Techniques d'enquête, décembre 2006
Vol. 32, N° 2, pp. 149-160
Statistique Canada, N° 12-001-XPB au catalogue

Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture

Phillip S. Kott [1]

Résumé

La pondération par calage peut être utilisée pour corriger la non-réponse totale et (ou) les erreurs de couverture sous des modèles appropriés de quasi-randomisation. Divers ajustements par calage qui sont asymptotiquement identiques dans un contexte d'échantillonnage pur peuvent diverger lorsqu'ils sont utilisés de cette manière. L'introduction de variables instrumentales dans la pondération par calage permet que la non-réponse (disons) soit une fonction d'un ensemble de caractéristiques différentes de celles comprises dans le vecteur de calage. Si l'ajustement par calage a une forme non linéaire, une variante du jackknife permet d'éliminer le besoin d'itération dans l'estimation de la variance.

Mots clés : Modèle prédictif; modèle de quasi-randomisation; convergent sous quasi-randomisation; variable instrumentale; ajustement proportionnel itératif (raking) généralisé.

1. Introduction

La méthode de pondération par calage a été mise au point au départ en vue de réduire les erreurs d'échantillonnage en maintenant la convergence sous randomisation. Deville et Särndal (1992) ont démontré que de nombreuses formes de pondération par calage sont asymptotiquement identiques dans le contexte de l'échantillonnage, ce qui a fait progresser grandement notre compréhension des méthodes courantes de repondération, telles que la méthode itérative du quotient aussi appelée l'ajustement proportionnel itératif (API, "raking" en anglais), qui ne se trouve pas sous le format de l'estimateur par la régression généralisée (GREG).

Folsom et Singh (2000) ont montré que la pondération par calage peut aussi être utilisée pour corriger les erreurs connues de couverture et (ou) la non-réponse totale sous des modèles appropriés de quasi-randomisation. Ces travaux n'ont été publiés dans aucune revue avec comité de lecture. Le cœur du présent article est une répétition des principaux résultats publiés dans Folsom et Singh, y compris une modification nécessaire de l'approche de Deville-Särndal en vue de modéliser l'estimation de la variance ou de l'erreur quadratique moyenne sous randomisation dans ce contexte élargi. Une version antérieure, strictement linéaire, de la pondération par calage pour l'ajustement pour la non-réponse totale peut être consultée dans Fuller, Loughin et Baker (1994). Voir aussi Lundström et Särndal (1999).

Nous faisons une distinction entre le modèle prédictif qui sous-tend habituellement le calage et le modèle de quasi-randomisation de Folsom et Singh. Toutefois, contrairement à ces deux auteurs, nous examinons ici les propriétés dans les deux cas. En outre, les variables explicatives du modèle de quasi-randomisation peuvent différer des variables de calage, ce qui est également permis dans Lundström et Särndal.

Nous proposons un nouveau jackknife qui est analogue à l'estimateur de la variance par linéarisation de Deville-Särndal. Il repose sur l'utilisation de poids de rééchantillonnage calculés en une étape, quoique les poids de calage proprement dits puissent être déterminés itérativement.

Après la présentation de la notion bien connue de pondération par calage, à la section 2, nous passons en revue le cas particulier de l'estimateur GREG dans un contexte d'échantillonnage pur. À la section 3, nous décrivons l'extension de la pondération par calage d'Estevao et Särndal (2000) dans sa forme linéaire, afin d'inclure des variables instrumentales. À la section 4, nous étendons le traitement de la pondération par calage de Deville et Särndal, afin d'inclure la possibilité de variables instrumentales. À la section 5, nous passons en revue l'estimation de la variance ou de l'erreur quadratique moyenne, et proposons un nouveau jackknife pour certains plans d'échantillonnage. À la section 6, nous décrivons comment la pondération par calage peut être utilisée pour la correction de la non-réponse. Dans ce contexte, les diverses formes fonctionnelles de la pondération par calage ne doivent plus nécessairement être asymptotiquement identiques. À la section 7, nous discutons des modèles de quasirandomisation pour les erreurs de couverture, c'est-à-dire le sous- ou le surdénombrement dans la base de sondage. À la section 8, nous donnons un petit exemple empirique appuyant le nouveau jackknife. Enfin, à la section 9, nous présentons une discussion des diverses approches et des domaines dans lesquels les travaux de recherche doivent se poursuivre.

