimateur par régression; inférence basée sur le plan; information auxiliaire; modèles; non-réponse; plan de sondage complexe; pondération.

1. Introduction

1.1 Définition du calage

Il serait bon, pour les besoins du présent article, de faire référence à une définition du calage. Voici celle proposée ici.

Définition. L'application de la *méthode de calage* à l'estimation pour population finie consiste à :

a) calculer des poids qui tiennent compte de l'*information auxiliaire* spécifiée et sont soumis à des contraintes précisées par une ou plusieurs *équations de calage*;

b) utiliser ces poids pour calculer des estimations linéairement pondérées des totaux et d'autres paramètres de population finie, c'est-à-dire multiplier la valeur de la variable par le poids et faire la somme sur un ensemble d'unités observées;

c) se fixer l'objectif d'obtenir des estimations presque sans biais sous le plan de sondage, à condition qu'il n'y ait pas d'erreur de non-réponse ni d'autres erreurs non dues à l'échantillonnage.

Dans la littérature, le terme « calage » fait souvent référence à la partie a) seulement. Ici, le terme sera souvent appliqué aux parties a) à c) regroupées. Les définitions antérieures, quoique moins générales, concordent essentiellement avec la présente définition. Ardilly (2006) définit le calage (ou, plus précisément, le « calage généralisé ») comme une méthode de repondération utilisée lorsqu'on a accès à plusieurs variables, qualitatives ou quantitatives, sur lesquelles on souhaite effectuer, conjointement, un ajustement.

Kott (2006) définit les poids de calage comme un ensemble de poids, pour les unités de l'échantillon, qui sont calées sur les totaux connus de population de sorte que l'estimateur résultant soit convergent dans des conditions de randomisation (convergent sous le plan) ou, de manière plus rigoureuse, que le biais par rapport au plan de sondage représente, sous des conditions faibles, un apport asymptotiquement non significatif à l'erreur quadratique moyenne de l'estimateur, propriété appelée ici « presque sans biais sous le plan de sondage ».

La quatrième édition des Lignes directrices concernant la qualité de Statistique Canada (2003) dit ce qui suit : « Le *calage aux marges* est une procédure qu'on peut appliquer pour incorporer des données auxiliaires. Cette procédure rajuste les poids d'échantillonnage au moyen de multiplicateurs appelés les *facteurs de calage*, lesquels font correspondre les estimations aux totaux connus. Les poids

1. Carl-Erik Särndal, 2115 Erinbrook Crescent, #44, Ottawa, (Ontario), K1B 4J5, Canada. Courriel : carl.sarndal@rogers.com.

obtenus sont appelés les *poids de calage* ou les *poids d'estimation finaux*. De façon générale, ces poids de calage donneront des estimations convergentes qui admettront une variance d'échantillonnage plus faible que celle obtenue au moyen de l'estimateur Horvitz Thompson. »

La partie c) de la définition appelle un commentaire. Rien n'empêche de produire des poids calés sur des données auxiliaires données sans que c) soit nécessaire. Toutefois, la plupart des travaux publiés sur le calage sont dans l'esprit de c), d'où l'intérêt de l'inclure. En présence d'erreurs non dues à l'échantillonnage, les estimations sont inévitablement entachées d'un biais, qu'elles soient produites par calage ou par toute autre méthode. Allant dans le sens de c), l'inférence fondée sur le plan de sondage est considérée comme étant la situation de référence dans le présent article. La variance d'un estimateur par rapport au plan de sondage est donc une caractéristique importante. Toutefois, l'article est axé sur les « raisons qui motivent l'estimation (ponctuelle) » et faute d'espace, l'importante question de l'estimation de la variance n'y est pas abordée.

1.2 Commentaires

La définition de la section 1.1 suscite certains commentaires et des renvois aux études publiées antérieurement.

1) *Le calage comme méthode de pondération linéaire*. Le calage est intimement lié à la pratique. La fixation des grands organismes statistiques nationaux sur les méthodes de pondération est une puissante force poussant le calage. Attribuer un poids approprié à une valeur observée d'une variable et faire la sommation des valeurs pondérées de la variable pour former des agrégats appropriés est une pratique fermement enracinée. Elle est utilisée par les organismes statistiques pour estimer divers paramètres descriptifs de population finie, dont les totaux, les moyennes et les fonctions de totaux. La pondération est facile à expliquer aux utilisateurs des données et aux autres intervenants des organismes statistiques.

La pondération des unités par l'inverse de leur probabilité d'inclusion trouve depuis longtemps ses fondements scientifiques dans des articles tels que ceux de Hansen et Hurwitz (1943) et d'Horwitz et Thompson (1952). La pondération a fini par être généralement reconnue. Plus tard, la pondération par poststratification l'a été à son tour. La pondération par calage s'inscrit dans le prolongement de ces deux idées. La pondération par calage est fonction du résultat et les poids dépendent de l'échantillon observé.

Par définition, les poids correspondant à l'inverse de la probabilité d'inclusion sont égaux ou supérieurs à l'unité. Une interprétation courante est qu'une unité observée est représentative d'elle-même et d'un certain nombre d'autres unités, non observées. En revanche, les poids calés ne sont pas nécessairement égaux ou supérieurs à l'unité, à moins que l'on ne prenne tout spécialement soin dans les calculs d'obtenir cette propriété.

Le terme « calage » est nouveau en échantillonnage - il remonte à une quinzaine d'années - mais l'application de la méthode pour produire des pondérations ne l'est pas. Il y a du vrai dans ce que disent ceux qui affirment avoir fait du calage bien avant qu'on l'ait baptisé ainsi. La méthode a gagné en portée et en attrait au cours des 15 dernières années. Une méthode de pondération voisine du calage est utilisée de longue date par les instituts de sondage privés, par exemple, dans le contexte de l'échantillonnage par quota, forme d'échantillonnage probabiliste qui dépasse le cadre du présent exposé.

La pondération des valeurs observées d'une variable était déjà un sujet important avant que le calage devienne à la mode. Certains auteurs calculaient les pondérations en posant qu'elles devaient s'écarter aussi peu que possible des poids de sondage sans biais (c'est-à-dire l'inverse des probabilités d'inclusion). D'autres obtenaient les poids en reconnaissant qu'un estimateur par la régression linéaire pouvait s'écrire sous la forme d'une somme linéairement pondérée des valeurs observées de la variable étudiée. Ils ont utilisé des expressions telles que « pondération par les poids de sondage » (*survey sample weighting*), « pondération par régression » (*regression weighting*) et « pondération de cas » (*case weighting*). Ces « premiers articles » comptent Alexander (1987), Bankier, Rathwell et Majkowski (1992), Bethlehem et Keller (1987), Chambers (1996), Fuller, Loughin et Baker (1994), Kalton et Flores-Cervantes (1998), Lemaître et Dufour (1987), Särndal (1982) et Zieschang (1990). La méthode de « pondération répétée », avancée par l'organisme statistique national des Pays-Bas (CBS) sera commentée plus loin. Le terme « calage », plus récent, communique un message plus précis et une orientation plus catégorique que le terme « pondération ».

2) *Le calage comme moyen systématique d'utiliser l'information auxiliaire*. Le calage représente un moyen systématique de tenir compte des données auxiliaires. Comme le souligne Rueda et coll. (2007), dans de nombreuses conditions ordinaires, le calage offre un moyen simple et pratique d'intégrer l'information auxiliaire dans l'estimation.

L'information auxiliaire a été utilisée pour améliorer l'exactitude des estimations par sondage bien avant que le calage gagne en popularité. De nombreux articles ont été rédigés en poursuivant cet objectif, dans le contexte de situations plus ou moins spécialisées. Aujourd'hui, le calage donne une vue systématique des utilisations de l'information auxiliaire. Par exemple, il permet de traiter efficacement les données d'enquêtes pour lesquelles l'information auxiliaire existe à divers niveaux. Dans

l'échantillonnage à deux degrés, des données peuvent exister pour les unités d'échantillonnage de premier degré (grappes) et d'autres, pour les unités d'échantillonnage de deuxième degré. Dans les enquêtes avec non réponse (c'est-à-dire essentiellement toutes les enquêtes), des données peuvent exister « au niveau de la population » (totaux de population connus) et d'autres, « au niveau de l'échantillon » (valeurs des variables auxiliaires pour toutes les unités échantillonnées, répondantes et non répondantes). Le calage à l'aide de « données composites » est examiné aux sections 8 et 9.

L'estimation par la régression, ou estimation par la régression généralisée (GREG) fait concurrence au calage en tant que moyen systématique d'intégrer l'information auxiliaire. Il importe donc de faire la distinction entre l'estimation GREG (décrite à la section 3) et l'estimation par calage (décrite à la section 4). Les deux approches sont différentes.

3) *Le calage pour obtenir la convergence*. Le calage est souvent décrit comme un moyen d'obtenir des estimations convergentes. (Ici « convergence » ne signifie pas « convergence sous la plan d'échantillonnage (randomisation) », mais « convergence avec des agrégats connus ».) Les *équations de calage* imposent la convergence sur le système de poids, de sorte que, lorsqu'ils sont appliqués aux variables auxiliaires, ils confirment (concordent avec) les agrégats connus pour ces mêmes variables auxiliaires. Le désir de rendre les statistiques publiées plus crédibles est dans bien des cas le motif cité de la recherche de convergence. Certains utilisateurs des statistiques n'aiment pas constater qu'une même grandeur de population est estimée par deux chiffres ou plus qui ne concordent pas.

Les totaux par rapport auxquels est recherchée la convergence sont parfois appelés totaux de contrôle. Des expressions comme « poids contrôlés » ou « poids calés » donnent une impression d'estimation améliorée, plus exacte. Le terme « calage » a aussi la connotation de « stabilité ».

Obtenir la convergence par calage a une incidence plus générale que la simple concordance avec des totaux de population auxiliaires connus. On pourrait, par exemple, rechercher la convergence avec des totaux *estimés* de manière appropriée, d'après les données de l'enquête courante ou d'autres enquêtes.

La convergence entre les tableaux estimés d'après des enquêtes différentes est la raison qui motive la *pondération répétée*, méthode mise au point par l'organisme statistique national des Pays-Bas (CBS) et exposée dans plusieurs articles, dont Renssen et Nieuwenbroek (1997), Nieuwenbroek, Renssen et Hofman (2000), Renssen, Kroese et Willeboordse (2001), ainsi que Knottnerus et van Duin (2006). L'objectif énoncé est de répondre aux demandes des utilisateurs de produire des données de sortie numériquement convergentes. Comme le soulignent les auteurs du dernier article mentionné, la pondération répétée peut être considérée comme une étape de calage supplémentaire en vue d'un nouvel ajustement des poids déjà calés. Les poids finaux réalisent la convergence avec les valeurs de marge données.

La convergence avec des totaux connus ou estimés peut avoir l'avantage supplémentaire d'améliorer l'exactitude (réduction de la variance et/ou du biais de non réponse). Toutefois, dans certains articles, particulièrement ceux publiés par les organismes statistiques, la convergence en vue de satisfaire les utilisateurs semble un motif plus impératif que la perspective d'une plus grande exactitude.

Si la principale motivation du calage n'est pas tant la concordance avec d'autres statistiques que la réduction de la variance et/ou du biais de non-réponse, l'expression « système de poids équilibrés » est une description plus appropriée que « système de poids convergents », parce que l'objectif est alors d'équilibrer les poids afin de refléter le résultat de l'échantillonnage, la réponse à l'enquête et l'information disponible.

4) *Calage de commodité et de transparence*. Comme le font remarquer Harms et Duchesne (2006), le recours à la méthode de calage s'est répandu dans les applications réelles, parce que les estimations résultantes sont faciles à interpréter et à justifier puisqu'elles s'appuient sur les poids de sondage et des contraintes de calage naturelles. À l'utilisateur moyen, le calage sur des totaux connus paraît transparent et naturel. Les utilisateurs qui comprennent la pondération d'échantillon apprécient le fait que le calage « ne fait que modifier légèrement » les poids de sondage, tout en respectant les valeurs de contrôle. L'absence de biais n'est perturbée que de manière négligeable. Les formes de calage les plus simples ne font appel à aucune hypothèse et ne s'appuient que sur des « contraintes naturelles ». Un autre avantage qu'apprécient les utilisateurs est que, dans de nombreuses applications, le calage produit un système de poids unique, applicable à toutes les variables étudiées, lesquelles sont habituellement nombreuses dans le cas des grandes enquêtes menées par les organismes statistiques publics.

5) *Le calage combiné à d'autres termes*. Certains auteurs utilisent le mot « calage » en combinaison avec d'autres termes afin de décrire diverses lignes de pensée. Voici des exemples de cette prolifération d'expressions : calage basé sur un modèle (Wu et Sitter 2001), calage *g* (Vanderhoeft, Waeytens et Museux 2000), calage harmonisé (Webber, Latouche et Rancourt 2000), calage de plus haut niveau (Singh, Horn et Yu 1998), calage par régression (Demnati et

Rao 2004), calage non linéaire (Plikusas 2006), calage supergénéralisé (Ardilly 2006), estimateur par calage fondé sur un modèle de réseau neuronal et estimateur par calage fondé sur un modèle polynomial local (Montanari et Ranalli 2003, 2005), estimateur du pseudo-maximum de vraisemblance empirique calé sur un modèle (Wu 2003), et ainsi de suite. En outre, le calage joue un rôle important dans les méthodes de sondage indirectes proposées dans Lavalleé (2006). Dans un esprit légèrement différent, certains auteurs proposent des concepts, que nous n'examinerons pas ici, tels que l'imputation calée (Beaumont 2005a) ou le calage du biais (Chambers, Dorfman et Wehrly 1993, Zheng et Little 2003). Les pages qui suivent ne rendent pas justice à toutes les innovations qui ont lieu dans le domaine du calage, mais l'énumération des noms des méthodes donne à elle seule une idée des diverses voies qui ont été explorées.

(6) *Le calage en tant que nouvelle ligne de pensée*. Si le calage constitue « une nouvelle approche » qui se distingue nettement de celles qui l'ont précédée, nous devons nous poser des questions telles que : le calage généralise-t-il les théories ou les approches antérieures? Le calage fournit-il de meilleures réponses, plus satisfaisantes, aux questions importantes que les approches reconnues antérieurement? Aux sections 4.5 et 7.1, nous illustrons dans quelle mesure les réponses offertes par le calage sont comparables ou différent de celles résultant des courants de pensée antérieurs.

Le praticien de l'échantillonnage se heurte à des « nuisances », telles que la non-réponse, les déficiences des bases de sondage et les erreurs de mesure. Certes, l'imputation et la repondération pour corriger la non-réponse sont d'usage répandu et se font selon une foule de méthodes. Mais il s'agit dans une certaine mesure de « problèmes distincts », qui attendent encore d'être intégrés plus pleinement dans une théorie générale, plus satisfaisante, de l'inférence dans le contexte des sondages. Nombre d'articles théoriques traitent de l'estimation dans le cas d'un sondage idéal imaginaire, inexistant dans la pratique, qui ne souffre pas de la non-réponse et d'autres erreurs non dues à l'échantillonnage. Le propos n'est pas de critiquer toutes ces études théoriques excellentes, mais idéalisantes. Mais il faut aussi explorer les fondements.

Les sections 9 et 10 indiquent que le calage peut offrir une perspective plus systématique de l'inférence dans les sondages, même en présence de diverses erreurs non dues à l'échantillonnage. De fructueux développement sont à prévoir à cet égard.

2. Conditions de base pour l'estimation fondée sur le plan dans les sondages

La présente section a pour but d'établir le contexte des sections 3 à 7. Par « conditions de base », nous entendons ici un échantillonnage probabiliste d'éléments à une seule phase et une réponse complète. Les conditions réelles de sondage ne sont pas aussi simples et parfaites, mais de nombreux articles théoriques traitent néanmoins de cette situation.

Soit un échantillon probabiliste s tiré de la population finie $U = \{1, 2, ..., k, ..., N\}$. Le plan d'échantillonnage probabiliste produit, pour l'élément k, une probabilité d'inclusion connue, $\pi_k > 0$, et un poids de sondage correspondant $d_k = 1/\pi_k$. La valeur y_k de la variable étudiée y est enregistrée pour tous les $k \in s$ (réponse complète). L'objectif est d'estimer un total de population $Y = \sum_U y_k$ en se servant de l'information auxiliaire. La variable étudiée y peut être continue ou, comme dans le cas de nombreuses enquêtes menées par des organismes publics, catégorique. Par exemple, si y est dichotomique et prend la valeur $y_k = 0$ ou $y_k = 1$ selon que la personne k est occupée ou chômeuse, le paramètre $Y = \sum_U y_k$ à estimer est le chiffre de population des chômeurs. (Si $A \subseteq U$ est un ensemble d'éléments, nous écrivons $\sum_A$ pour $\sum_{k \in A}$.) L'estimateur de base sans biais sous le plan de Y est l'estimateur d'Horwitz-Thompson $\hat{Y}_{\mathrm{HT}} = \sum_s d_k y_k$. Cependant, il est inefficace si des données auxiliaires puissantes sont disponibles à l'étape de l'estimation.