1. Phillip S. Kott, National Agricultural Statistics Service, 3251 Old Lee Highway, Fairfax, VA 22030, États-Unis. Courriel : pkott@nass.usda.gov.

2. Pondération par calage et l'estimateur GREG

Supposons que nous connaissions la probabilité de sélection, π_k, pour chaque élément d'échantillonnage k dans l'échantillon S. Nous pouvons estimer tout total de population, $T_y = \sum_U y_k$, où U dénote la population, au moyen de l'estimateur à facteur d'extension $t_{y\_E} = \sum_S y_k/\pi_k = \sum_U y_k I_k/\pi_k$, où $I_k = 1$ quand $k \in S$ et 0 autrement. En traitant les I_k comme des variables aléatoires, il est facile de voir que t_{y_E} est un estimateur sans biais de T_y. Les propriétés qui découlent du fait que les I_k sont traitées comme des variables aléatoires sont dites *fondées sur la randomisation*. Nous pouvons également écrire $t_{y\_E} = \sum_U a_k y_k = \sum_S a_k y_k$, où $a_k = I_k/\pi_k$ est le *poids d'échantillonnage* de l'élément k.

Deville et Särndal (1992) ont inventé l'expression « estimateur par calage » pour décrire un estimateur de la forme $t_{y\_\mathrm{CAL}} = \sum_S w_k y_k$, où $\sum_S w_k \mathbf{x}_k = \sum_U \mathbf{x}_k = T_\mathbf{x}$ pour un certain vecteur ligne de variables auxiliaires, $\mathbf{x}_k = (x_{1k}, \ldots, x_{Pk})$, au sujet duquel $T_\mathbf{x}$ est connu. Puisqu'il existe généralement un continuum d'ensembles $\{w_k | k \in S\}$ qui satisfont l'*équation de calage* :

$$\sum_{k\in S} w_k \mathbf{x}_k = T_\mathbf{x}, \tag{1}$$

Deville et Särndal ont imposé comme condition que la différence entre l'ensemble de poids, $\{w_k | k \in S\}$, satisfaisant l'équation (1) et $\{a_k | k \in S\}$ minimise une fonction de perte.

Une autre approche de l'échantillonnage consiste à traiter les y_k comme des variables aléatoires satisfaisant le modèle prédictif linéaire :

$$y_k = \mathbf{x}_k \boldsymbol{\beta} + \varepsilon_k, \tag{2}$$

où $E(\varepsilon_k | \{\mathbf{x}_g, I_g | g \in U\}) = 0$ pour tout $k \in U$. En conditionnant cette espérance sur les I_g, nous supposons que l'on peut ignorer le mécanisme d'échantillonnage. Il s'agit d'un aspect critique, et parfois déraisonnable, du cadre (prédictif) *fondé sur un modèle*.

Il est facile de voir que t_{y_CAL} est un estimateur sans biais de T_y sous le modèle en ce sens que $E_\varepsilon(t_{y\_\mathrm{CAL}} - T_y) = 0$ (en supprimant le conditionnement pour simplifier la notation); l'indice ε fait référence au traitement des ε_k comme des variables aléatoires (et des I_k comme des constantes prédéterminées).

Aux fins de notre étude, l'estimateur par la régression généralisée ou GREG a la forme :

$$t_{y\_\mathrm{GREG}} = t_{y\_E} + \left(T_\mathbf{x} - \sum_{k\in S} a_k \mathbf{x}_k\right)\left(\sum_{k\in S} c_k a_k \mathbf{x}_k' \mathbf{x}_k\right)^{-1} \sum_{k\in S} c_k a_k \mathbf{x}_k' y_k, \tag{3}$$

où c_k est une constante arbitraire qui peut ou non être une fonction de $\mathbf{x}_k$, et $\lim_{N\to\infty} \sum_U c_k \mathbf{x}_k' \mathbf{x}_k / N = \boldsymbol{\Lambda}$ est une matrice définie positive, où N est la taille de U. Cette dernière condition signifie que $\sum_S c_k a_k \mathbf{x}_k' \mathbf{x}_k$ sera habituellement inversible en pratique. Par souci de commodité, nous supposerons qu'elle l'est toujours.

L'estimateur GREG de l'équation (3) peut être réécrit sous une forme de calage comme étant $t_{y\_\mathrm{GREG}} = \sum_S w_k y_k$, où

$$w_k = a_k + \left(T_\mathbf{x} - \sum_{j\in S} a_j \mathbf{x}_j\right)\left(\sum_{j\in S} c_j a_j \mathbf{x}_j' \mathbf{x}_j\right)^{-1} c_k a_k \mathbf{x}_k'.$$

Strictement parlant, les w_k sont des fonctions de l'échantillon réalisé, S, et des $c_k a_k$, mais nous supprimons cela dans la notation pour simplifier. Observons que les poids de calage peuvent être exprimés sous la forme

$$w_k = a_k(1 + c_k \mathbf{x}_k \mathbf{q}), \tag{4}$$

où $\mathbf{q} = [(\sum_S a_j c_j \mathbf{x}_j' \mathbf{x}_j)^{-1}]'(T_\mathbf{x} - \sum_S a_j \mathbf{x}_j)'$ est un vecteur colonne, puisque $\mathbf{x}_k \mathbf{q} = \mathbf{q}' \mathbf{x}_k'$.

Supposons que des conditions de régularité raisonnables soient vérifiées (voir, par exemple, Kott 2004a pour un traitement plus approfondi) et que le plan d'échantillonnage est tel que $t_{y\_E} - T_y = O_P(N/\sqrt{n})$, où n est la taille prévue de S (la taille réelle peut être aléatoire), $\sum_S a_k \mathbf{x}_k - T_\mathbf{x} = \mathbf{O}_P(N/\sqrt{n})$, et $\sum_S a_k c_k \mathbf{x}_k' \mathbf{f}_k - \sum_U c_k \mathbf{x}_k' \mathbf{f}_k = \mathbf{O}_P(N/\sqrt{n})$, où $\mathbf{f}_k$ peut être $\mathbf{x}_k$ ou y_k. Soit $e_k = y_k - \mathbf{x}_k (\sum_U c_i \mathbf{x}_i' \mathbf{x}_i)^{-1} \sum_U c_i \mathbf{x}_i' y_i$, de sorte que $\sum_U c_i \mathbf{x}_i' e_i = 0$, et $\sum_S a_k c_k \mathbf{x}_k' e_k = \mathbf{O}_P(N/\sqrt{n})$. Nous pouvons exprimer l'erreur de t_{y_GREG} sous la forme

$$\begin{aligned}
&t_{y_\mathrm{GREG}} - T_y \\
&= \sum_{k\in S} w_k y_k - \sum_{k\in U} y_k \\
&= \sum_{k\in S} w_k e_k - \sum_{k\in U} e_k \quad \left(\text{car } \sum_{k\in S} w_k \mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k\right) \\
&= \sum_{k\in S} a_k e_k + \left(T_\mathbf{x} - \sum_{k\in S} a_k \mathbf{x}_k\right)\left(\sum_{k\in S} a_k c_k \mathbf{x}_k' \mathbf{x}_k\right)^{-1} \sum_{k\in S} a_k c_k \mathbf{x}_k' e_k \\
&\qquad - \sum_{k\in U} e_k \\
&= \sum_{k\in S} a_k e_k - \sum_{k\in U} e_k + O_P(N/n).
\end{aligned} \tag{5}$$

Il est maintenant aisé de voir que l'estimateur GREG est convergent sous randomisation; autrement dit, $p\lim_{n\to\infty} [(t_{y\_\mathrm{GREG}} - T_y)/N] = 0$. En outre, le biais relatif et l'erreur quadratique moyenne relative de l'estimateur GREG dans les conditions de randomisation sont d'ordre $1/n$. Puisque l'erreur quadratique moyenne = biais^2 + variance, nous pouvons conclure que le biais sous randomisation de l'estimateur GREG est habituellement un contributeur asymptotiquement non significatif à l'erreur quadratique moyenne de cet estimateur.

3. Redéfinition des poids de calage

Dans leur définition originale des poids de calage, Deville et Särndal (1992) posaient comme condition que l'ensemble des poids de calage, $\{w_k | k \in S\}$, minimisent une certaine fonction de distance entre les membres de l'ensemble et les poids d'échantillonnage originaux, les a_k, sous la contrainte qu'ils satisfassent l'équation de calage. Par conséquent, l'estimateur par calage, $t_{y\_\mathrm{CAL}} = \sum_S w_k y_k$, était à la fois sans biais sous le modèle donné par l'équation (2) et habituellement convergent sous randomisation.