La notation générale du vecteur auxiliaire sera $\mathbf{x}_k$. Dans certains pays, pour certaines enquêtes, les sources de données auxiliaires permettent de construire des vecteurs $\mathbf{x}_k$ de grande portée. Voici néanmoins quelques exemples de vecteurs simples : 1) $\mathbf{x}_k = (1, x_k)'$, où x_k est la valeur pour l'élément k d'une variable auxiliaire continue x; 2) le vecteur de classification utilisé pour coder l'appartenance à l'un de P groupes mutuellement exclusifs et exhaustifs, $\mathbf{x}_k = \boldsymbol{\gamma}_k = (\gamma_{1k}, ..., \gamma_{pk}, ..., \gamma_{Pk})'$, de sorte que, pour $p = 1, 2, ..., P$, $\gamma_{pk} = 1$ si k appartient au groupe p, et $\gamma_{pk} = 0$ autrement; 3) la combinaison de 1) et 2), $\mathbf{x}_k = (\boldsymbol{\gamma}_k', x_k \boldsymbol{\gamma}_k')'$; 4) le vecteur $\mathbf{x}_k$ qui codifie deux classifications déployées « côte à côte », la dimension de $\mathbf{x}_k$ étant $P + Q$ - 1, où P et Q sont les nombres respectifs de catégories, et où le « -1 » évite d'obtenir une matrice singulière dans le calcul des poids calés « sur les marges »; 5) l'extension de 4) à plus de deux classifications déployées « côte à côte ». Les cas 4 et 5 sont particulièrement importants en production dans les organismes statistiques nationaux.

Dans l'approche du calage, il importe au plus haut point de spécifier exactement l'*information auxiliaire*. Dans les

conditions de base, nous devons distinguer deux cas en ce qui concerne $\mathbf{x}_k$:

i) $\mathbf{x}_k$ est une valeur du vecteur auxiliaire connue pour chaque $k \in U$ (information auxiliaire complète);

ii) $\sum_U \mathbf{x}_k$ est un total connu (importé) et $\mathbf{x}_k$ est connu (observé) pour chaque $k \in s$.

C'est souvent le contexte de l'enquête qui dicte si i) ou ii) s'applique. Le cas i), celui de l'information auxiliaire complète, se présente quand $\mathbf{x}_k$ est spécifié dans la base de sondage pour chaque $k \in U$ (et est donc connu pour chaque $k \in s$). Ce contexte est typique des enquêtes auprès des particuliers et des ménages menées en Scandinavie et dans d'autres pays nord-européens dotés de registres administratifs de haute qualité qui, appariés à la base de sondage, donnent un grand nombre de variables auxiliaires possibles. Le total de population $\sum_U \mathbf{x}_k$ s'obtient simplement par sommation des $\mathbf{x}_k$.

Le cas i) offre une grande latitude pour la structuration du vecteur auxiliaire $\mathbf{x}_k$. Ainsi, si x_k est une valeur de variable continue spécifiée pour chaque $k \in U$, nous sommes appelés à considérer l'inclusion de x_k^2 et d'autres fonctions de x_k dans $\mathbf{x}_k$, parce que des totaux tels que $\sum_U x_k^2$ et $\sum_U \log x_k$ se calculent facilement. Si la relation avec la variable étudiée y est curviligne, ce serait une grave omission de ne pas tenir compte de totaux connus tels que la forme quadratique ou la forme logarithmique.

Le cas ii) est celui qui s'applique dans les enquêtes où la condition i) n'est pas satisfaite, mais où $\sum_U \mathbf{x}_k$ est importé d'une source externe jugée suffisamment fiable et où la valeur individuelle $\mathbf{x}_k$ est disponible (observée pendant la collecte des données) pour chaque $k \in s$. Alors, $\sum_U \mathbf{x}_k$ est parfois appelé « total de contrôle indépendant », pour préciser que sa source est externe à l'enquête. Il est moins souple que le cas i) : si x_k est une variable pour laquelle un total $\sum_U x_k$ est importé d'une source fiable, $\sum_U x_k^2$ peut ne pas être disponible, ce qui empêchera d'inclure x_k^2 dans $\mathbf{x}_k$.

3. Estimation par la régression généralisée dans les conditions de base

3.1 Concept d'estimation par la regression généralisée

Avant de parler du calage, nous considérerons l'*estimation par la régression généralisée* (GREG) (ou simplement l'*estimation par la régression*) pour deux bonnes raisons : 1) l'estimation GREG peut aussi être considérée comme un moyen systématique de tenir compte de l'information auxiliaire et 2) certains estimateurs GREG (mais pas tous) sont des estimateurs par calage, en ce sens qu'ils peuvent être exprimés en fonction d'une pondération linéaire calée.

Les estimateurs GREG et les estimateurs par calage ont été étudiés abondamment au cours des deux dernières décennies. Rien que la terminologie, « estimation GREG » et « estimation par calage », reflète des processus de réflexion différents. Les statisticiens spécialisés dans le domaine appartiennent à deux écoles de pensée, celle de la « régression généralisée GREG » et celle du « calage ». La distinction n'est peut-être pas aussi nette, mais nous l'utiliserons ici car elle aide à structurer l'exposé. Nous n'irons pas jusqu'à soutenir que la seconde école compte plus d'adeptes parmi les organismes statistiques nationaux et la première, dans les milieux universitaires, mais il se pourrait fort bien qu'une telle tendance existe.

Le concept de l'estimateur GREG a évolué progressivement depuis le milieu des années 1970. Särndal, Swensson et Wretman (1992) expliquent l'estimation GREG simple (linéaire), tandis que Fuller (2002) présente une revue détaillée de l'estimation par la régression. L'idée centrale est que les valeurs prévues de y, $\hat{y}_k$, peuvent être produites pour chacun des N éléments de la population par ajustement d'un *modèle auxiliaire* et utilisation des valeurs du vecteur auxiliaire $\mathbf{x}_k$ connues pour tous les $k \in U$. Ces valeurs prévues servent à élaborer un estimateur presque sans biais sous le plan du total de population $Y = \sum_U y_k$ de la forme

$$\begin{aligned} \hat{Y}_{\text{GREG}} &= \sum_U \hat{y}_k + \sum_s d_k (y_k - \hat{y}_k) \\ &= \sum_s d_k y_k + \left(\sum_U \hat{y}_k - \sum_s d_k \hat{y}_k \right). \end{aligned} \qquad (3.1)$$

Le motif évident de cette construction est la perspective d'une estimation très précise $\hat{Y}_{\text{GREG}}$ grâce à un ajustement étroit du modèle auxiliaire qui aboutit à de petits résidus $y_k - \hat{y}_k$. Cette modélisation est la pierre angulaire de l'approche GREG. Certains auteurs donnent à la forme (3.1) le nom (également justifiable) d'*estimateur par différence généralisé*.

La grande variété de modèles auxiliaires possibles engendre une vaste famille d'estimateurs GREG de la forme (3.1). Le modèle auxiliaire, qui traduit la relation imaginée entre $\mathbf{x}$ et y, peut prendre de nombreuses formes : linéaire, non linéaire, linéaire généralisée, mixte (modèle contenant des termes d'effets fixes et d'effets aléatoires) et ainsi de suite. Quel que soit le choix, le modèle « ne fait qu'assister »; bien qu'il ne puisse être qualifié absolument de « vrai », l'estimateur (3.1) est presque sans biais sous le plan dans le cas de contraintes faibles sur le modèle auxiliaire et sur le plan d'échantillonnage, de sorte que $(\hat{Y}_{\text{GREG}} - Y)/N = O_p(n^{-1/2})$ et $(\hat{Y}_{\text{GREG}} - Y)/N = (\hat{Y}_{\text{GREG,lin}} - Y)/N + O_p(n^{-1})$, où la statistique $\hat{Y}_{\text{GREG,lin}}$, qui est le résultat de la linéarisation de $\hat{Y}_{\text{GREG}}$, est sans biais pour Y.

3.2 Estimateur GREG linéaire

Par estimateur GREG linéaire, nous entendons un estimateur produit par un modèle auxiliaire linéaire à effets fixes. Les prédictions sont $\hat{y}_k = \mathbf{x}_k' \mathbf{B}_{s;dq}$ avec

$$\mathbf{B}_{s;dq} = \left(\sum_s d_k q_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\sum_s d_k q_k \mathbf{x}_k y_k\right)$$

de sorte que (3.1) devient

$$\hat{Y}_{\text{GREG}} = \left(\sum_U \mathbf{x}_k\right)' \mathbf{B}_{s;dq} + \sum_s d_k (y_k - \mathbf{x}_k' \mathbf{B}_{s;dq}). \qquad (3.2)$$

Les q_k sont des facteurs d'échelle, choisis par le statisticien. Le choix type est $q_k = 1$ pour tout k. Le choix de q_k a une certaine incidence (mais souvent limitée) sur l'exactitude de $\hat{Y}_{\text{GREG}}$; la quasi-absence de biais tient pour toute spécification de q_k. (sauf choix outranciers). Bien que le modèle soit simple, l'expression GREG linéaire (3.2) contient de nombreux estimateurs, étant donné les nombreuses options possibles pour le vecteur auxiliaire $\mathbf{x}_k$ et les facteurs d'échelle q_k. Dans des conditions générales,

$$(\hat{Y}_{\text{GREG}} - Y)/N = \left(\sum_s d_k E_k - \sum_U E_k\right)\Big/ N + O_p(n^{-1})$$

où $\sum_s d_k E_k$ est l'estimateur de Horvitz-Thompson dans les résidus $E_k = y_k - \mathbf{x}_k' \mathbf{B}_{U;q}$ avec $\mathbf{B}_{U;q} = (\sum_U q_k \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_U q_k \mathbf{x}_k y_k)$. D'où, les propriétés fondées sur le plan $E(\hat{Y}_{\text{GREG}}) \approx Y$ et $\text{Var}(\hat{Y}_{\text{GREG}}) \approx \text{Var}(\sum_s d_k E_k)$. Une régression *linéaire* étroitement ajustée de y sur $\mathbf{x}$ est la clé d'une variance faible pour $\hat{Y}_{\text{GREG}}$ (ce qui est loin d'affirmer qu'« une régression linéaire est la régression vraie »).

L'estimateur GREG linéaire de Särndal, Swensson et Wretman (1992) était motivé par le modèle auxiliaire linéaire ξ énonçant que $E_\xi(y_k) = \boldsymbol{\beta}' \mathbf{x}_k$ et $V_\xi(y_k) = \sigma_k^2$. L'ajustement par les moindres carrés généralisés donne l'estimateur (3.2) avec $q_k = 1/\sigma_k^2$. Dans ce contexte, une estimation éclairée quant à la variation des résidus $y_k - \boldsymbol{\beta}' \mathbf{x}_k$ détermine les q_k. Si le vecteur $\mathbf{x}_k$ est fixe, l'effort de modélisation se résume à une opinion quant à la forme des résidus. Le choix $\sigma_k^2 = \sigma^2 x_k$ donne l'estimateur par le ratio classique. Si $q_k = \boldsymbol{\mu}' \mathbf{x}_k$ pour tout $k \in U$ et un vecteur constant $\boldsymbol{\mu}$, alors (3.2) se réduit à la « forme esthétique » $(\sum_U \mathbf{x}_k)' \mathbf{B}_{s;dq}$.

Comme l'ont fait remarquer Beaumont et Alavi (2004) et d'autres, l'estimateur GREG linéaire est robuste au biais (presque sans biais, bien que le modèle auxiliaire ne soit pas tout à fait « correct »), mais peut être considérablement moins efficace (avoir une plus grande erreur quadratique moyenne) que les estimateurs de rechange sensibles au modèle qui, quoique biaisés, ont parfois une variance beaucoup plus faible. Donc, on pourrait affirmer que l'estimateur GREG linéaire n'est pas robuste à la variance; néanmoins, il s'agit d'un concept fondamental dans la théorie de l'estimation fondée sur le plan de sondage.

La spécification de $\mathbf{x}_k$ devrait inclure les variables (dont les totaux de population sont connus) qui ont déjà servi à définir le plan de sondage. Les données issues de l'élaboration du plan de sondage ne devraient pas être abandonnées à l'étape de l'estimation; au contraire, une « utilisation répétée » est recommandée. Par exemple, dans le cas de l'échantillonnage aléatoire simple stratifié (EASS), le vecteur $\mathbf{x}_k$ dans l'estimateur (3.2) devrait inclure, en même temps que les autres variables disponibles, la variable muette servant d'identificateur de strate, $\boldsymbol{\gamma}_k = (\gamma_{k1}, \gamma_{k2}, ..., \gamma_{kh}, ..., \gamma_{kH})'$, où $\gamma_{kh} = 1$ si l'élément k appartient à la strate h et $\gamma_{kh} = 0$ sinon; $h = 1, ..., H$.

Nous pouvons écrire l'estimateur GREG linéaire (3.2) sous la forme d'une somme pondérée par les poids de sondage, $\hat{Y}_{\text{GREG}} = \sum_s w_k y_k$, avec

$$w_k = d_k g_k;\ g_k = 1 + q_k \boldsymbol{\lambda}' \mathbf{x}_k;$$
$$\boldsymbol{\lambda}' = \left(\sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k\right)' \left(\sum_s d_k q_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1}. \qquad (3.3)$$

Les poids w_k sont *calés* (convergents) sur le total $\mathbf{x}$ de population connu : $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. Que $\hat{Y}_{\text{GREG}}$ puisse être exprimé sous la forme d'une somme linéairement pondérée à l'aide de poids calés est un sous produit fortuit. Cette propriété ne fait pas partie du raisonnement GREG, dont l'idée centrale, formulée dans (3.1), est l'ajustement d'un modèle auxiliaire. Outre l'estimateur linéaire simple, quelques autres estimateurs GREG ont la propriété de calage, comme nous le mentionnerons plus loin.

3.3 Estimateur GREG non linéaire

Deux caractéristiques de l'estimateur GREG linéaire (3.2) en font une option de choix pour la production courante des organismes statistiques : i) le total de population auxiliaire $\sum_U \mathbf{x}_k$ est exclu, de sorte que l'estimation peut avoir lieu à condition qu'une valeur exacte de ce total puisse être calculée ou importée, et ii) lorsqu'il est écrit sous forme de la somme pondérée linéairement $\hat{Y}_{\text{GREG}} = \sum_s w_k y_k$, le système de poids (3.3) est indépendant de la variable y et peut donc être appliquée à toutes les variables y de l'enquête. Il n'est pas nécessaire de connaître $\mathbf{x}_k$ individuellement pour tout $k \in U$; connaître $\sum_U \mathbf{x}_k$ suffit. Naturellement, si nous connaissons toutes les valeurs de $\mathbf{x}_k$, nous pouvons rechercher des membres plus efficaces (mais toujours presque sans biais sous le plan) de la famille d'estimateurs GREG (3.1), ce qui permettra également d'écarter une autre critique de l'estimateur GREG linéaire, à savoir qu'un modèle linéaire n'est pas raisonnable pour certains types de données. Par exemple, pour une variable y dichotomique, un modèle auxiliaire

logistique pourrait à la fois être plus réaliste et aboutir à un estimateur GREG plus précis.

Par estimateur GREG non linéaire, il est entendu ici qu'un estimateur est produit comme en (3.1) avec l'aide d'un modèle d'un autre type que « linéaire en $\mathbf{x}_k$ avec effets fixes ». Firth et Bennett (1998), ainsi que Lehtonen et Veijanen (1998) ont été parmi les premiers à étendre le concept GREG dans cette direction; à cet égard, voir aussi Chambers, Dorfman et Wehrly (1993). Ces dernières années, plusieurs auteurs ont étudié les estimateurs GREG non linéaires assistés par modèle.

La notion d'estimation GREG non linéaire est souple; une gamme d'estimateurs deviennent possibles par la voie de modèles auxiliaires ξ du type suivant :

$$E_\xi(y_k|\mathbf{x}_k) = \mu_k \quad \text{pour} \quad k \in U \qquad (3.4)$$

où la moyenne sous le modèle μ_k et la variance sous le modèle $V_\xi(y_k|\mathbf{x}_k)$ se voient attribuer chacune une formule appropriée.

Le modèle (3.4) s'applique notamment quand $\mu_k = \mu_k(\mathbf{x}_k, \boldsymbol{\theta})$ est une fonction non linéaire en $\mathbf{x}_k$ spécifiée. Si l'on estime $\boldsymbol{\theta}$ par $\hat{\boldsymbol{\theta}}$, les valeurs ajustées nécessaires pour $\hat{Y}_{\text{GREG}}$ dans (3.1) sont $\hat{y}_k = \mu(\mathbf{x}_k, \hat{\boldsymbol{\theta}})$ pour $k \in U$. Par exemple, si le modélisateur spécifie $\log \mu_k = \alpha + \beta x_k$, les prédictions à utiliser dans (3.1) sont, après estimation des paramètres $\hat{y}_k = \exp(\hat{\alpha} + \hat{\beta} x_k)$.

Les autres applications de (3.4) incluent les modèles linéaires généralisés, tels que $g(\mu_k) = \mathbf{x}_k'\boldsymbol{\theta}$ pour une fonction lien spécifiée $g(\cdot)$ et qu'une structure appropriée est donnée à $V_\xi(y_k|\mathbf{x}_k) = v(\mu_k)$. Nous estimons $\boldsymbol{\theta}$ par $\hat{\boldsymbol{\theta}}$; les valeurs ajustées nécessaires pour l'estimateur GREG non linéaire (3.1) sont $\hat{y}_k = \hat{\mu}_k = g^{-1}(\mathbf{x}_k'\hat{\boldsymbol{\theta}})$. Par exemple, si l'on utilise un modèle auxiliaire logistique, $\mathbf{x}_k'\boldsymbol{\theta} = \text{logit}(\mu_k) = \log(\mu_k/(1-\mu_k))$, et $\hat{y}_k = \hat{\mu}_k = \exp(\mathbf{x}_k'\hat{\boldsymbol{\theta}})/(1+\exp(\mathbf{x}_k'\hat{\boldsymbol{\theta}}))$.

Lehtonen et Veijanen (1998) considèrent le cas d'une variable étudiée catégorique comportant I catégories, $i = 1, 2, ..., I$, $y_{ik} = 1$ si l'élément k appartient à la catégorie i, et $y_{ik} = 0$ sinon. Par exemple, dans une enquête sur la population active comportant $I = 3$ catégories, à savoir « occupé », « chômeur » et « inactif », l'objectif est d'estimer les chiffres de population respectifs $Y_i = \sum_U y_{ik}$, $i = 1, 2, 3$. Ces auteurs utilisent le modèle auxiliaire logistique

$$E_\xi(y_{ik}|\mathbf{x}_k) = \mu_{ik};\ \mu_{ik} = \exp(\mathbf{x}_k'\boldsymbol{\theta}_i)\Big/\left(1+\sum_{i=2}^{I}\exp(\mathbf{x}_k'\boldsymbol{\theta}_i)\right). \qquad (3.5)$$

Ils obtiennent les estimations $\hat{\boldsymbol{\theta}}_i$ de $\boldsymbol{\theta}_i$ en maximisant la log-vraisemblance pondérée par les poids de sondage. Les prédictions résultantes $\hat{y}_{ik} = \hat{\mu}_{ik}$ sont utilisées pour former $\hat{Y}_{i\,\text{GREG}} = \sum_U \hat{y}_{ik} + \sum_s d_k(y_{ik} - \hat{y}_{ik})$, pour $i = 1, 2, ..., I$.

Un autre développement est l'application de l'approche GREG à l'estimation pour des domaines, comme dans Lehtonen, Särndal et Veijanen (2003, 2005), ainsi que Myrskylä (2007). Dans les deux premiers de ces articles, les auteurs utilisent des modèles mixtes pour assister l'estimateur GREG non linéaire. Soit U_a un domaine, $U_a \subset U$, dont nous souhaitons estimer le total $Y_{ia} = \sum_{U_a} y_{ik}$ pour $i = 1, 2, ..., I$. Dans l'article de 2005, les prédictions pour l'estimateur GREG non linéaire sont calculées d'après le modèle mixte logistique énonçant que, pour $k \in U_a$,

$$E_\xi(y_{ik}|\mathbf{x}_k; \mathbf{u}_{ia}) = \exp(\mathbf{x}_k'\boldsymbol{\theta}_{ia})\Big/\left(1+\sum_{i=2}^{I}\exp(\mathbf{x}_k'\boldsymbol{\theta}_{ia})\right) \qquad (3.6)$$

avec $\boldsymbol{\theta}_{ia} = \boldsymbol{\beta}_i + \mathbf{u}_{ia}$, où $\mathbf{u}_{ia}$ est un vecteur d'écarts aléatoires particuliers au domaine par rapport au vecteur d'effets fixes $\boldsymbol{\beta}_i$.

Les estimateurs GREG non linéaires assistés par modèle, tels que (3.5) et (3.6), nécessitent l'ajustement individuel du modèle pour chaque variable y, car il n'existe aucun système de pondérations applicable uniformément. Cependant, la question se pose de savoir s'il existe des exemples d'estimateurs GREG non linéaires tels que les avantages pratiques de l'estimateur GREG linéaire soient préservés, autrement dit une forme linéairement pondérée à l'aide de poids calés indépendants de la variable y. La réponse est affirmative. À cet égard, deux orientations intéressantes se dégagent de la littérature récente.

Breidt et Opsomer (2000), et Montanari et Ranalli (2005) considèrent des estimateurs GREG assistés par modèle polynomial local dans le cas d'une variable auxiliaire continue unique dont les valeurs x_k sont connues pour tout $k \in U$. La méthode requiert plusieurs choix, dont 1) l'ordre q de l'expression polynomiale locale, 2) la spécification de la fonction noyau et 3) la largeur de fenêtre. L'estimateur résultant peut être exprimé en fonction des poids calés sur les totaux de population des puissances de x_k, de sorte que $\sum_s w_k x_k^j = \sum_U x_k^j$ pour $j = 0, 1, ..., q$.

Breidt, Claeskens et Opsomer (2005) élaborent un estimateur GREG à fonction spline pénalisée pour une variable x unique; le modèle auxiliaire est $m(x; \boldsymbol{\beta}) = \beta_0 + \beta_1 x + ... + \beta_q x^q + \sum_{j=1}^{K} \beta_{q+j}(x - \kappa_j)_+^q$, où $(t)_+^q = t^q$ si $t > 0$ et 0 autrement, q est le degré de la spline et les κ_k sont des nœuds espacés de manière appropriée, par exemple, les quantiles d'échantillon, uniformément espacés, des valeurs x_k. Après avoir estimé les paramètres β, ils obtiennent les prédictions $\hat{y}_k = m(x_k; \hat{\boldsymbol{\beta}})$ nécessaires pour la formule GREG générale (3.1). Les auteurs soulignent que l'estimateur GREG résultant est calé pour la partie paramétrique du modèle, c'est-à-dire $\sum_s w_k x_k^j = \sum_U x_k^j$ pour $j = 0, 1, ..., q$, ainsi que pour les termes polynomiaux tronqués dans le modèle, à condition qu'ils ne soient pas pénalisés.

Nous pouvons résumer l'estimation GREG comme il suit. L'estimateur GREG linéaire offre des avantages pratiques pour la production de statistiques à grande échelle. Il peut être exprimé sous la forme d'une somme pondérée linéairement de valeurs y_k au moyen de poids calés sur $\sum_U \mathbf{x}_k$; les poids sont indépendants des valeurs y_k et peuvent être appliqués à toutes les variables y de l'enquête. Il suffit de connaître un total de population auxiliaire $\sum_U \mathbf{x}_k$, importé d'une source fiable. L'estimateur GREG non linéaire peut donner lieu à une variance considérablement réduite, grâce aux modèles plus perfectionnés qui peuvent être envisagés lorsqu'on dispose d'information auxiliaire complète ($\mathbf{x}_k$ connu pour tout $k \in U$). La quasi-absence de biais sous le plan est préservée. Certains estimateurs GREG non linéaires peuvent s'écrire sous forme de sommes linéairement pondérées.

Dans les exercices théoriques portant sur des populations et des relations créées artificiellement, il est possible de provoquer des situations où un estimateur GREG non linéaire présente un grand avantage par rapport à l'estimateur GREG linéaire en ce qui concerne la variance. Les expériences de ce type sont importantes pour l'illustration. Cependant, pour répondre aux exigences quotidiennes de production des organismes statistiques nationaux, les estimateurs GREG non linéaires « extravagants » ne semblent présenter qu'un intérêt assez lointain à l'heure actuelle. Les modèles auxiliaires spécifiés pour l'estimation GREG doivent satisfaire aux exigences de robustesse et de faisabilité. L'attrait d'une petite réduction de la variance d'échantillonnage est balayé par les préoccupations au sujet d'autres erreurs (non due à l'échantillonnage) et aux difficultés rencontrées dans le processus de production quotidienne.

Le passage des estimateurs GREG linéaires aux estimateurs GREG non linéaires offre des possibilités et soulève des questions. Quelle est la formule la plus appropriée de l'espérance du modèle μ_k? À quel point les résultats sont-ils sensibles à la spécification de la partie variance du modèle auxiliaire? Dans quelle mesure la rapidité des calculs est-elle un problème? Des travaux de recherche plus approfondis permettront de mieux répondre à ces questions.

4. Méthode de l'estimation par calage

4.1 Calage dans les conditions de base

Dans la méthode d'estimation GREG examinée à la section précédente, une étape essentielle consiste à produire les valeurs prévues $\hat{y}_k$ par ajustement d'un modèle auxiliaire. Par contre, telle qu'elle est définie à la section 1.1, la méthode de calage ne fait directement référence à aucun modèle. Elle met plutôt l'accent sur les données sur lesquelles peut se faire le calage. Un élément clé de l'« approche du calage » est la pondération linéaire des valeurs y observées, avec alignement des poids sur des agrégats calculables. Cette différence d'ordre conceptuel donne parfois des estimateurs différents pour les approches GREG et de calage.

La méthode de calage offre un haut degré de généralisation. Elle peut s'appliquer dans des conditions diverses, dont les plans d'échantillonnage complexes, la correction de la non-réponse et les erreurs dans les bases de sondage. Néanmoins, à la présente section, nous nous attachons aux conditions de base énoncées à la section 2, c'est-à-dire celles d'échantillonnage à une phase et de réponse complète. Nous reprenons également la notation utilisée dans cette section. Les données dont nous disposons pour estimer le total de population $Y = \sum_U y_k$ sont i) les valeurs de la variable étudiée y_k observées pour $k \in s$, ii) les poids de sondage connus $d_k = 1/\pi_k$ pour $k \in U$ et iii) les valeurs du secteur auxiliaire connues $\mathbf{x}_k$ pour $k \in U$ (ou un total importé $\sum_U \mathbf{x}_k$). Ces conditions simples sont celles énoncées dans les articles publiés par Deville et Särndal (1992), ainsi que par Deville, Särndal et Sautory (1993), auxquels la méthode doit son nom et qui ont inspiré les travaux qui ont suivi. Bien que le contexte soit simple, le calage soulève plusieurs questions, dont certaines ayant trait aux calculs, que nous passons en revue à la section 5.

Notre objectif, dans les sections 4.2 et 4.3, est de déterminer les poids w_k qui satisfont l'équation de calage $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$, puis de les utiliser pour obtenir l'estimateur par calage de Y de la forme $\hat{Y}_{\mathrm{CAL}} = \sum_s w_k y_k$, que nous pouvons comparer à l'estimateur de Horvitz-Thompson sans biais en écrivant $\hat{Y}_{\mathrm{CAL}} = \hat{Y}_{\mathrm{HT}} + \sum_s (w_k - d_k) y_k$. Il s'ensuit que le biais de $\hat{Y}_{\mathrm{CAL}}$ est $E(\hat{Y}_{\mathrm{CAL}}) - Y = E(\sum_s (w_k - d_k) y_k)$. Réaliser l'objectif de quasi-absence de biais sous le plan de sondage exige que $E(\sum_s (w_k - d_k) y_k) \approx 0$, quelle que soit la variable y. Naturellement, le calage devrait viser à produire de petits écarts $w_k - d_k$.

Nous pouvons atteindre l'objectif de « calage pour obtenir la convergence sur des totaux de population auxiliaires connus » de nombreuses façons. Nous pouvons créer de nombreux ensembles de poids calés sur le total connu $\sum_U \mathbf{x}_k$. À la présente section, nous examinerons cette prolifération des méthodes sous deux angles relevés dans la littérature, à savoir la *méthode de la distance minimale* et la *méthode du vecteur instrumental*. Demnati et Rao (2004) proposent encore un autre moyen de construire une variété de poids calés.

4.2 Méthode de la distance minimale

Dans cette méthode, le but du calage est de modifier les poids initiaux $d_k = 1/\pi_k$ afin d'obtenir de nouveaux poids

w_k, jugés comme étant « proches » des poids d_k. Pour cela, considérons la fonction de distance $G_k(w, d)$ définie pour chaque $w > 0$ de façon que $G_k(w, d) \geq 0$, $G_k(d, d) = 0$, dérivable par rapport à w, strictement convexe, à dérivée continue $g_k(w, d) = \partial G_k(w, d)/\partial w$ telle que $g_k(d, d) = 0$. Habituellement, on choisit la fonction de distance de manière que $g_k(w, d) = g(w/d)/q_k$, où les q_k sont des facteurs d'échelle positifs convenablement choisis, $g(\cdot)$ est une fonction d'un argument unique, continue, strictement croissante, avec $g(1) = 0, g'(1) = 0$. Soit $F(u) = g^{-1}(u)$ la fonction inverse de $g(\cdot)$. Minimiser la distance totale $\sum_s G_k(w_k, d_k)$ sous la contrainte de l'équation de calage $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$ mène à $w_k = d_k F(q_k \mathbf{x}_k' \boldsymbol{\lambda})$, où $\boldsymbol{\lambda}$ est la solution (à supposer qu'elle existe) de

$$\sum_s d_k \mathbf{x}_k F(q_k \mathbf{x}_k' \boldsymbol{\lambda}) = \sum_U \mathbf{x}_k. \qquad (4.1)$$

Les poids ont une propriété d'optimalité, parce qu'une fonction objective dûment spécifiée est minimisée, mais il s'agit d'une « optimalité faible » en ce sens que les spécifications possibles de la fonction de distance et des facteurs d'échelle q_k sont nombreuses.

La fonction de distance $G_k(w_k, d_k) = (w_k - d_k)^2/2d_k q_k$ a suscité beaucoup d'intérêt. Elle donne $g_k(w_k, d_k) = (w_k/d_k - 1)/q_k$; $g(w/d) = w/d - 1$; $F(u) = g^{-1}(u) = 1 + u$. L'expression « cas linéaire » est donc appropriée. La tâche consiste alors à minimiser la « distance du chi-carré » $\sum_s (w_k - d_k)^2/2d_k q_k$, sachant $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. L'équation (4.1) se lit $\sum_s d_k \mathbf{x}_k (1 + q_k \mathbf{x}_k' \boldsymbol{\lambda}) = \sum_U \mathbf{x}_k$, qui se résout facilement pour obtenir $\boldsymbol{\lambda}$. L'estimateur résultant de $Y = \sum_U y_k$ est $\hat{Y}_{\text{CAL}} = \sum_s w_k y_k$ avec les poids $w_k = d_k g_k$ donnés par (3.3). Autrement dit, $\hat{Y}_{\text{CAL}} = \hat{Y}_{\text{GREG}}$ tel qu'il est donné par (3.2) et les résidus qui déterminent la variance asymptotique sont $E_k = y_k - \mathbf{x}_k' \mathbf{B}_{U;q}$ conformément à la section 3.2. Il arrive d'obtenir certains poids négatifs w_k.

L'estimateur GREG linéaire implique l'utilisation de poids calés (sur $\sum_U \mathbf{x}_k$) et le revers de cette médaille est que le cas linéaire du calage (avec la distance du chi-carré) donne l'estimateur GREG linéaire. La tendance, dans certaines publications et applications, à entremêler l'approche GREG et l'approche du calage émane de ce fait. Bon nombre d'applications fructueuses de l'utilisation de l'information auxiliaire découlent, en tout cas, de cette linéarité bilatérale. L'Enquête sur la population active menée au Canada en est un exemple et un fait récent intéressant en ce qui la concerne est le recours à des estimateurs composites où une partie de l'information provient des résultats d'enquête des mois antérieurs, comme le décrivent Fuller et Rao (2001).

L'équation de calage est satisfaite pour tout choix des facteurs d'échelle positifs q_k dans (3.2). Un choix simple est $q_k = 1$ pour tout k, mais ce n'est pas toujours celui qui est privilégié. Par exemple, s'il existe une variable auxiliaire unique, systématiquement positive et que $\mathbf{x}_k = x_k$, nombreux sont ceux qui s'attendront intuitivement à ce que $\hat{Y}_{\text{CAL}} = \sum_s w_k y_k$ mène à l'estimateur par le ratio habituel $\sum_U x_k (\sum_s d_k y_k)/(\sum_s d_k x_k)$, ce qu'il fait, mais si l'on prend $q_k = x_k^{-1}$ et non $q_k = 1$.

Une autre fonction de distance d'un intérêt considérable est $G_k(w_k, d_k) = \{w_k \log(w_k/d_k) - w_k + d_k\}/q_k$. Elle mène à $F(u) = g^{-1}(u) = \exp(u)$, qui est le « cas exponentiel ». Alors, (4.1) se lit $\sum_s d_k \mathbf{x}_k \exp(q_k \mathbf{x}_k' \boldsymbol{\lambda}) = \sum_U \mathbf{x}_k$. Des méthodes de résolution numérique doivent être appliquées pour trouver $\boldsymbol{\lambda}$, afin d'obtenir les poids $w_k = d_k \exp(q_k \mathbf{x}_k' \boldsymbol{\lambda})$. Aucun w_k négatif ne sera obtenu.

Deville et Särndal (1992) montrent qu'une gamme de fonctions de distance satisfaisant des conditions faibles produisent des estimateurs par calage asymptotiquement équivalents. Diverses fonctions de distance sont comparées dans Deville, Särndal et Sautory (1993), Singh et Mohl (1996), ainsi que Stukel, Hidiroglou et Särndal (1996). Certaines de ces fonctions garantissent que les poids se situent entre des bornes spécifiées, de façon à exclure toute valeur trop grande ou trop faible (négative). Les changements apportés à la fonction de distance n'ont souvent qu'un effet mineur sur la variance de l'estimateur par calage $\hat{Y}_{\text{CAL}} = \sum_s w_k y_k$, même si la taille d'échantillon est assez petite. Les questions relatives à l'existence d'une solution à l'équation de calage sont abordées dans Théberge (2000).

4.3 Méthode du vecteur instrumental

La méthode du vecteur instrumental est une alternative à la minimisation de la distance. Elle est considérée dans Deville (1998), Estevao et Särndal (2000, 2006) et Kott (2006). Elle permet également de produire de nombreux ensembles de poids tous calés sur la même information.

Considérons des poids de la forme $w_k = d_k F(\boldsymbol{\lambda}' \mathbf{z}_k)$, où $\mathbf{z}_k$ est un vecteur dont les valeurs sont définies pour $k \in s$ et ayant la même dimension que le vecteur auxiliaire spécifié $\mathbf{x}_k$, et où le vecteur $\boldsymbol{\lambda}$ est déterminé d'après l'équation de calage $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. La fonction $F(\cdot)$ joue le même rôle que pour la méthode de minimisation de la distance; plusieurs choix de $F(\cdot)$ sont intéressants, par exemple, $F(u) = 1 + u$ et $F(u) = \exp(u)$.