Estevao et Särndal (2002) ont proposé d'éliminer l'exigence que les poids de calage minimisent une fonction de distance. À la place, ils ont essentiellement proposé que les w_k soient seulement obligés de satisfaire l'équation de calage et d'avoir la « forme fonctionnelle » suivante :

$$w_k = a_k(1 + \mathbf{h}_k \mathbf{q}), \qquad (6)$$

où $\mathbf{h}_k$ est un vecteur ligne de même dimension que $\mathbf{x}_k$, tel que $\sum_S a_k \mathbf{h}_k' \mathbf{x}_k$ est inversible, et $\mathbf{q}$ est un vecteur colonne de même dimension. L'équation (6) est une généralisation faible de (4), où $\mathbf{h}_k$ remplace effectivement $c_k \mathbf{x}_k$.

Il n'est pas difficile de voir que $\mathbf{q} = [(\sum_S a_j \mathbf{h}_j' \mathbf{x}_j)^{-1}]' (T_\mathbf{x} - \sum_S a_j \mathbf{x}_j)'$. En outre, sous des contraintes faibles que nous supposons vérifiées, $t_{y\_\mathrm{CAL}} = \sum_S w_k y_k = \sum_S a_k y_k + (T_\mathbf{x} - \sum_S a_j \mathbf{x}_j)(\sum_S a_j \mathbf{h}_j' \mathbf{x}_j)^{-1} \sum_S a_k \mathbf{h}_k' y_k$ est convergent sous randomisation quand t_{y_E} l'est. Il est sans biais sous le modèle prédictif linéaire donné par l'équation (2) quand $E(\varepsilon_k | \{\mathbf{x}_g, \mathbf{h}_g | g \in S\}, \{I_g | g \in U\}) = 0$ pour tout $k \in U$.

Cela suggère une définition de rechange des poids de calage : un ensemble de poids, $\{w_k | k \in S\}$, tel que

i. les w_k satisfassent l'équation de calage pour $\{\mathbf{x}_k | k \in U\}$ et,

ii. $t_{y\_\mathrm{CAL}} = \sum_S w_k y_k$ soit convergent sous randomisation quand t_{y_E} l'est sous des contraintes faibles.

Cette définition est celle que nous utiliserons. Cette définition élargie de la pondération par calage s'avérera fort utile lors du calage pour la correction de la non-réponse ou des erreurs de couverture.

Il découle de notre nouvelle définition que le calage à forme fonctionnelle d'Estevao et Särdnal est, en réalité, une forme de pondération par calage. En nous inspirant de la théorie économétrique, nous donnons aux composantes de $\mathbf{h}_k$ qui ne sont pas des combinaisons linéaires des composantes de $\mathbf{x}_k$ le nom de « variables instrumentales ».

4. Calage éventuellement non linéaire

En partant des idées de Deville et Särndal (1992), nous pouvons généraliser la forme linéaire des poids de calage donnée par l'équation (6) à

$$w_{k\_\mathrm{GEN}} = a_k f(\mathbf{h}_k \mathbf{q}^*), \qquad (7)$$

où f est une fonction monotone, dérivable deux fois avec $f(0) = 1, f'(0) = 1$ ($f'(0)$ est la dérivée première de f évaluée à 0) et $\mathbf{q}^*$ est choisi de sorte que l'équation de calage soit vérifiée. Contrairement à l'équation des poids de calage susmentionnés, l'équation de calage proprement dite, $\sum_S w_k \mathbf{x}_k = T_\mathbf{x}$, demeure linéaire. Notons que, puisque $f(0) = 1, f'(0) = 1, f(\mathbf{h}_k \mathbf{q}^*) \approx 1 + \mathbf{h}_k \mathbf{q}^*$.

Strictement parlant, nous devrions utiliser un symbole supplémentaire sur w_{k_GEN} (et plus tard sur w_{k_LIN}) pour dénoter le choix particulier de $\mathbf{h}_k$. Nous l'avons laissé tomber pour simplifier.