Si nous optons pour la fonction linéaire $F(u) = 1 + u$, nous obtenons $w_k = d_k(1 + \boldsymbol{\lambda}' \mathbf{z}_k)$. Il est facile de déterminer $\boldsymbol{\lambda}$ de façon à satisfaire l'équation de calage $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. L'estimateur par calage résultant est

$$\hat{Y}_{\text{CAL}} = \sum_s w_k y_k;\ w_k = \sum_s d_k (1 + \boldsymbol{\lambda}' \mathbf{z}_k) y_k;$$
$$\boldsymbol{\lambda}' = \left(\sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k\right)' \left(\sum_s d_k \mathbf{z}_k \mathbf{x}_k'\right)^{-1}. \qquad (4.2)$$

Quel que soit le choix de $\mathbf{z}_k$, les poids $w_k = d_k(1+\boldsymbol{\lambda}'\mathbf{z}_k)$ satisfont l'équation de calage. Le choix typique est $\mathbf{z}_k = \mathbf{x}_k$. En particulier, fixer $\mathbf{z}_k = q_k\mathbf{x}_k$, pour les valeurs q_k spécifiées, donne les poids (3.3).

Même les « choix délibérément maladroits » pour $\mathbf{z}_k$ donnent des résultats étonnamment bons. Par exemple, posons que x_k est une variable auxiliaire continue unique et que $\mathbf{z}_k = c_k x_k^{p-1}$. Supposons que $p=3$ et $c_k = 1$ pour quatre éléments seulement, choisis au hasard parmi les $n=100$ éléments d'un échantillon réalisé s et que $c_k = 0$ pour les 96 autres. La quasi-absence de biais de $\hat{Y}_{\mathrm{CAL}} = \sum_s d_k(1+\boldsymbol{\lambda}'\mathbf{z}_k)y_k$ existe encore. Même dans le cas d'un vecteur $\mathbf{z}$ aussi parcimonieux, l'accroissement de la variance comparativement à de meilleurs choix de $\mathbf{z}_k$ n'est pas nécessairement excessif.

Si le plan d'échantillonnage et le vecteur $\mathbf{x}$ sont tous deux fixes, Estevao et Särndal (2004) et Kott (2004) font remarquer qu'il existe un vecteur $\mathbf{z}$ asymptotiquement optimal donné par

$$\mathbf{z}_k = \mathbf{z}_{0k} = d_k^{-1}\sum_{\ell\in s}(d_k d_\ell - d_{k\ell})\mathbf{x}_\ell$$

où $d_{k\ell}$ est l'inverse de la probabilité d'inclusion de deuxième ordre $\pi_{k\ell} = P(k\ \&\ \ell \in s)$, supposée strictement positive. L'estimateur par calage résultant $\hat{Y}_{\mathrm{CAL}} = \sum_s d_k(1+\boldsymbol{\lambda}'\mathbf{z}_{0k})y_k$ est essentiellement l'« estimateur optimal sous "randomisation" » proposé au départ par Montanari (1987) et discuté depuis par de nombreux auteurs.

Andersson et Thorburn (2005) considèrent la question sous l'angle opposé et se demandent si, dans la méthode de la distance minimale, il est possible de spécifier une fonction de distance telle que sa minimisation produise l'estimateur optimal sous « randomisation ». Ils trouvent cette distance qui - ce qui n'est pas entièrement surprenant - est reliée à la distance du chi-carré (mais n'y est pas identique).

4.4 Le calage nécessite-t-il l'énoncé explicite d'un modèle

La méthode de calage présentée aux sections 4.2 et 4.3 consiste simplement à calculer les poids qui reproduisent les totaux auxiliaires spécifiés. Elle ne requiert aucun modèle auxiliaire explicite, à moins que l'on veuille à tout prix que le choix de certaines variables à inclure dans le vecteur $\mathbf{x}_k$ représente un sérieux effort de modélisation. La justification des poids repose plutôt en majeure partie sur leur convergence avec les totaux de contrôle précisés. Les premiers travaux reflètent cette attitude, d'abord ceux de Deming (1943), puis ceux d'Alexander (1987), Zieschang (1990) et d'autres. D'où la question : N'est-il pas néanmoins important de justifier ce « calage sans modèle » à l'aide d'une formulation explicite de modèle? Il est vrai que les statisticiens ont l'habitude de réfléchir en fonction de modèles et se sentent plus ou moins obligés de toujours assortir une procédure statistique de la formulation d'un modèle. Énoncer la relation connexe entre y et $\mathbf{x}$, même si elle est aussi simple que le modèle linéaire courant, pourrait effectivement avoir une certaine valeur pédagogique pour l'explication du calage.

Mais l'énoncé d'un modèle aidera-t-il les utilisateurs et les praticiens à mieux comprendre la méthode de calage? La plupart d'entre eux la considère parfaitement claire et transparente de toute façon. Ils n'ont besoin d'aucune autre justification que celle de la convergence avec les valeurs de contrôle précisées. La recherche du « vrai modèle ayant la vraie structure de variance » se traduira-t-elle par une précision sensiblement meilleure pour la majeure partie des nombreuses estimations produites dans le cadre d'une grande enquête menée par un organisme public? Peu probablement.

La section suivante traite du calage fondé sur un modèle. Dans cette variante, proposée par Wu et Sitter (2001), la modélisation joue effectivement un rôle explicite et important. Ces auteurs dénomment l'estimateur par calage linéaire $\hat{Y}_{\mathrm{CAL}} = \sum_s w_k y_k$ où les poids w_k sont donnés par (3.3) « une application de routine sans modélisation ». Cette description est pertinente, puisque la seule exigence est d'identifier les variables x à leurs totaux de population connus.

4.5 Calage fondé sur un modèle

L'idée du calage fondé sur un modèle est avancée dans Wu et Sitter (2001) et examinée plus en profondeur dans Wu (2003) et dans Montanari et Ranalli (2003, 2005). Le facteur qui motive cette approche est que l'existence d'information auxiliaire complète permet d'utiliser plus efficacement les valeurs connues de $\mathbf{x}_k$ pour chaque $k \in U$ qui ne l'est possible dans le calage sans modèle, où un total connu $\sum_U \mathbf{x}_k$ suffit. Les poids sont contraints de converger vers le total de population calculable des prédictions $\hat{y}_k$ calculées d'après un modèle formulé convenablement. Donc, le système de poids n'est pas nécessairement convergent avec le total de population connu de chaque variable auxiliaire, à moins qu'une mesure particulière soit prise afin de retenir cette propriété. Le calage fondé sur un modèle satisfait encore les trois éléments a) à c) de la définition du calage proposée à la section 3.1; en particulier, les estimateurs sont presque sans biais sous le plan de sondage.

Considérons un modèle auxiliaire non linéaire du type (3.4). Nous estimons le paramètre inconnu $\boldsymbol{\theta}$ par $\hat{\boldsymbol{\theta}}$, ce qui donne les valeurs ajustées $\hat{y}_k = \hat{\mu}_k = \mu(\mathbf{x}_k, \hat{\boldsymbol{\theta}})$ calculées à l'aide des $\mathbf{x}_k$ connus pour tout $k \in U$. Il s'ensuit que la taille de population N est connue et devrait

jouer un rôle important dans le calage. Si nous utilisons la distance du chi-carré minimale, nous trouvons les poids de l'estimateur par calage sous un modèle $\hat{Y}_{\text{MCAL}} = \sum_s w_k y_k$ en minimisant $\sum_s (w_k - d_k)^2/(2d_k q_k)$ pour la valeur spécifiée de q_k et $d_k = 1/\pi_k$, sous la contrainte des équations de calage

$$\sum_s w_k = N; \sum_s w_k \hat{y}_k = \sum_U \hat{y}_k. \qquad (4.3)$$

Pour simplifier, prenons $q_k = 1$ pour tout k; nous calculons les poids calés, réarrangeons les termes et trouvons que l'estimateur par calage sous un modèle peut s'écrire sous la forme

$$\hat{Y}_{\text{MCAL}} = N\{\bar{y}_{s;d} + (\bar{\hat{y}}_U - \bar{\hat{y}}_{s;d})\tilde{B}_{s;d}\} \qquad (4.4)$$

où $\bar{y}_{s;d} = \sum_s d_k y_k / \sum_s d_k$; $\bar{\hat{y}}_{s;d} = \sum_s d_k \hat{y}_k / \sum_s d_k$, et

$$\tilde{B}_{s;d} = \left(\sum_s d_k (\hat{y}_k - \bar{\hat{y}}_{s;d}) y_k\right) \Big/ \sum_s d_k (\hat{y}_k - \bar{\hat{y}}_{s;d})^2.$$

La régression impliquée par $\tilde{B}_{s;d}$ est celle des valeurs observées de y sur les valeurs prévues de y. L'idée de cette régression viendrait rarement à l'esprit du modélisateur qui essaye de structurer la relation entre y_k et $\mathbf{x}_k$, mais elle s'avère efficace dans l'élaboration de l'estimateur par calage. Wu et Sitter (2001) démontrent que

$$(\hat{Y}_{\text{MCAL}} - Y)/N = \left(\sum_s d_k \tilde{E}_k - \sum_U \tilde{E}_k\right) \Big/ N + O_p(n^{-1})$$

avec $\tilde{E}_k = y_k - \bar{y}_U - (\mu_k - \bar{\mu}_U)\tilde{B}_U$, où $\tilde{B}_U = (\sum_U (\mu_k - \bar{\mu}_U) y_k) / \sum_U (\mu_k - \bar{\mu}_U)^2$, et $\bar{\mu}_U = \sum_U \mu_k / N$. Le coefficient $\tilde{B}_U$ peut ne pas être proche de l'unité, même dans le cas de grands échantillons. Il exprime la régression de y_k sur sa moyenne sous le modèle auxiliaire $\mu_k = \mu(\mathbf{x}_k, \boldsymbol{\beta})$. Autrement dit, $\hat{Y}_{\text{MCAL}}$ peut être considéré comme un estimateur par la régression qui utilise l'espérance du modèle μ_k comme variable auxiliaire, ce qui laisse à $\tilde{E}_k$ le rôle de résidu déterminant la variance asymptotique de $\hat{Y}_{\text{MCAL}}$.

Comment cette variance asymptotique se compare-t-elle à la formulation GREG non linéaire (3.1) sous le même modèle auxiliaire non linéaire et les mêmes $\hat{y}_k = \hat{\mu}_k$? La formule (3.1) implique l'existence d'une pente égale à l'unité dans la régression de y_k sur $\hat{y}_k = \hat{\mu}_k$. Vu sous cet angle, $\hat{Y}_{\text{GREG}}$ est un estimateur par différence plutôt qu'un estimateur par régression et est donc moins sensible aux structures présentes dans les données. L'estimateur GREG non linéaire $\hat{Y}_{\text{GREG}}$ est généralement moins efficace que $\hat{Y}_{\text{MCAL}}$. (Il est évidemment possible de modifier $\hat{Y}_{\text{GREG}}$ afin de tenir compte également de l'information contenue dans la taille connue de population N.)

Par ailleurs, comparativement à l'estimateur par calage linéaire (sans modèle) $\hat{Y}_{\text{CAL}} = \sum_s w_k y_k$ avec les poids tels qu'en (3.3), l'estimateur par calage fondé sur un modèle $\hat{Y}_{\text{MCAL}}$ donné par (4.4) peut avoir un avantage considérable pour ce qui est de la variance, mais entraîne la perte des avantages pratiques que sont la convergence vers le total connu de population $\sum_U \mathbf{x}_k$ et un système de poids polyvalent applicable à toutes les variables y. Dans (4.4) les valeurs de y sont pondérées linéairement, mais maintenant, les poids dépendent aussi des valeurs de y. On peut donc se demander si $\hat{Y}_{\text{MCAL}}$ est un estimateur par calage authentique.

Dans une étude empirique, Wu et Sitter (2001) comparent $\hat{Y}_{\text{MCAL}} = \sum_s w_k y_k$, calé conformément à (4.3), à l'estimateur GREG non linéaire, $\hat{Y}_{\text{GREG}} = \sum_U \hat{y}_k + \sum_s d_k (y_k - \hat{y}_k)$ donné par (3.1), pour le même modèle auxiliaire non linéaire et les mêmes $\hat{y}_k = \hat{\mu}_k$. L'étude confirme que $\hat{Y}_{\text{MCAL}}$ a des meilleures propriétés de variance que l'estimateur non linéaire $\hat{Y}_{\text{GREG}}$. Les auteurs créent une population finie U de taille $N = 2\,000$ avec les valeurs (y_k, x_k), $k = 1, ..., 2\,000$, telles que $\log(y_k) = 1 + x_k + \varepsilon_k$; les 2 000 valeurs de x_k sont des réalisations de la variable aléatoire Gamma (1,1) et ε_k est une erreur normalement distribuée. L'information auxiliaire consiste en la taille de population N et les valeurs connues x_k pour $k = 1, ..., 2\,000$. Ils tirent ensuite des échantillons aléatoires simples répétés de taille n = 100. Pour les deux estimateurs, le modèle auxiliaire est le modèle loglinéaire $E_\xi(y_k | x_k) = \mu_k$ avec $\log(\mu_k) = \alpha + \beta x_k$. Ce modèle est ajusté à chaque échantillon, en utilisant la méthode d'estimation du pseudomaximum de quasi-vraisemblance. Les valeurs ajustées $\hat{y}_k = \exp(\hat{\alpha} + \hat{\beta} x_k)$ sont utilisées pour former $\hat{Y}_{\text{MCAL}}$ ainsi que $\hat{Y}_{\text{GREG}}$. La variance de simulation est nettement plus faible pour $\hat{Y}_{\text{MCAL}}$. (Le modèle GREG linéaire (3.2), identique à l'estimateur par calage sans modèle, est également inclus dans l'étude de Wu et Sitter; fait peu étonnant, il est encore moins efficace que l'estimateur GREG non linéaire sous la forte relation non linéaire imposée dans leur expérience.

Montanari et Ranalli (2005) fournissent d'autres données, pour plusieurs populations créées artificiellement, sur la comparaison entre $\hat{Y}_{\text{MCAL}}$ et l'estimateur non linéaire $\hat{Y}_{\text{GREG}}$. Leur modèle auxiliaire, $y_k = \mu_k + \varepsilon_k$, est ajusté par régression non paramétrique (lissage polynomial local) donnant les prédictions $\hat{y}_k = \hat{\mu}_k$ pour $k \in U$. Dans le cas de ce type d'ajustement du modèle, les prédictions $\hat{y}_k = \hat{\mu}_k$ sont hautement exactes. Naturellement, l'estimateur par calage fondé sur un modèle $\hat{Y}_{\text{MCAL}}$ ne donne lieu qu'à une amélioration marginale comparativement à l'estimateur non linéaire $\hat{Y}_{\text{GREG}}$.

Nous pouvons résumer la méthode du calage de la façon suivante. L'estimateur de $Y = \sum_U y_k$ a la forme linéairement pondérée $\hat{Y} = \sum_s w_k y_k$. Dans le calage linéaire (sans modèle), l'équation de calage s'écrit $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$; un total de population auxiliaire connu $\sum_U \mathbf{x}_k$ est requis, mais non une information auxiliaire complète ($\mathbf{x}_k$

connu pour tout $k \in U$). Les mêmes poids peuvent être appliqués à toutes les variables y (pondération polyvalente); l'estimateur est identique à l'estimateur GREG linéaire (mais dérivé selon un raisonnement différent). Dans le calage fondé sur un modèle, la moyenne sous le modèle auxiliaire μ_k est non linéaire en $\mathbf{x}_k$; l'information auxiliaire complète est habituellement requise; les contraintes de calage comprennent l'équation $\sum_s w_k \hat{y}_k = \sum_U \hat{y}_k$; les poids w_k dépendent des valeurs y_k, ce qui implique la perte de la propriété de polyvalence.

5. Aspects du calcul, poids extrêmes et valeurs aberrantes

Le calcul des poids calés soulève d'importantes questions d'ordre pratique qui sont traitées dans un certain nombre d'articles. Dans la production à grande échelle de statistiques d'un organisme statistique national, tous les calculs doivent se dérouler harmonieusement, de manière routinière. Les valeurs de pondération inappropriées (ou indûment variables) doivent être évitées. De nombreux praticiens soutiennent raisonnablement que tous les poids doivent être positifs (voir même supérieurs à l'unité) et que les valeurs très élevées doivent être évitées.

Quelques-uns des poids calculés selon (4.2) peuvent s'avérer très grands ou négatifs. Huang et Fuller (1978), ainsi que Park et Fuller (2005) ont proposé des méthodes permettant d'éviter ces pondérations indésirables.

Dans la méthode de minimisation de la distance, la fonction de distance peut être formulée de manière à exclure les poids négatifs, tout en satisfaisant les équations de calage données. Le logiciel CALMAR (Deville, Särndal et Sautory 1993) permet d'utiliser plusieurs fonctions de distance de ce type. Une version étendue, CALMAR2, est décrite dans LeGuennec et Sautory (2002). D'autres organismes statistiques ont développé leur propre logiciel pour le calcul des pondérations. Le SGE de Statistique Canada, le CLAN97 de Statistique Suède, le Bascula 4.0 du Bureau central de la statistique des Pays-Bas et le *g*-CALIB-S de Statistique Belgique en sont des exemples. Chacun à leur façon, ces logiciels visent à résoudre les problèmes de calcul qui se posent. Dans chaque cas, l'utilisateur doit consulter le guide de l'utilisateur afin de savoir exactement comment sont traités les problèmes de calcul, y compris la manière d'éviter les poids indésirables.

Dans le SGE, un programme mathématique minimise la distance du chi-carré, conditionnellement aux contraintes de calage et aux bornes individuelles sur les poids, de façon que ceux-ci satisfassent $A_k \le w_k \le B_k$ pour les valeurs spécifiées A_k, B_k. Bascula 4.0 est décrit dans Nieuwenbroek et Boonstra (2002). Le logiciel *g*-CALIB-S, décrit dans Vanderhoeft, Waeytens et Museux (2001), ainsi que Vanderhoeft (2001), s'appuie sur l'inverse généralisée (la Moore-Penrose) pour le calcul des poids, si bien qu'il n'y a pas lieu de s'inquiéter d'une redondance éventuelle dans l'information auxiliaire.

Dans Bankier, Houle et Luc (1997), l'objectif est double, à savoir maintenir les poids calculés entre les bornes souhaitées et laisser tomber certaines variables x afin d'éliminer les dépendances presque linéaires. Isaki, Tsay et Fuller (2004) considèrent la programmation quadratique pour obtenir, pour les ménages ainsi que pour les personnes, des poids qui sont compris entre les bornes spécifiées.

Une intervention touchant les poids (afin d'éliminer les valeurs de pondération indésirables) amène à se demander dans quelle mesure on peut s'écarter des poids de sondage d_k sans compromettre la propriété désirable d'estimation presque sans biais sous le plan. Une idée qui a été mise à l'épreuve consiste à modifier l'ensemble de contraintes de façon que la différence entre l'estimateur pour les variables auxiliaires et pour les totaux de population connus correspondants soit comprise entre des marges de tolérance. Ainsi, Chambers (1996) minimise une fonction de perte biaisée ou fonction de perte ridge.

Des valeurs aberrantes dans les variables auxiliaires peuvent être la cause des poids extrêmes. Le calage en présence de valeurs aberrantes est examiné par Duchesne (1999). Sa méthode de « calage robuste » peut introduire dans les estimations un biais qui pourrait toutefois être plus que compensé par une réduction de la variance.

Si l'ensemble de contraintes est étendu de façon à limiter les poids à des intervalles précis, il n'est pas certain que le problème d'optimisation aura une solution. L'existence de celle-ci est considérée dans Théberge (2000), qui propose aussi des méthodes de traitement des valeurs aberrantes.

6. Estimation par calage de paramètres plus complexes

La méthode de calage peut être adaptée à l'estimation de paramètres plus complexes qu'un total de population. Nous examinons certains exemples à la présente section. Nous continuons de supposer que nous nous trouvons dans des conditions d'échantillonnage à une seule phase et de réponse complète, et nous utilisons la même notation qu'à la section 2. Un exemple est l'estimation des quantiles de population (section 6.1), un autre est l'estimation des fonctions de totaux (section 6.2). D'autres exemples rentrant dans cette catégorie, que nous ne passons pas en revue ici, sont ceux de Théberge (1999), pour l'estimation de paramètres bilinéaires et de Tracy, Singh et Arnab (2003), pour le calage par rapport aux moments de deuxième ordre.

6.1 Calage de l'estimation des quantiles

La médiane et d'autres quantiles de la population finie sont des mesures descriptives importantes, particulièrement dans le cas des enquêtes économiques. Afin d'estimer les quantiles, il faut d'abord estimer la fonction de répartition de la population finie. Avant que l'usage du calage se répande, plusieurs auteurs ont considéré l'estimation des quantiles, avec ou sans l'utilisation d'information auxiliaire. Les auteurs d'articles plus récents se sont tournés vers la méthode de calage dans le même but, dont Kovaĉević (1997), Wu et Sitter (2001), Ren (2002), Tillé (2002), Harms (2003), Harms et Duchesne (2006), et Rueda et coll. (2007). Comme l'illustrent ces articles, il existe plus d'un moyen d'appliquer cette méthode. Le caractère non lisse de la fonction de répartition de la population finie cause certaines complexités dont la résolution varie selon les auteurs.

Soit $\Delta(\cdot)$ la fonction de Heaviside, définie pour tout réel z de manière que $\Delta(z)=1$ si $z \geq 0$ et $\Delta(z)=0$ si $z<0$. La fonction de répartition inconnue de la variable étudiée y est

$$F_y(t)=\frac{1}{N}\sum_U \Delta(t-y_k). \qquad (6.1)$$

Le quantile α de la population finie est défini comme étant $Q_{y\alpha}=\inf\{t|F_y(t)\geq\alpha\}$. La variable auxiliaire x_j, qui prend les valeurs x_{jk}, possède la fonction de répartition $F_{x_j}(t)=(1/N)\sum_U \Delta(t-x_{jk})$ avec le quantile α dénoté $Q_{x_j\alpha}$, $j=1,2,...,J$. Un estimateur naturel de $F_y(t)$ basé sur les poids de sondage $d_k=1/\pi_k$ est

$$\hat{F}_y(t)=\frac{1}{\sum_s d_k}\sum_s d_k\Delta(t-y_k).$$

Un estimateur par calage de $F_y(t)$ prend la forme

$$\hat{F}_{y\mathrm{CAL}}(t)=\frac{1}{\sum_s w_k}\sum_s w_k\Delta(t-y_k) \qquad (6.2)$$

où les poids w_k sont calés comme il convient sur une information auxiliaire spécifiée. Puis, nous tirons de $\hat{F}_{y\mathrm{CAL}}(t)$ l'estimateur du quantile α donné par $\hat{Q}_{y\alpha}=\inf\{t|\hat{F}_{y\mathrm{CAL}}(t)\geq\alpha\}$. Une formule analogue à (6.2) est vérifiée pour $\hat{F}_{x_j\mathrm{CAL}}(t)$.

Sans référence explicite à un modèle, Harms et Duchesne (2006) spécifient l'information disponible pour le calage comme étant une taille de population connue, N, et les quantiles de population connue $Q_{x_j\alpha}$ pour $j=1,2,...,J$. L'information auxiliaire complète, avec les valeurs $\mathbf{x}_k=(x_{k1},...,x_{kJ})'$ connues pour $k\in U$, n'est pas nécessaire. (Toutefois, dans la pratique, l'information complète serait généralement requise, parce qu'il est peu probable que les valeurs exactes des quantiles de plusieurs variables x puissent être importées de sources externes.) Ils déterminent les poids w_k de façon à minimiser la distance du chi-carré $\sum_s(w_k-d_k)^2/2d_kq_k$, pour la valeur spécifiée de q_k, sous la contrainte des équations de calage

$$\sum_s w_k=N;\ \hat{Q}_{x_j\mathrm{CAL},\alpha}=Q_{x_j\alpha},\ j=1,2,...,J$$

pour des estimations définies convenablement $\hat{Q}_{x_j\mathrm{CAL},\alpha}$. Maintenant, si nous décidions de spécifier $\hat{Q}_{x_j\mathrm{CAL},\alpha}$ comme $\hat{Q}_{x_j\alpha}=\inf\{t|\hat{F}_{x_j\mathrm{CAL}}(t)\geq\alpha\}$, il serait généralement impossible de trouver une solution exacte au problème de calage tel qu'il est énoncé. Harms et Duchesne choisissent plutôt de lui substituer des estimateurs lissés, qu'ils appellent « estimateurs interpolés de répartition » des fonctions de répartition $F_{x_j}(t)$, $j=1,2,...,J$. Ils remplacent $\Delta(\cdot)$ par une fonction légèrement modifiée. Il est alors possible d'obtenir les poids w_k, ainsi qu'une fonction de répartition estimée correspondante $\hat{F}_{y\mathrm{CAL}}(t)$; enfin, ils estiment $Q_{y\alpha}$ comme $\hat{Q}_{y\alpha}=\hat{F}^{-1}_{y\mathrm{CAL}}(\alpha)$.

Les poids calés résultants w_k nous permettent d'extraire les quantiles de population connus des variables auxiliaires. La chose est rassurante, car on s'attendrait à ce que ces poids produisent des estimateurs raisonnables des quantiles de la variable étudiée y. De surcroît, dans le cas d'une variable auxiliaire scalaire unique x, l'estimateur par calage résultant donne des quantiles de population exacts pour y quand la relation entre y et x est parfaitement linéaire, c'est-à-dire quand $y_k=\beta x_k$ pour tout $k\in U$. Une idée faisant intervenir des fonctions de répartition lissées est également mentionnée dans Tillé (2002).

La méthode mathématiquement plus simple de Rueda et coll. (2007) est une application du calage fondée sur un modèle en ce que leur calage se fait par rapport à un total de population des valeurs *prévues* de y. Elle requiert une information auxiliaire complète. Partant de la valeur connue $\mathbf{x}_k$, elle consiste à calculer d'abord les prédictions linéaires $\hat{y}_k=\hat{\boldsymbol{\beta}}'\mathbf{x}_k$ pour $k\in U$, avec $\hat{\boldsymbol{\beta}}=(\sum_s d_kq_k\mathbf{x}_k\mathbf{x}_k')^{-1}(\sum_s d_kq_k\mathbf{x}_ky_k)$, où $d_k=1/\pi_k$ et les q_k sont des facteurs d'échelle spécifiés. Les poids w_k sont obtenus par minimisation de la distance du chi-carré sous la contrainte des équations de calage énoncées en fonction des prédictions, de façon à réaliser la convergence à J points choisis arbitrairement t_j, $j=1,...,J$:

$$\frac{1}{N}\sum_s w_k\Delta(t_j-\hat{y}_k)=F_{\hat{y}}(t_j),\ j=1,...,J$$

où $F_{\hat{y}}(t_j)$ est la fonction de répartition en population finie des prédictions $\hat{y}_k$, évaluées à t_j. Les auteurs pensent qu'un assez petit nombre de points sélectionnés arbitrairement t_j peut suffire, disons moins de 10. Une fois que les w_k sont déterminés, l'estimation du quantile α est obtenue d'après $\hat{F}_{y\mathrm{CAL}}(t)=(1/N)\sum_s w_k\Delta(t-y_k)$.

L'estimation des quantiles illustre bien le fait que la méthode de calage peut être exécutée de plus d'une façon lorsqu'on estime des paramètres un peu plus complexes. Les deux méthodes mentionnées ici donnent une estimation presque sans biais sous le plan. Les poids de Harms et Duchesne (2006) sont polyvalents, indépendants de la variable y. Par contre, la méthode de Rueda et coll. (2007) nécessite un nouvel ensemble de poids pour chaque nouvelle variable y. Des données empiriques, obtenues par simulation, donnent à penser que les deux méthodes se comparent favorablement aux méthodes antérieures d'estimation des quantiles, non fondées explicitement sur l'approche du calage (mais sur la même information auxiliaire).

L'extension de la méthode du calage à l'estimation d'autre paramètres complexes, tels que le coefficient de Gini, est esquissée dans Harms et Duchesne (2006).

6.2 Calage pour d'autres paramètres complexes

Plikusas (2006), ainsi que Krapavickaitė et Plikusas (2005) examinent l'estimation par calage de certaines fonctions de totaux de population. (Leur expression « calage non linéaire » signifie « fonction non linéaire de totaux » et n'est pas utilisée ici.) Un exemple simple est l'estimation du ratio de deux totaux, $R = \sum_U y_{1k} / \sum_U y_{2k}$, où y_{1k} et y_{2k} sont les valeurs, pour l'élément k, des variables y_1 et y_2, respectivement. (En fait, la fonction de répartition (6.1) est également du type ratio, avec $y_{2k} = 1$ et $N = \sum_U 1$ comme total au dénominateur.) Ces auteurs étudient l'estimateur par calage $\hat{R}_{\text{CAL}} = \sum_s w_k y_{1k} / \sum_s w_k y_{2k}$. Les poids w_k, communs au numérateur et au dénominateur, sont déterminés par calage sur l'information auxiliaire énoncée comme il suit : Il existe une variable auxiliaire, x_{1k}, pour y_{1k}, et une autre, x_{2k}, pour y_{2k}; le ratio des totaux $R_0 = \sum_U x_{1k} / \sum_U x_{2k}$ est une valeur connue, obtenue par un dénombrement complet antérieur ou provenant d'une autre source fiable. L'équation de calage proposée est $\sum_s w_k e_k = 0$, où $e_k = x_{1k} - R_0 x_{2k}$. Parce que $\sum_U e_k = 0$, selon la méthode de la distance du chi-carré minimale, les poids sont

$$w_k = d_k \left\{ 1 - \left(\sum\nolimits_s d_k e_k \right) \left(\sum\nolimits_s d_k e_k^2 \right)^{-1} e_k \right\}.$$

Ces poids extraient correctement la valeur connue du ratio R_0; en posant que $y_{1k} = x_{1k}$ et $y_{2k} = x_{2k}$ dans $\hat{R}_{\text{CAL}}$, nous obtenons

$$\frac{\sum_s w_k x_{1k}}{\sum_s w_k x_{2k}} - R_0 = \frac{\sum_s w_k e_k}{\sum_s w_k x_{2k}} = 0.$$

Les données empiriques présentées dans Plikusas (2006), ainsi que dans Krapavickaitė et Plikusas (2005) donnent à penser que l'estimateur par calage se compare favorablement (variance plus faible, tout en maintenant la quasi-absence de biais sous le plan) à d'autres estimateurs, établis suivant d'autres arguments que le calage, tout en s'appuyant sur la même information auxiliaire.

7. Comparaison du calage à d'autres approches

Comme beaucoup l'ont souligné, les utilisateurs voient dans le calage un moyen simple et convaincant d'intégrer l'information auxiliaire, pour des paramètres simples (section 4), ainsi que pour des paramètres plus complexes, tels que les quantiles, les ratios et d'autres (section 6). Sa simplicité et son caractère pratique sont des avantages indéniables, mais peut-on aussi affirmer que le calage est « supérieur du point de vue théorique »? Existe-t-il des situations où l'on peut montrer que le calage donne des réponses plus exactes et/ou plus satisfaisantes à des questions d'importance que d'autres approches fondées sur le plan de sondage?

La section 4.5 donne une raison de penser que l'approche du calage pourrait être supérieure à l'approche GREG, parce que le calage basé sur un modèle pourrait donner des estimations plus précises que l'estimateur GREG non linéaire, pour le même modèle auxiliaire. La section 7.1 qui suit offre un autre exemple où l'approche du calage et l'approche GREG produisent des réponses divergentes, l'avantage allant à la méthode de calage.

7.1 Un exemple d'estimation par domaine

L'exemple présenté ici, tiré d'Estevao et Särndal (2004), illustre, pour une situation pratique simple, un conflit entre les résultats de l'approche GREG et de celle du calage. Le contexte est l'estimation du total y d'une sous-population (domaine).

Un échantillon probabiliste s est tiré de $\{1, 2, ..., k, ..., N\}$; les poids de sondage connus sont $d_k = 1/\pi_k$. Soit U_a un domaine; $U_a \subset U$. L'indicateur de domaine est δ_{ak} dont la valeur est $\delta_{ak} = 1$ si $k \in U_a$ et $\delta_{ak} = 0$ autrement. La grandeur estimée est le total de domaine $Y_a = \sum_U y_{ak}$, où $y_{ak} = \delta_{ak} y_k$, et y_k est observé pour $k \in s$. L'estimateur $\hat{Y}_{\text{HT}} = \sum_s d_k y_{ak}$ de Horvitz-Thompson est sans biais sous le plan de sondage, mais sa précision est faible, surtout si le domaine est petit, et l'utilisation d'information auxiliaire permettra de l'améliorer. Nous spécifions une valeur du vecteur auxiliaire $\mathbf{x}_k$ pour chaque $k \in U$.

Comme il est fréquent en pratique, les éléments qui appartiennent à un domaine d'intérêt ne sont pas précisés dans la base de sondage (s'ils le sont, on dispose d'information très puissante dès le départ, mais souvent, les conditions réelles ne sont pas aussi favorables). Néanmoins,

supposons que les éléments d'un groupe plus grand U_C sont identifiables; $U_a \subset U_C \subset U$. Par exemple, supposons que y est le « revenu » et que U_C est un groupe professionnel spécifié pour les personnes énumérées dans la base de sondage, tandis que U_a est un sous-groupe professionnel non défini dans la base de sondage. Nous pouvons identifier les sous-ensembles d'échantillons $s_C = s \cap U_C$ et $s_a = s \cap U_a$ et tirer parti du fait que nous connaissons le total $\sum_U \mathbf{x}_{Ck}$, qui peut être estimé sans biais par $\sum_s d_k \mathbf{x}_{Ck}$, où $\mathbf{x}_{Ck} = \delta_{Ck}\mathbf{x}_k$, et δ_{Ck} est l'indicateur de groupe informatif : $\delta_{Ck} = 1$ si $k \in U_C$ et $\delta_{Ck} = 0$ autrement. Le total auxiliaire de domaine $\sum_U \mathbf{x}_{ak}$ n'est pas disponible, parce que U_a n'est pas défini. Le calage en vue de satisfaire $\sum_s w_k \mathbf{x}_{Ck} = \sum_U \mathbf{x}_{Ck}$ donne l'estimateur presque sans biais sous le plan $\hat{Y}_{a\text{CAL}} = \sum_U w_k y_{ak}$, où $w_k = d_k(1+\boldsymbol{\lambda}'\mathbf{z}_k)$, avec $\boldsymbol{\lambda}' = (\sum_U \mathbf{x}_{Ck} - \sum_s d_k \mathbf{x}_{Ck})' \, (\sum_s d_k \mathbf{z}_k \mathbf{x}'_{Ck})^{-1}$. L'instrument asymptotiquement optimal pour le vecteur donné $\mathbf{x}_k$ est (voir la section 4.3) $\mathbf{z}_k = \mathbf{z}_{0Ck} = d_k^{-1} \sum_{\ell \in s}(d_k d_\ell - d_{k\ell})\mathbf{x}_{C\ell}$.