Une solution, $\mathbf{q}^*$, de l'équation (7) peut souvent être obtenue de façon itérative. On peut partir de $\mathbf{q}^{(0)} = \mathbf{0}$; c'est-à-dire $\sum_S w_k^{(0)} y_k$, où $w_k^{(0)} = a_k f(0)$. Pour $r = 1, 2, \ldots$, on fixe alors $\mathbf{q}^{(r)} = \mathbf{q}^{(r-1)} + \{[\sum_s f'(\mathbf{h}_k \mathbf{q}^{(r-1)}) a_k \mathbf{x}_k' \mathbf{h}_k]^{-1}\}' \times (T_\mathbf{x} - \sum_S w_k^{(r-1)} \mathbf{x}_k)'$, et $w_k^{(r)} = a_k f(\mathbf{h}_k \mathbf{q}^{(r)})$. L'itération s'arrête à r^* quand $T_\mathbf{x} = \sum_S w_k^{(r^*)} \mathbf{x}_k$, à toutes fins utiles. Cependant, il faut se souvenir qu'*il pourrait ne pas exister d'ensemble de poids pouvant être exprimé sous la forme de l'équation* (7) *tout en satisfaisant l'équation de calage.*

Soulignons que $\mathbf{q}^{(1)}$ susmentionné est égal à $\mathbf{q}$ dans $w_{k\_\mathrm{LIN}} = a_k (1 + \mathbf{h}_k \mathbf{q})$. Un développement en série de Taylor autour de zéro révèle $f(\mathbf{h}_k \mathbf{q}^{(1)}) = 1 + \mathbf{h}_k \mathbf{q}^{(1)} + O_P(1/n)$ sous des contraintes faibles, de sorte que $\sum_S w_k^{(1)} y_k = \sum_S w_{k\_\mathrm{LIN}} y_k + O_P(N/n) = T_y [1 + O_P(1/n)]$. En outre, il n'est pas difficile de voir que $w_{k\_\mathrm{GEN}} = w_{k\_\mathrm{LIN}}[1 + O_P(1/n)]$, une égalité qui s'avère utile dans l'estimation de la variance.

L'exemple le plus courant en pratique d'une fonction f non linéaire est $f(\mathbf{h}_k \mathbf{q}) = \exp(\mathbf{x}_k \mathbf{q})$, où les valeurs de chaque composante de $\mathbf{x}_k$, dénotées $x_{1k}, \ldots, x_{Pk}$, sont 0 ou 1. Cela est effectivement la forme des poids de calage sur marges (API) de Deming et Stephan (1940) calculés par ajustement proportionnel itératif. De nombreux auteurs ont constaté que la routine itérative décrite plus haut peut être utilisée même si les composantes de $\mathbf{x}_k$ ne sont pas binaires, comme elles le sont dans Deming et Stephan. Soulignons que les poids de calage par *raking généralisé* résultants sont systématiquement non négatifs.

5. Estimation de la variance

Särndal, Swensson et Wretman (1989) ont proposé cet estimateur de la variance ou de l'erreur quadratique moyenne sous randomisation à modèle *plug-in* pour t_{y_GREG} sous un plan d'échantillonnage arbitraire :

$$v_{\mathrm{SSW}} = \sum_{k\in S}\sum_{j\in S}[(\pi_{kj}-\pi_k\pi_j)/\pi_{kj}](w_k r_k)(w_j r_j). \quad (8)$$

Le terme *plug-in* vient du fait que r_k est « introduit dans » (plugged into) $v_{\mathbf{SSW}}$ à la place des $e_k = y_k - \mathbf{x}_k(\sum_U \mathbf{h}_i'\mathbf{x}_i)^{-1}\sum_U \mathbf{h}_i' y_i$ inconnus pour l'estimation de l'erreur quadratique moyenne sous randomisation.

En utilisant des arguments parallèles à ceux de Deville et Särndal (1992), $v_{\mathbf{SSW}}$ s'applique aussi de façon générale à t_{y_CAL} avec les poids de calage définis par l'équation (7) avec

$$r_k = y_k - \mathbf{x}_k\left(\sum_{j\in S} a_j\mathbf{h}_j'\mathbf{x}_j\right)^{-1}\sum_{j\in S} a_j\mathbf{h}_j'\mathbf{x}_j. \quad (9)$$

Il en est ainsi parce que $w_{k\_\mathrm{GEN}} = w_{k\_\mathrm{LIN}}[1+O_P(1/n)]$, donc $\sum_S w_{k\_\mathrm{GEN}}\, e_k = \sum_S w_{k\_\mathrm{LIN}}\, e_k + O_P(N/n) = \sum_S a_k e_k + O_P(N/n)$. La dernière étape s'appuie sur le raisonnement exprimé dans l'équation (5), avec $\mathbf{h}_j$ remplaçant les $c_j\mathbf{x}_j$.