Par contre, l'approche par la régression en s'appuyant sur la même information auxiliaire mène à $\hat{Y}_{a\text{GREG}} = \sum_s d_k y_{ak} + (\sum_U \mathbf{x}_{Ck} - \sum_s d_k \mathbf{x}_{Ck})'\mathbf{B}_{\tilde{s};d}$, également presque sans biais sous le plan, où le coefficient de régression $\mathbf{B}_{\tilde{s};d} = (\sum_{\tilde{s}} d_k \mathbf{x}_k \mathbf{x}'_k)^{-1} \sum_{\tilde{s}} d_k \mathbf{x}_k y_k$ est le résultat d'un ajustement par les moindres carrés pondérés à un niveau approprié, en utilisant tous les points de données $(y_k, \mathbf{x}_k)$ disponibles pour $k \in s$ (quand $\tilde{s} = s$) ou une partie d'entre eux (quand $\tilde{s} \subset s$).

Ainsi, le modélisateur pourrait opter pour un ajustement de la régression « s'étendant au-delà du domaine » (de sorte que $\tilde{s} \supset s_a = s \cap U_a$) pour essayer d'emprunter de l'information pour $\hat{Y}_{a\text{GREG}}$ en le laissant dépendre également des données sur y provenant de l'extérieur du domaine. En revanche, $\hat{Y}_{q\text{CAL}}$ dépend exclusivement des données sur y contenues dans le domaine, ce qui est effectivement mieux. Estevao et Särndal (2004) montrent que $\hat{Y}_{a\text{CAL}}$ avec $\mathbf{z}_k = \mathbf{z}_{0Ck}$ a une variance (asymptotique) plus faible que $\hat{Y}_{a\text{GREG}}$, quelle que soit la façon dont $\tilde{s}$ est choisi. L'apport de données sur y en provenance de l'extérieur n'aide pas; l'approche du calage et l'approche de la régression ne s'accordent pas.

8. Estimation par calage en présence d'information composite

Comme l'ont montré les sections précédentes, beaucoup d'auteurs ont choisi d'étudier l'estimation dans le cas d'un échantillonnage direct, à une seule phase, d'éléments, sans aucune non-réponse. L'information disponible pour le calage est simple. Le k^{e} élément de la population finie $U = \{1, 2, ..., k, ..., N\}$ est associé à une valeur $\mathbf{x}_k$ du vecteur auxiliaire.

Toutefois, dans une importante catégorie de situations, l'information auxiliaire a une *structure composite* et sa complexité augmente avec celle du plan d'échantillonnage. Dans les plans à deux phases ou plus, ou à deux degrés ou plus, l'information comporte habituellement plusieurs composantes, qui reflètent les caractéristiques du plan de sondage. Elle est énoncée en fonction de plus d'un vecteur de variables auxiliaires. Par exemple, dans l'échantillonnage à deux degrés, une partie de l'information pourrait être disponible au sujet des unités d'échantillonnage de premier degré (grappes), et une autre partie, au sujet des unités de deuxième degré (éléments).

Par conséquent, l'estimation par calage (ou par toute autre méthode) doit systématiquement tenir compte de la structure de l'information. L'information complète comporte plusieurs éléments et le calage peut être effectué de plus d'une façon. Tous les éléments pertinents doivent être pris en compte, afin d'obtenir les estimations les plus exactes possibles. Accomplir cela de façon générale ou « optimale » n'est pas une tâche banale. L'approche du calage offre un moyen de le faire.

L'approche de la régression, appuyée sur un modèle auxiliaire dûment formulé, en est un autre, mais certains utilisateurs pourraient la juger moins directe. Donc, les enquêtes qui permettent l'utilisation d'information auxiliaire composite apportent un éclairage supplémentaire sur ce qui distingue l'approche du calage de l'approche GREG.

Nous discutons de l'échantillonnage à deux phases et de l'échantillonnage à deux degrés à la présente section. À la section 9, nous examinons un autre exemple d'information composite, dans le contexte de la correction du biais de non-réponse.

Un autre aspect de l'information composite est celui où le but est de combiner des renseignements provenant de plusieurs enquêtes, ce qui constitue aussi un moyen de renforcer les estimations et d'en accroître l'exactitude. Ce facteur est l'un de ceux (outre le désir de réaliser la cohérence entre les enquêtes pour le bénéfice de l'utilisateur des données) qui motivent la méthode de pondération répétée mentionnée antérieurement, qui a été élaborée par l'organisme statistique des Pays-Bas. L'utilisation d'informations auxiliaires combinées dans le contexte de l'estimation GREG est examinée dans Merkouris (2004).

8.1 Information composite pour les plans d'échantillonnage à deux phases

Par échantillonnage double, on entend un plan de sondage comportant le tirage de deux échantillons probabilistes, s_1 et s_2, à partir de la même population $U = \{1, ..., k, ..., N\}$. Des données auxiliaires peuvent être relevées pour les deux échantillons, mais les valeurs de la variable étudiée y_k ne sont relevées que pour $k \in s_2$ dans le but d'estimer $Y = \sum_U y_k$. Hidiroglou (2001) distingue

plusieurs formes d'échantillonnage double. Dans le cas *hiérarchique* (*ou emboîté*) (échantillonnage à deux phases classique), l'échantillon de première phase s_1 est tiré de U, et l'échantillon de seconde phase s_2 est un sous-échantillon tiré de s_1, de sorte que $U \supset s_1 \supset s_2$. Nous pouvons distinguer deux *cas non hiérarchiques* (*ou non emboîtés*). Dans le premier, s_1 est tiré de la base de sondage U_1 et s_2, de la base de sondage U_2, où U_1 et U_2 couvrent la même population U, et les unités d'échantillonnage peuvent être définies différemment dans les deux bases. Dans le deuxième cas non hiérarchique, s_1 et s_2 sont tirés indépendamment de U.

Afin d'illustrer comment l'information composite intervient dans l'estimation, considérons le cas hiérarchique. Les poids de sondage sont $d_{1k} = 1/\pi_{1k}$ (s_1 échantillonné dans U); $d_{2k} = 1/\pi_{2k} (\pi_{2k} = \pi_{k|s_1}$ dans le sous-échantillonnage de s_2 à partir de s_1). Le poids de sondage combiné est $d_k = d_{1k} d_{2k}$. L'estimateur sans biais élémentaire $\hat{Y} = \sum_{s_2} d_k y_k$ peut être amélioré grâce à l'utilisation d'information auxiliaire, spécifiée ici à deux niveaux.

Niveau de la population : La valeur du vecteur $\mathbf{x}_{1k}$ est connue (donnée dans la base de sondage) pour chaque $k \in U$, de sorte qu'elle est connue pour chaque $k \in s_1$ et pour chaque $k \in s_2$; $\sum_U \mathbf{x}_{1k}$ est un total vectoriel de population connu.

Niveau du premier échantillon : La valeur du vecteur auxiliaire $\mathbf{x}_{2k}$ est connue (observée) pour chaque $k \in s_1$ et, par conséquent, pour chaque $k \in s_2$; le total inconnu $\sum_U \mathbf{x}_{2k}$ est estimé sans biais par $\sum_{s_1} d_{1k} \mathbf{x}_{2k}$.

Quel est le meilleur moyen de tenir compte de cette information composite? Dans une adaptation de l'approche GREG, Särndal et Swensson (1987) ont formulé deux modèles auxiliaires linéaires, le premier spécifié en fonction du vecteur $\mathbf{x}_{1k}$ et le second tenant compte aussi du vecteur $\mathbf{x}_{2k}$. Les deux modèles sont ajustés et les prédictions résultantes, de deux types, sont utilisées pour créer un estimateur GREG approprié $\hat{Y}_{\text{GREG}}$ de $Y = \sum_U y_k$.

Dupont (1995) fait le commentaire important que l'information composite donnée appelle « deux approches naturelles différentes ». Outre l'approche GREG, il existe une approche par calage qui donnera les poids finaux w_k pour un estimateur par calage $\hat{Y}_{\text{CAL}} = \sum_{s_2} w_k y_k$. Il est intéressant de comparer les résultats de ces deux approches. L'une et l'autre offrent plus d'une option. Dans l'approche GREG, il existe divers moyens de formuler les modèles auxiliaires linéaires et leur structure de variance respective. Dans l'approche par le calage, diverses formulations des équations de calage sont possibles.

Voici, par exemple, une option de *calage en deux étapes* : d'abord trouver les poids intermédiaires w_{1k} qui satisfont $\sum_{s_1} w_{1k} \mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$, puis utiliser ces poids à la deuxième étape pour calculer les poids finaux w_k qui satisfont

$$\sum_{s_2} w_k \mathbf{x}_k = \sum_{s_1} w_{1k} \mathbf{x}_k = \begin{pmatrix} \sum_U \mathbf{x}_{1k} \\ \sum_{s_1} w_{1k} \mathbf{x}_{2k} \end{pmatrix}$$

où $\mathbf{x}_k$ est le vecteur auxiliaire combiné

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_{1k} \\ \mathbf{x}_{2k} \end{pmatrix}.$$

Alternativement, dans une option *à une seule étape*, nous déterminons les poids w_k directement pour satisfaire

$$\sum_{s_2} w_k \mathbf{x}_k = \begin{pmatrix} \sum_U \mathbf{x}_{1k} \\ \sum_{s_1} d_{1k} \mathbf{x}_{2k} \end{pmatrix}.$$

En général, les poids finaux w_k ne sont pas identiques dans les deux options. Supposons que $\sum_U \mathbf{x}_{1k}$ est un total $\mathbf{x}_1$ importé. Si nous examinons la situation de plus près, nous constatons que l'option en deux étapes nécessite plus d'information, parce que les valeurs connues $\mathbf{x}_{1k}$ sont requises individuellement pour $k \in s_1$, tandis que dans l'option à une seule étape, il suffit qu'elle soit disponible pour $k \in s_2$. Nous pourrions donc nous attendre à un certain avantage de l'option à deux étapes en ce qui concerne la variance, puisque $\sum_{s_1} w_{1k} \mathbf{x}_{2k}$ est souvent plus exact (en tant qu'estimateur de $\sum_U \mathbf{x}_{2k}$) que $\sum_{s_1} d_{1k} \mathbf{x}_{2k}$ dans la procédure à une seule étape. Néanmoins, cette attente n'est pas toujours confirmée et la méthode à une seule étape peut être supérieure, par exemple quand $\mathbf{x}_1$ et $\mathbf{x}_2$ sont faiblement corrélés.

Dupont (1995), ainsi qu'Hidiroglou et Särndal (1998) examinent les liens qui existent, sans surprise, entre les deux approches. Un estimateur GREG, dérivé de modèles adjoints ayant des structures de variance spécifiques, pourrait être identique à l'estimateur par calage si les poids de ce dernier sont calés d'une certaine façon. Dans d'autres cas, les différences pourraient être faibles.

L'efficacité de diverses options dépend, de manière assez subtile, de la configuration des corrélations entre y_k, $\mathbf{x}_{1k}$ et $\mathbf{x}_{2k}$. Par exemple, dans quelle mesure $\mathbf{x}_1$ et $\mathbf{x}_2$ sont-ils complémentaires, dans quelle mesure sont-ils des substituts l'un de l'autre? Dans l'approche GREG, il est difficile, voire même futile, de définir avec précision une structure de variance qui reflète vraiment la « réalité » qui sous-tend les données. L'approche du calage est plus directe et Estevao et Särndal (2002, 2006) explorent certaines de ses possibilités.

8.2 Information composite dans les plans d'échantillonnage à deux degrés

Le point commun entre l'échantillonnage à deux degrés classique (grappes échantillonnées au premier degré,

éléments sous-échantillonnés dans les grappes sélectionnées au deuxième degré) et l'échantillonnage à deux phases tient au fait que l'information totale peut comporter plus d'une composante. Il peut exister a) de l'information au niveau de la grappe (au sujet des grappes), b) de l'information au niveau de l'élément pour toutes les grappes et c) de l'information au niveau de l'élément pour les grappes selectionnées uniquement. Ici encore, les auteurs appartiennent à deux écoles, certains exploitant l'information par l'approche du calage et les autres choisissant la route de la régression généralisée GREG.

Estevao et Särndal (2006) conçoivent l'estimation par calage sous échantillonnage à deux degrés classique, où l'information composite est spécifiée comme il suit : i) pour la population de grappes $U_{\rm I}$, il existe un total connu $\sum_{U_{\rm I}} \mathbf{x}_{(c)i}$, où $\mathbf{x}_{(c)i}$ est une valeur du vecteur auxiliaire associé à la grappe U_i, pour $i \in U_{\rm I}$; ii) pour la population d'éléments $U = \bigcup_{i \in U_{\rm I}} U_i$, il existe un total connu $\sum_U \mathbf{x}_k$, où la valeur du vecteur auxiliaire $\mathbf{x}_k$ est associée à l'élément $k \in U$. Supposons que l'on doit produire à la fois les statistiques de grappe et les statistiques d'élément dans une enquête, si bien qu'il faut estimer le total de population de grappes $Y_{\rm I} = \sum_{U_{\rm I}} y_{(c)i}$ et le total de population d'éléments $Y = \sum_U y_k$.

Si aucune relation n'est imposée entre les poids des grappes $w_{{\rm I}i}$ et les poids des éléments w_k, les premiers sont calés de manière à satisfaire $\sum_{s_{\rm I}} w_{{\rm I}i} \mathbf{x}_{(c)i} = \sum_{U_{\rm I}} \mathbf{x}_{(c)i}$ et les seconds, de manière à satisfaire $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. (Ici, $s_{\rm I}$ est l'échantillon de grappes tiré de $U_{\rm I}$; s_i est l'échantillon d'éléments tiré de U_i; et $s = \bigcup_{i \in s_{\rm I}} s_i$ est l'échantillon complet d'éléments.) Alors, $\hat{Y}_{\rm ICAL} = \sum_{s_{\rm I}} w_{{\rm I}i} y_{(c)i}$ estime le total de population de grappes $Y_{\rm I}$ et $\hat{Y}_{\rm CAL} = \sum_s w_k y_k$ estime le total de population d'éléments Y.

Dans la pratique, on recourt souvent à la *pondération intégrée*, qui consiste à imposer une relation commode entre les poids des grappes $w_{{\rm I}i}$ et les poids w_k des éléments compris dans les grappes sélectionnées. Estevao et Särndal (2006) examinent deux formes de pondération intégrée.

L'une d'elles consiste à imposer $w_k = d_{k|i} w_{{\rm I}i}$, où $d_{k|i}$ est l'inverse de la probabilité d'inclusion de l'élément *k* dans la grappe *i*. (Par exemple, dans l'échantillonnage en grappes à un degré, quand tous les éléments *k* d'une grappe échantillonnée sont sélectionnés, alors $d_{k|i} = 1$. Par conséquent, la relation imposée est $w_k = w_{{\rm I}i}$ et tous les éléments de la grappe reçoivent le même poids pour le calcul des statistiques d'élément, et ce même poids est également utilisé pour calculer les statistiques de grappe.) L'équation de calage $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$ se lit alors $\sum_{s_{\rm I}} w_{{\rm I}i} \sum_{s_i} d_{k|i} \mathbf{x}_k = \sum_U \mathbf{x}_k$. Les poids de grappe $w_{{\rm I}i}$ sont maintenant calculés en minimisant $\sum_{s_{\rm I}} (w_{{\rm I}i} - d_{{\rm I}i})^2 / d_{{\rm I}i}$ sous la contrainte de l'équation de calage qui tient compte des deux types d'information :

$$\begin{pmatrix} \sum_{s_{\rm I}} w_{{\rm I}i} \mathbf{x}_{(c)i} \\ \sum_{s_{\rm I}} w_{{\rm I}i} \sum_{s_i} d_{k|i} \mathbf{x}_k \end{pmatrix} = \begin{pmatrix} \sum_{U_{\rm I}} \mathbf{x}_{(c)i} \\ \sum_U \mathbf{x}_k \end{pmatrix}. \qquad (8.1)$$

Une fois les poids de grappe $w_{{\rm I}i}$ déterminés, le calcul des poids d'élément $w_k = d_{k|i} w_{{\rm I}i}$ s'ensuit.

Une autre méthode de pondération intégrée raisonnable consiste à imposer que $\sum_{s_i} w_k = N_i w_{{\rm I}i}$. Par exemple, pour l'échantillonnage en grappes à un degré, cela implique que le poids de grappe $w_{{\rm I}i}$ est égal à la moyenne des poids d'élément w_k contenus dans la grappe.

L'échantillonnage à deux degrés est également traité dans Kim, Breidt et Opsomer (2006). Ces auteurs supposent que des données auxiliaires existent pour les grappes, par la voie d'une variable de grappe quantitative unique $x_{(c)i}$, mais non pour les éléments. Ils développent et étudient un estimateur de type GREG du total de population d'éléments $Y = \sum_U y_k$, $\hat{Y} = \sum_{i \in U_{\rm I}} \hat{\mu}_i + \sum_{i \in s_{\rm I}} d_i (\hat{t}_i - \hat{\mu}_i)$, où $\hat{t}_i$ est sans biais sous le plan pour le total de population de grappes $t_i = \sum_{U_i} y_k$, et $\hat{\mu}_i$ est obtenu par un ajustement par régression polynomiale locale. L'estimateur peut être exprimé sous la forme linéairement pondérée en utilisant des poids qui sont calés sur les totaux de population des puissances de la variable de grappe $x_{(c)i}$.