Dans leur article, Deville et Särndal ont, en réalité, remplacé les a_j de l'équation (9) par $w_j = a_j f(\mathbf{h}_j\mathbf{q}^*)$. Une version différente est donnée dans Demanti et Rao (2004), où les a_j de l'équation sont remplacés par $a_j f'(\mathbf{h}_j\mathbf{q}^*)$. Ces auteurs soulignent dans un commentaire accompagnant cette dernière expression que les trois versions des r_k sont asymptotiquement identiques puisque $f(0) = f'(0) = 1$ et $\mathbf{q}^*$ est asymptotiquement égal à $\mathbf{0}$. Ces identités asymptotiques pourraient ne plus être vérifiées lorsque la pondération par calage est utilisée pour corriger pour la non-réponse, comme nous le verrons à la section suivante.

L'établissement de propriétés asymptotiques pour $v_{\mathbf{SSW}}$ sous échantillonnage aléatoire simple stratifié est un exercice simple. Dans le présent contexte, $v_{\mathbf{SSW}}$ se réduit à

$$v_{\mathrm{ST1}} = \sum_{\alpha=1}^{A}(n_\alpha/[n_\alpha-1])\sum_{k\in S_\alpha}(1-n_\alpha/N_\alpha) \times \left(w_k r_k - \sum_{j\in S_\alpha} w_j r_j/n_\alpha\right)^2,$$

où S_α dénote l'échantillon de n_α unités dans la strate $\alpha(\alpha = 1,\ldots,A)$, et U_α, la population de la strate contenant N_α éléments.

Dans le cas d'un échantillon à plusieurs degrés, il est logique de permettre que, dans le modèle prédictif, ε_k et ε_j soient corrélés quand k et j sont des éléments de la même UPE, mais autrement pas. Si l'on peut ignorer la correction pour population finie, la variance fondée sur un modèle d'un estimateur par calage est approximativement $V_m = \sum_{i\in S'} E_\varepsilon[(\sum_{k\in S(i)} w_k\varepsilon_k)^2]$ sous des contraintes faibles, où $S(i)$ est l'ensemble d'éléments échantillonnés dans l'UPE i, et S' est l'ensemble d'UPE sélectionnées au premier degré de l'échantillonnage.

L'estimateur de la variance qui suit, qui n'est pas strictement égal à $v_{\mathbf{SSW}}$, a souvent de bonnes propriétés sous randomisation et sous un modèle (lorsque les probabilités de sélection de premier degré sont toutes faibles) :

$$v_{\mathrm{ST2}} = \sum_{\alpha=1}^{A}(n_\alpha/[n_\alpha-1]) \times \left\{\sum_{j\in S_\alpha} - \left(\sum_{k\in S_{\alpha j}} w_k r_k\right)^2 \frac{\left(\sum_{j\in S_\alpha}\sum_{k\in S_{\alpha j}} w_k r_k\right)^2}{n_\alpha}\right\}, \quad (10)$$

où α dénote une strate d'UPE de premier degré, $n_{1\alpha}$ est le nombre d'UPE échantillonnées dans la strate α, S_α est l'ensemble d'UPE échantillonnées dans α, et $S_{\alpha j}$ est l'ensemble d'éléments sous-échantillonnés dans l'UPE j de la strate α. Le nombre de degrés d'échantillonnage peut être élevé.

Il n'est pas difficile de montrer que v_{ST2} est asymptotiquement indistinguable de l'estimateur de la variance par le jackknife :

$$v_J = \sum_{\alpha=1}^{A}([n_\alpha-1]/n_\alpha)\left\{\sum_{j\in S_\alpha}(t_{y\_\mathrm{CAL}(\alpha j)} - t_{y\_\mathrm{CAL}})^2\right\}, \quad (11)$$

où $t_{y\_\mathrm{CAL}(\alpha j)} = \sum_{k\in S} w_{k(\alpha j)} y_k$, et les *poids de calage par rééchantillonnage jackknife* sont

$$w_{k(\alpha j)} = w_k a_{k(\alpha j)}/a_k + \left(\sum_{m\in U}\mathbf{x}_m - \sum_{m\in S} w_m[a_{m(\alpha j)}/a_m]\mathbf{x}_m\right) \times \left(\sum_{m\in S} a_{m(\alpha j)}\mathbf{h}_m'\mathbf{x}_m\right)^{-1} a_{k(\alpha j)}\mathbf{h}_k', \quad (12)$$

où $a_{k(\alpha j)} = 0$ quand k est dans l'UPE j de la strate α, $a_{k(\alpha j)} = a_k$ quand k n'est pas dans la strate α du tout, et $a_{k(\alpha j)} = (n_\alpha/[n_\alpha-1])a_k$ autrement. Les $w_{k(\alpha j)}$ sont contraints de telle sorte que $\sum_{k\in S} w_{k(\alpha j)}\,\mathbf{x}_k = \sum_{k\in U}\mathbf{x}_k$ pour tout αj.