8.3 Pondération des ménages et pondération des personnes

Dans certaines grandes enquêtes sociales, l'objectif est de produire des estimations au niveau du ménage ainsi qu'au niveau de la personne, si bien que certaines variables étudiées sont des variables du ménage (grappe) et d'autres, des variables de la personne (élément). Par conséquent, un certain nombre d'auteurs ont étudié la situation de l'échantillonnage en grappes à un degré ($d_{k|i} = 1$) et de la pondération intégrée qui consiste à attribuer le même poids à tous les membres d'un ménage sélectionné, poids qui est également utilisé pour produire les statistiques au niveau du ménage. Une solution générale de ce problème de pondération, si l'information au niveau du ménage et l'information au niveau de la personne sont l'une et l'autre spécifiées, consiste à obtenir les poids des ménages $w_{{\rm I}i}$, calés comme dans l'équation (8.1) avec $d_{k|i} = 1$, puis à prendre $w_k = w_{{\rm I}i}$.

Dans plusieurs articles, l'accent est mis sur les valeurs du vecteur auxiliaire $\mathbf{x}_k$ attribuées aux personnes. Alexander (1987) calcule des poids qui minimisent la distance du chi-carré, tandis que Lemaître et Dufour (1987) et Niewenbrook (1993) calculent les poids intégrés à l'aide d'un estimateur GREG. La méthode de Lemaître et Dufour procède par construction indirecte d'une « valeur de vecteur auxiliaire équipondérée » s'appliquant à toutes les personnes membres d'un ménage sélectionné. Leur résultat est calculable par la méthode directe exposée à la section 8.2.

Le auteurs d'articles plus récents reviennent sur la question de la double pondération ménage et personne. Certains adoptent l'approche du calage et d'autres, celle de la régression généralisée GREG. Isaki, Tsay et Fuller (2004) en font un problème de pondération par calage; ils appliquent des poids calés aux totaux de contrôle au niveau du ménage, ainsi qu'au niveau de la personne et ne formulent aucun modèle auxiliaire explicite. Par contre, Steel et Clark (2007) suivent l'approche GREG, avec spécification de modèles auxiliaires linéaires et des structures de variance connexes.

9. Calage pour corriger la non-réponse

9.1 Correction classique de la non-réponse

Nombre de bons articles théoriques ont pour contexte le cas simple de la section 2, qui inclut l'absence totale de non-réponse. Ce sont de bonnes théories, mais applicables dans des conditions qui ne se réalisent que rarement, voire jamais. (À titre d'auteur d'articles dans ce domaine, je ne suis moi-même pas irréprochable.) La non-réponse existe dans pratiquement toutes les enquêtes. Bien qu'elle ne soit pas souhaitable, il s'agit d'un phénomène naturel dont la théorie devrait tenir compte d'emblée, en adoptant une perspective de sélection à deux phases.

Dans de nombreuses enquêtes, les taux de non-réponse sont aujourd'hui très élevés si on les compare à ceux d'il y a 40 ans, qui étaient si faibles qu'on pouvait essentiellement ne pas en tenir compte. De nos jours, la théorie de l'échantillonnage doit s'attacher de plus en plus souvent aux conséquences indésirables de la non-réponse. En particulier, un objectif immédiat est d'examiner le biais et d'essayer de le réduire autant que possible

Soit un échantillon probabiliste s tiré de $U = \{1, 2, ..., k, ..., N\}$; le poids de sondage connu de l'élément k est $d_k = 1/\pi_k$. Il y a non-réponse, ce qui donne un ensemble de réponses r, qui est un sous-ensemble de s; la valeur de la variable étudiée y_k est observée pour $k \in r$ uniquement. La probabilité de réponse inconnue de l'élément k est $\Pr(k \in r|s) = \theta_k$. Nous écartons l'estimateur sans biais $\hat{Y} = \sum_r d_k \phi_k y_k$, parce que $\phi_k = 1/\theta_k$ est inconnue. Si nous voulons retenir la notion de somme linéairement pondérée, comment devons-nous construire les poids? La correction par pondération de la non-réponse totale, ou non-réponse d'une unité, en se fondant sur la « modélisation de la non-réponse » se fait depuis longtemps. Le calage offre une nouvelle perspective.

Dans ce que nous pourrions appeler la « méthode classique », les poids de sondage probabilistes $d_k = 1/\pi_k$ sont d'abord corrigés de la non-réponse et, éventuellement, d'autres imperfections, telles que les valeurs aberrantes. L'information utilisée à cette étape provient souvent du regroupement des éléments échantillonnés. Enfin, si des totaux de population fiables sont disponibles, les poids de sondage corrigés sont calés sur ces totaux.

Au Canada, la méthodologie de l'Enquête sur la population active, décrite dans Statistique Canada (1998), est un exemple de cette pratique répandue. Un poids de sondage (modifié) est d'abord calculé pour un ménage donné, par multiplication de trois facteurs. Le produit du poids de sondage et d'un facteur de correction de la non-réponse est appelé le sous-poids. À la dernière étape, les sous poids sont calés sur des estimations de population postcensitaire très précises par groupe d'âge, sexe et région infraprovinciale. Les poids finaux ont les propriétés souhaitées de convergence, dans les régions d'une province, avec les estimations postcensitaires. Le biais de non-réponse qui persiste dans les estimations résultantes est inconnu, mais considéré comme modeste.

La méthode classique est intégrée dans le type d'estimateur $\hat{Y} = \sum_r d_k (1/\hat{\theta}_k) y_k$, où θ_k a été estimé par $\hat{\theta}_k$ à une étape préliminaire, en modélisant la réponse (c'est- à-dire la propension à répondre). Ce que demande la théorie au statisticien, à savoir formuler le « modèle de réponse réelle », capable de fournir des valeurs $\hat{\theta}_k$ exactes, sans biais, n'est pas une tâche facile. Toutefois, dans de nombreuses enquêtes, les facteurs $1/\hat{\theta}_k$ sont appliqués machinalement, sans esprit critique, par exemple par extension directe dans les strates déjà utilisées pour la sélection de l'échantillon.

La méthode classique est appliquée, par exemple, dans Ekholm et Laaksonen (1991) et dans Rizzo, Kalton et Brick (1996).

Souvent, les praticiens agissent comme si l'estimateur résultant $\hat{Y} = \sum_r d_k (1/\hat{\theta}_k) y_k$ (découlant d'une modélisation plus ou moins pénétrante de la réponse pour essayer d'obtenir les $\hat{\theta}_k$) était essentiellement sans biais, ce qu'il n'est pas (à moins qu'on ait eu la chance de spécifier le modèle idéal). Ils se comportent (pour les besoins de l'estimation de la variance, par exemple) comme si $\pi_k \hat{\theta}_k$ était la probabilité de sélection réelle de l'élément k dans une étape unique de sélection, ce qui n'est définitivement pas le cas. Cette pratique, dont l'origine remonte à un passé idyllique, devient de plus en plus contestable à mesure que grimpent subrepticement les taux de non réponse.

Le remplacement de θ_k par $\hat{\theta}_k$ donne inévitablement lieu à un biais. Il y a plusieurs décennies, les taux de non-réponse n'étaient habituellement que de quelques pour cents et il était justifiable d'ignorer ce biais, mais leur croissance galopante rend cette pratique indéfendable aujourd'hui. Selon les premiers principes, l'objectif est de produire une estimation sans biais et non une estimation où le carré du biais est un contributeur dominant (et inconnu) à l'erreur

quadratique moyenne. Nous devons décider de limiter le biais autant que possible. L'approche du calage peut nous aider à construire un vecteur auxiliaire qui répond à cet objectif.

9.2 Calage pour la correction du biais de non-réponse

Faisant plus ou moins contraste avec la procédure classique, un certain nombre d'articles récents mettent l'accent sur l'approche du calage pour corriger la non-réponse. Certaines références récentes à cet égard sont Deville (1998, 2002), Ardilly (2006), chapitre 3, Skinner (1998), Folsom et Singh (2000), Fuller (2002), Lundström et Särndal (1999), Särndal et Lundström (2005), et Kott (2006).

L'approche du calage débute par l'évaluation de l'information auxiliaire totale disponible au niveau de l'échantillon (valeurs des variables auxiliaires observées pour les répondants et les non-répondants) et au niveau de la population (totaux auxiliaires connus de population). L'objectif est de tirer le meilleur parti possible des deux sources combinées, afin de réduire le biais ainsi que la variance. Les poids de sondage sont modifiés, en une ou en deux étapes de calage, de façon qu'ils reflètent i) le résultat de la phase de réponse, ii) les caractéristiques individuelles des répondants et iii) l'information auxiliaire spécifiée. L'information peut se résumer comme il suit.

Niveau de la population : La valeur du vecteur auxiliaire $\mathbf{x}_k^*$ est connue (spécifiée dans la base de sondage) pour chaque $k \in U$, donc est connue pour chaque $k \in s$ et chaque $k \in r$; $\sum_U \mathbf{x}_k^*$ est un total de population connu.

Niveau de l'échantillon : La valeur du vecteur auxiliaire $\mathbf{x}_k^\circ$ est connue (observée) pour chaque $k \in s$, et est donc connue pour chaque $k \in r$; le total inconnu $\sum_U \mathbf{x}_k^\circ$ est estimé sans biais par $\sum_s d_k \mathbf{x}_k^\circ$.

Le calage sur cette information composite peut se faire en deux étapes (calcul de poids intermédiaires pour commencer, puis utilisation de ces poids à la deuxième étape pour produire les poids finaux) ou directement en une seule étape. En principe, les différences de biais et de variance des estimations devraient être modestes. Dans l'option en une seule étape, le vecteur auxiliaire combiné et l'information correspondante sont

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix}; \mathbf{X} = \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_s d_k \mathbf{x}_k^\circ \end{pmatrix}.$$

En utilisant une extension de la méthode du vecteur instrumental décrite à la section 4.3, nous recherchons les poids calés $w_k = d_k v_k$, où $v_k = F(\boldsymbol{\lambda}' \mathbf{z}_k)$ est le facteur de correction de la non-réponse, avec un vecteur $\boldsymbol{\lambda}$ déterminé par l'équation de calage $\sum_r w_k \mathbf{x}_k = \mathbf{X}$; l'estimateur par calage résultant est $\hat{Y}_{\text{CAL}} = \sum_r w_k y_k$. Il suffit de spécifier la valeur du vecteur instrumental $\mathbf{z}_k$ pour les répondants; $\mathbf{z}_k$ peut être différent de $\mathbf{x}_k$. La fonction $F(\cdot)$ joue le même rôle qu'aux section 4.2 et 4.3. Ici, $F(\boldsymbol{\lambda}' \mathbf{z}_k)$ estime implicitement la probabilité de réponse inverse, $\phi_k = 1/\theta_k$ comme l'ont souligné Deville (2002), Dupont (1995) et Kott (2006). Dans le cas linéaire, $F(u) = 1 + u$ et $v_k = 1 + \boldsymbol{\lambda}' \mathbf{z}_k$, avec $\boldsymbol{\lambda}' = (\sum_U \mathbf{x}_k - \sum_r d_k \mathbf{x}_k)' (\sum_r d_k \mathbf{z}_k \mathbf{x}_k')^{-1}$.

Bien qu'elles ne soient observées que pour les éléments échantillonnés, les variables qui constituent le vecteur $\mathbf{x}_k^\circ$ peuvent être de la plus haute importance pour la réduction du biais de non-réponse (quoique moins importantes que les $\mathbf{x}_k^*$ pour la réduction de la variance). Par exemple, Beaumont (2005b) discute des variables du processus de collecte des données qui peuvent être utilisées pour construire la composante vectorielle $\mathbf{x}_k^\circ$.

9.3 Construction du vecteur auxiliaire

Dans certaines enquêtes, les variables auxiliaires possibles abondent, comme le signalent, par exemple, Rizzo, Kalton et Brick (1996), de même que Särndal et Lundström (2005). Ainsi, pour les enquêtes auprès des ménages et des particuliers en Scandinavie, on peut obtenir une foule de variables auxiliaires éventuelles par appariement des données d'enquête à celles des registres administratifs de haute qualité existants. Il convient ensuite de décider lesquelles de ces variables devraient être incluses dans le vecteur auxiliaire $\mathbf{x}_k$ afin que celui-ci soit aussi efficace que possible, notamment en ce qui concerne la réduction du biais. Comme le mentionnent Rizzo, Kalton et Brick (1996), le choix des variables auxiliaires est probablement plus important que celui de la méthode de pondération.

Examinons le biais quand $\mathbf{z}_k = \mathbf{x}_k$. Nous devons comparer divers vecteurs $\mathbf{x}_k$ afin de choisir, en dernière analyse, celui qui produira vraisemblablement le biais le plus faible. (Nous posons que $\mathbf{x}_k$ est tel que $\boldsymbol{\mu}' \mathbf{x}_k = 1$ pour tout k et un vecteur constant $\boldsymbol{\mu}$, ce qui est le cas de nombreux vecteurs $\mathbf{x}_k$, y compris les exemples 1 à 5 présentés au début de la section 2.) Nous obtenons une bonne approximation du biais de $\hat{Y}_{\text{CAL}}$ par linéarisation de Taylor de la forme $\text{biaisapprox}(\hat{Y}_{\text{CAL}}) = (\sum_U \mathbf{x}_k)' (\mathbf{B}_{U;\theta} - \mathbf{B}_U)$, où intervient la différence entre le coefficient de régression pondéré $\mathbf{B}_{U;\theta} = \left(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_U \theta_k \mathbf{x}_k y_k$ et le coefficient non pondéré $\mathbf{B}_U = (\sum_U \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_U \mathbf{x}_k y_k)$. À moins que tous les θ_k soient égaux, le biais causé par l'écart entre les deux vecteurs de coefficients de régression pourrait être important, même si $\mathbf{x}_k$ semble être un « bon vecteur auxiliaire ». Cette expression du biais approximatif (*nearbias*) est donnée dans Särndal et Lundström (2005). Des expressions apparentées du biais (sous des conditions

différentes) sont proposées dans Bethlehem (1988) et dans Fuller, Loughin et Baker (1994). Nous pouvons aussi écrire $\text{biaisapprox}(\hat{Y}_{\text{CAL}}) = \sum_U (\theta_k M_k - 1) y_k$, où $M_k = (\sum_U \mathbf{x}_k)'(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k$. Dans la comparaison des variantes possibles de $\mathbf{x}_k$, une référence commode est le « vecteur auxiliaire primitif », $\mathbf{x}_k = 1$ pour tout $k \in U$, qui donne $\hat{Y}_{\text{CAL}} = N \bar{y}_r = N \sum_r y_k / n_r$, où n_r est le nombre de répondants, avec $\text{biaisapprox}(N \bar{y}_r) = N(\bar{y}_{U;\theta} - \bar{y}_U)$, où $\bar{y}_{U;\theta} = \sum_U \theta_k y_k / \sum_U \theta_k$ et $\bar{y}_U = \sum_U y_k / N$. Le ratio

$$\text{biaisrel}(\hat{Y}_{\text{CAL}}) = \frac{\text{biaisapprox}(\hat{Y}_{\text{CAL}})}{\text{biaisapprox}(N \bar{y}_r)} = \frac{\sum_U (\theta_k M_k - 1) y_k}{N(\bar{y}_{U;\theta} - \bar{y}_U)}$$

indique dans quelle mesure un vecteur candidat $\mathbf{x}_k$ réussit à contrôler le biais, comparativement au vecteur primitif. Nous recherchons un vecteur $\mathbf{x}_k$ qui donne un biais faible. Cependant, $\text{biaisrel}(\hat{Y}_{\text{CAL}})$ n'est pas un indicateur de biais calculable, car il dépend de y_k non observé et de θ_k inobservable. Il nous faut un indicateur calculable qui s'approche de $\text{biaisrel}(\hat{Y}_{\text{CAL}})$ et dépend du vecteur $\mathbf{x}$, mais non des variables y, qui peuvent être nombreuses dans l'enquête.

Il est facile de voir que $\text{biaisrel}(\hat{Y}_{\text{CAL}}) = 0$ dans le cas d'un vecteur $\mathbf{x}$ idéal (probablement inexistant) tel que $\phi_k = 1/\theta_k = \boldsymbol{\lambda}' \mathbf{x}_k$ pour tout $k \in U$ et un vecteur constant $\boldsymbol{\lambda}$.

Pour un vecteur $\mathbf{x}$ qui peut effectivement être construit dans le contexte de l'enquête, nous pouvons au moins obtenir les prédictions de ϕ_k. Si nous déterminons $\boldsymbol{\lambda}$ de manière à minimiser $\sum_U \theta_k (\phi_k - \boldsymbol{\lambda}' \mathbf{x}_k)^2$, nous obtenons $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}_U$, où $\hat{\boldsymbol{\lambda}}_U' = (\sum_U \mathbf{x}_k)'(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k')^{-1}$; la valeur prévue de ϕ_k est $\hat{\phi}_{kU} = \hat{\boldsymbol{\lambda}}_U' \mathbf{x}_k = M_k$. Les premier et deuxième moments (pondérés par thêta) des prédictions $\hat{\phi}_{kU} = M_k$ sont, respectivement, $\bar{M}_{U;\theta} = \sum_U \theta_k M_k / \sum_U \theta_k = N / \sum_U \theta_k = 1/\bar{\theta}_U$ et

$$Q = \frac{1}{\sum_U \theta_k} \sum_U \theta_k (M_k - \bar{M}_{U;\theta})^2 = (1/\bar{\theta}_U)(\bar{M}_U - 1/\bar{\theta}_U)$$

où $\bar{M}_U = \sum_U M_k / N$. Särndal et Lundström (2007) montrent que, dans certaines conditions, la relation entre $\text{biaisrel}(\hat{Y}_{\text{CAL}})$ et Q est approximativement linéaire,

$$\text{biaisrel}(\hat{Y}_{\text{CAL}}) \approx 1 - \frac{Q}{Q_0}$$

où $\bar{\phi}_U = \sum_U \phi_k / N$ et $Q_0 = (1/\bar{\theta}_U)(\bar{\phi}_U - 1/\bar{\theta}_U)$ est la valeur maximale de Q. Donc, si Q est calculable, il pourrait servir d'indicateur pour comparer les différents vecteurs $\mathbf{x}_k$ candidats. Nous obtenons plutôt un analogue calculable $\hat{Q}$ de Q en tant que variance des prédictions basées sur l'échantillon correspondantes $\hat{\phi}_{ks} = \hat{\boldsymbol{\lambda}}_s' \mathbf{x}_k = (\sum_s d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k = m_k$, de sorte que

$$\hat{Q} = \frac{1}{\sum_r d_k} \sum_r d_k (m_k - \bar{m}_{r;d})^2 = \bar{m}_{r;d}(\bar{m}_{s;d} - \bar{m}_{r;d})$$

où

$$\bar{m}_{r;d} = \frac{\sum_r d_k m_k}{\sum_r d_k} = \frac{\sum_s d_k}{\sum_r d_k}; \bar{m}_{s;d} = \frac{\sum_s d_k m_k}{\sum_s d_k}.$$

Nous nous attendons à ce que le biais relatif *(relbias*) diminue de manière approximativement linéaire à mesure que $\hat{Q}$ augmente. Par conséquent, indépendamment des variables y, $\hat{Q}$ peut être utilisé comme outil pour classer les divers vecteurs $\mathbf{x}$ en fonction de leur capacité à réduire le biais.

Nous pouvons nous servir de $\hat{Q}$ comme outil de sélection des variables x qu'il convient d'inclure dans le vecteur $\mathbf{x}_k$, par exemple par régression multiple ascendante de sorte que les variables soient ajoutées à $\mathbf{x}_k$ une à la fois, celle qui entre à une étape donnée étant celle qui produit l'accroissement le plus important de $\hat{Q}$. La méthode est décrite dans Särndal et Lundström (2007).

10. Calage pour tenir compte d'autres erreurs non dues à l'échantillonnage

Les erreurs dues à la non-réponse jouent un rôle déterminant dans la qualité des statistiques publiées. En revanche, si nous examinons la place que pourrait tenir l'approche du calage dans le traitement des erreurs non dues à l'échantillonnage d'autres sources que la non-réponse, il n'est pas surprenant que la littérature soit jusqu'à présent moins abondante. Néanmoins, plusieurs auteurs esquissent une approche de calage en vue d'intégrer également les erreurs de base de sondage, les erreurs de mesure et les valeurs aberrantes. Le calage pourrait offrir une théorie plus générale de l'estimation dans le contexte des enquêtes qui engloberaient les diverses erreurs non dues à l'échantillonnage.

Comme le souligne Deville (2004), le concept de calage s'applique aisément et efficacement à une grande variété de problèmes posés par les sondages. Selon lui, sa portée dépasse celle de l'estimation par la régression, une notion à laquelle certains semblent vouloir réduire l'approche du calage. Il expose sommairement comment cette approche permet de traiter plusieurs erreurs dues à la non-réponse.

Folsom et Singh (2000) présentent une méthode de calage des poids s'appuyant sur ce qu'ils appellent le modèle exponentiel généralisé (GEM). Elle comporte trois volets : le traitement des valeurs extrêmes, la correction de la non-réponse et le calage par poststratification. Elle fournit des contrôles intégrés pour les valeurs extrêmes. Le calage pour corriger à la fois les erreurs de couverture (sous- ou

surdénombrement dans la base de sondage) et la non-réponse est exposé dans Särndal et Lundström (2005) et dans Kott (2006). Skinner (1998) discute de l'utilisation du calage en présence de non-réponse et d'erreurs de mesure. Son commentaire sur la nécessité de poursuivre les travaux de recherche en vue d'étudier les propriétés des estimations par calage en présence d'erreur non due à l'échantillonnage demeure un défi près de dix ans plus tard.

11. Conclusion

Si une question doit être choisie en vue de formuler la conclusion du présent exposé, il s'agit selon moi du concept d'information auxiliaire, qui est le concept central de l'article. Sans information auxiliaire, il n'y a pas de calage, car il n'existe rien sur quoi l'exécuter. J'ai mentionné par ailleurs que l'estimation par la régression est un autre moyen, qui suit un raisonnement différent, de tenir compte de l'information auxiliaire dans l'estimation.

L'un des objectifs du présent article était de brosser le tableau de deux types de raisonnement et de souligner en quoi ils se distinguent. Des exemples montrent comment la réalisation d'un objectif d'estimation essentiellement semblable est abordée par certains auteurs suivant la logique du calage et par d'autres, suivant celle de la régression généralisée GREG (ou du moins principalement suivant l'un ou l'autre de ces types de raisonnement). Les estimateurs respectifs que ces auteurs finissent par recommander peuvent ou non donner des résultats concordants. Que les écarts aient ou non des conséquences importantes (pour ce qui est de la variance, du biais, de questions pratiques telles que la convergence et la transparence) dépend de la situation. Le présent article arrivera peut-être à mieux faire comprendre ce qui distingue deux courants de pensée qui ont guidé les chercheurs spécialisés en échantillonnage.

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.

Andersson, P.G., et Thorburn, D. (2005). Une distance de calage optimale menant à un estimateur par la régression optimal. *Techniques d'enquête*, 31, 103-107.

Ardilly, P. (2006). *Les techniques de sondage*. Paris : Éditions Technip.

Bankier, M.D., Rathwell, S. et Majkowski, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census. Document de travail, Census Operations Section, Social Surveys Methods Division, Statistique Canada.

Bankier, M., Houle, A.M. et Luc, M. (1997). Calibration estimation in the 1991 and 1996 Canadian censuses. *Proceedings*, *Section on Survey Research Methods*, American Statistical Association, 66-75.

Beaumont, J.-F. (2005a). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society* B, 67, 445-458.

Beaumont, J.-F. (2005b). L'utilisation de renseignements sur le processus de collecte des données pour traiter la non-réponse totale au moyen de l'ajustement de poids. *Techniques d'enquête*, 31, 249-254.

Beaumont, J.-F., et Alavi, A. (2004). Estimation robuste par la régression généralisée. *Techniques d'enquête*, 30, 217-231.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

Bethlehem, J.G., et Keller, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.

Breidt, F.J., et Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28, 1026-1053.

Breidt, F.J., Claeskens, G. et Opsomer, J.D. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, 92, 831-846.

Chambers, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

Chambers, R.L., Dorfman, A.H. et Wehrly, T.E. (1993). Bias robust estimation in finite populations nonparametric calibration. *Journal of the American Statistical Association*, 88, 268-277.

Deming, W.E. (1943). *Statistical Adjustment of Data*. New York : John Wiley & Sons, Inc.

Demnati, A., et Rao, J.N.K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête. *Techniques d'enquête*, 30, 17-27.

Deville, J.-C. (1998). La correction de la nonréponse par calage ou par échantillonnage équilibré. Article présenté aux Congrès de l'ACFAS, Sherbrooke, Québec.

Deville, J.-C. (2002). La correction de la nonréponse par calage généralisé. *Actes des Journées de Méthodologie*, I.N.S.E.E., Paris.

Deville, J.-C. (2004). Calage, calage généralisé et hypercalage. Document interne, I.N.S.E.E., Paris

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Deville, J.-C., Särndal, C.-E. et Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

Duchesne, P. (1999). Estimateurs de calage robustes. *Techniques d'enquête*, 25, 47-60.

Dupont, F. (1995). Redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire. *Techniques d'enquête*, 21, 141-150.

Ekholm, A., et Laaksonen, S. (1991). Weighting via response modeling in the Finish Household Budget Survey. *Journal of Official Statistics*, 3, 325-337.

Estevao, V.M., et Särndal, C.-E. (2000). A functional form approach to calibration. *Journal of Official Statistics*, 16, 379-399.

Estevao, V.M., et Särndal, C.-E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics*, 18, 233-255.

Estevao, V.M., et Särndal, C.-E. (2004). Borrowing strength is not the best technique within a wide class of design-consistent domain estimators. *Journal of Official Statistics*, 20, 645-660.

Estevao, V.M., et Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *Revue Internationale de Statistique*, 74, 127-147.

Firth, D., et Bennett, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society*, 60, 3-21.

Folsom, R.E., et Singh, A.C. (2000). The generalized exponential model for design weight calibration for extreme values, nonresponse and poststratification. *Proceedings*, *Section on Survey Research Methods*, American Statistical Association, 598-603.

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

Fuller, W.A., Loughin, M.M. et Baker, H.D. (1994). Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey de 1987-1988. *Techniques d'enquête*, 20, 79-89.

Fuller, W.A., et Rao, J.N.K. (2001). Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada. *Techniques d'enquête*, 27, 49-56.

Hansen, M.H., et Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.

Harms, T. (2003). Extensions of the calibration approach: Calibration of distribution functions and its link to small area estimators. Chintex document de travail numéro 13, Federal Statistical Office, Germany.

Harms, T., et Duchesne, P. (2006). On calibration estimation for quantiles. *Techniques d'enquête*, 32, 37-52.

Hidiroglou, M.A. (2001). L'échantillonnage double. *Techniques d'enquête*, 27, 157–169.

Hidiroglou, M.A., et Särndal, C.-E. (1998). Emploi des données auxiliaires dans l'échantillonnage à deux phases. *Techniques d'enquête*, 24, 11-20.

Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Huang, E.T., et Fuller, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings*, *Social Statistics Section*, American Statistical Association, 300-305.

Isaki, C.T., Tsay, J.H. et Fuller, W.A. (2004). Pondération de données d'échantillon reposant sur des contrôles indépendants. *Techniques d'enquête*, 30, 39-49.

Kalton, G., et Flores-Cervantes, I. (1998). Weighting methods. Dans *New Methods for Survey Research* (Éds. A.Westlake, J. Martin, M. Rigg et C. Skinner), Berkeley, U.K.: Association for Survey Computing.

Kim, J., Breidt, F.J. et Opsomer, J.D. (2005). Nonparametric regression estimation of finite population totals under two-stage sampling. Manuscript non publié.

Knottnerus, P., et van Duin, C. (2006). Variances in repeated weighting with an application to the Dutch Labour Force Survey. *Journal of Official Statistics*, 22, 565-584.

Kott, P.S. (2004). Commentaire sur Demnati et Rao : Estimateurs de la variance par linéarisation pour des données d'enquête. *Techniques d'enquête*, 30, 29-30.

Kott, P.S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Techniques d'enquête*, 32, 149-160.

Kovaĉević, M.S. (1997). Calibration estimation of cumulative distribution and quantile functions from survey data. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 139-144.

Krapavickaitė, D., et Plikusas, A. (2005). Estimation of a ratio in the finite population. *Informatica*, 16, 347-364.

Lavallée, P. (2007). *Indirect Sampling*. New York: Springer-Verlag.

LeGuennec, J., et Sautory, O. (2002). CALMAR2 : une nouvelle version de la macro CALMAR de redressement d'échantillon par calage. *Actes des Journées de Méthodologie*, INSEE, Paris.

Lehtonen, R., et Veijanen, A. (1998). Estimateur de régression généralisés logistiques. *Techniques d'enquête*, 24, 53-58.

Lehtonen, R., Särndal, C.-E. et Veijanen, A. (2003). L'effet du choix d'un modèle dans l'estimation par domaine, dont les petits domaines. *Techniques d'enquête*, 29, 37-49.

Lehtonen, R., Särndal, C.-E. et Veijanen, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, 7, 649-674.

Lemaître, G., et Dufour, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 211-220.

Lundström, S., et Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, 15, 305-327.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99, 1131-1139.

Montanari, G.E. (1987). Post-sampling efficient prediction in large-scale surveys. *Revue Internationale de Statistique*, 55, 191-202.

Montanari, G.E., et Ranalli, M.G. (2003). On calibration methods for design-based finite population inferences. Bulletin of the International Statistical Institute, 54^{e} session, volume LX, articles contribute, livre 2, 81-82.

Montanari, G.E., et Ranalli, M.G. (2005). Nonparametric model-calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100, 1429-1442.

Myrskylä, M. (2007). Generalised regression estimation for domain class frequencies. Statistics Finland Research Reports 247.

Nieuwenbroek, N.J. (1993). An integrated method for weighting characteristics of persons and households using the linear regression estimator. Rapport, Central Bureau of Statistics, Pays-Bas.

Nieuwenbroek, N.J., et Boonstra, H.J. (2002). Bascula 4.0 for weighting sample survey data with estimation of variances. The Survey Statistician, Software Reviews, Juillet 2002.

Nieuwenbroek, N.J., Renssen, R.H. et Hofman, L. (2000). Towards a generalized weighting system. Dans *Proceedings*, *Second International Conference on Establishment Surveys*, American Statistical Association, Alexandria VA.

Park, M., et Fuller, W.A. (2005). Vers des poids de régression non négatifs pour les échantillons d'enquête. *Technique d'enquête*, 31, 93-101.

Plikusas, A. (2006). Non-linear calibration. *Proceedings, Workshop on Survey Sampling*, Venspils, Latvia. Riga: Central Statistical Bureau of Latvia.

Ren, R. (2002). Estimation de la fonction de répartition et des fractiles d'une population finie. *Actes des journées de méthodologie statistique*, INSEE Méthodes, tome 1, 100, 263-289.

Renssen, R.H., et Nieuwenbroek, N.J. (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92, 368-374.

Renssen, R.H., Kroese, A.H. et Willeboordse, A.J. (2001). Aligning estimates by repeated weighting. Rapport, Central Bureau of Statistics, Pays-Bas.

Rizzo, L., Kalton, G. et Brick, J.M. (1996). Comparaison de quelques méthodes de correction de la non-réponse d'un panel. *Techniques d'enquête*, 22, 43-53.

Rueda, M., Martínez, S., Martínez, H. et Arcos, A. (2007). Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference*, 137, 435-448.

Särndal, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning and Inference*, 7, 155-170.

Särndal, C.-E., et Swensson, B. (1987). A general view of estimation for two phases of selection with application to two-phase sampling and nonresponse. *Revue Internationale de Statistique*, 55, 279-294.

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

Särndal, C.-E., et Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York : John Wiley & Sons, Inc.

Särndal, C.-E., et Lundström, S. (2007). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. Statistics Sweden : Research and development - methodology report 2007:2.

Singh, A.C., et Mohl, C.A. (1996). Comprendre les estimateurs de calage dans les enquêtes par échantillonnage. *Techniques d'enquête*, 22, 107-116.

Singh, S., Horn. S. et Yu, F. (1998). Estimation de la variance de l'estimateur général de régression : approche de calage à niveau élevé. *Techniques d'enquête*, 24, 43-52.

Skinner, C. (1998). Calibration weighting and non-sampling errors. *Proceedings International Seminar on New Techniques for Statistics,* Sorrento, novembre 4-6, 1998, 55-62.

Statistique Canada (1998). Methodology of the Canadian Labour Force Survey. Statistique Canada, Division des méthodes d'enquêtes auprès des ménages. Ottawa : Minister of Industry, catalogue no. 71-526-XPB.

Statistique Canada (2003). Quality Guidelines (quatrième édition). Ottawa : Minister of Industry, numéro de catalogue 12-539-XIE.

Steel, D.G., et Clark, R.G. (2007). Estimation par la régression au niveau de la personne et au niveau du ménage dans les enquêtes ménages. *Techniques d'enquête*, 33, 59-69.

Stukel, D.M., Hidiroglou, M.A. et Särndal, C.-E. (1996). Estimation de la variance des estimateurs de calage : comparaison des méthodes du jackknife et de la linéarisation de Taylor. *Techniques d'enquête*, 22, 117-126.

Théberge, A. (1999). Extension of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94, 635-644.

Théberge, A. (2000). Calage et poids restreints. *Techniques d'enquête*, 26, 113-122.

Tillé, Y. (2002). Estimation sans biais par calage sur la répartition dans les plans simples sans remise. *Techniques d'enquête*, 28, 83-91.

Tracy, D.S., Singh, S. et Arnab, R. (2003). Note sur le calage sous échantillonnage stratifié et double. *Techniques d'enquête*, 29, 111-116.

Vanderhoeft, C. (2001). Generalised calibration at Statistics Belgium. SPSS Module g-CALIB-S and current practices. Statistics Belgium : Document de travail no. 3. Disponible à : www.statbel.fgov.be/studies/paper03_en.asp.

Vanderhoeft, C., Waeytens, E. et Museux, J.M. (2001). Generalised calibration with SPSS 9.0 for Windows baser. Dans *Enquêtes, Modèles et Applications* (Éds. J.J. Droesbeke et L. Lebart), Paris : Dunod.

Webber, M., Latouche, M. et Rancourt, E. (2000). Harmonised calibration of income statistics. Statistique Canada, document interne, avril 2000.

Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika*, 90, 937-951.

Wu, C., et Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Zheng, H., et Little, R.J.A. (2003). Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 99-117.

Zieschang, K.D. (1990). Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 85, 986-1001.