Soit $S(\alpha+)$ l'ensemble d'*éléments* dans la strate α (et non les UPE comme S_α) et $S(\alpha j)$, l'ensemble d'éléments dans l'UPE j de la strate α. Sous des contraintes faibles que nous supposons vérifiées,

$$\sum_U \mathbf{x}_m - \sum_S w_m[a_{m(\alpha j)}/a_m]\mathbf{x}_m = (n_\alpha/[n_\alpha-])\left(\sum_{S(\alpha j)} w_k\mathbf{x}_k - \sum_{S(\alpha+)} w_k\mathbf{x}_k/n_\alpha\right) = \mathbf{O}_P(N/n), \sum_S a_{m(\alpha j)}\mathbf{h}_m'\mathbf{x}_m = \mathbf{O}_P(N),$$

et

$$\sum_S a_{m(\alpha j)}\mathbf{h}'_m e_m = \mathbf{O}_P(N/\sqrt{n}).$$

Par conséquent,

$$t_{y\_\mathrm{CAL}} - t_{y\_\mathrm{CAL}} = \sum_s w_{k(\alpha j)}e_k - \sum_s w_k e_k$$
$$= (n_\alpha/[n_\alpha - 1])\left(\sum_{s(\alpha+)} w_k e_k / n_\alpha - \sum_{S(\alpha j)} w_k e_k\right)$$
$$+ O_P(N/n^{3/2}),$$

et $v_J = v_{\mathrm{ST2}}[1 + O_P(1/\sqrt{n})]$ quand $p\lim_{n\to\infty}(nv_{\mathrm{ST2}}/N^2) > 0$.

Les poids de rééchantillonnage définis par l'équation (12) ne nécessitent pas d'itération même si les poids de calage sont eux-mêmes produits de cette façon, ce qui est fort intéressant du point de vue informatique. Cela permet non seulement d'économiser du temps d'ordinateur, mais aussi d'éviter qu'une solution itérative puisse exister pour les w_k, mais non pour les poids de rééchantillonnage.

6. Non-réponse totale

6.1 Modèle de quasi-randomisation et modèle prédictif

À la présente section, nous examinons le traitement de la non-réponse totale (unité complète) en tant que phase supplémentaire de l'échantillonnage de Poisson. Il s'agit essentiellement d'un modèle de *quasi-randomisation*. Nous supposons que chaque élément k de l'échantillon original, maintenant dénoté F, a une probabilité de réponse, p_k. La probabilité que les éléments k et j répondent conjointement est $p_k p_j$, et le fait que l'élément k réponde (sachant un vecteur de covariables) est indépendant du fait qu'il soit choisi à partir de l'échantillon original.

Il est souvent possible de construire un ensemble de poids tel que l'estimateur par calage soit convergent par rapport au plan d'échantillonnage sous le modèle de quasi-randomisation. Nous recherchons ici un moyen particulier de construire ces poids. Pour cela, nous supposons que le modèle de quasi-randomisation est correct. Chaque élément est relié à un vecteur ligne de variables auxiliaires, $\mathbf{x}_k$, pour lequel $T_{\mathbf{x}} = \sum_U \mathbf{x}_j$ est connu. Enfin, nous supposons que chaque p_k est de forme :

$$p_k = 1/f(\mathbf{h}_k\boldsymbol{\phi}), \qquad (13)$$

où $\boldsymbol{\phi}$ est un vecteur colonne inconnu, $\mathbf{h}_k$ est un vecteur ligne de même dimension que $\mathbf{x}_k$, et $\sum_S a_k\mathbf{h}'_k\mathbf{x}_k/N$, où S représente maintenant le « sous-échantillon » de répondants, est inversible à la fois pour la taille de population réalisée, N, et pour la limite de probabilité.

La fonction $f(\cdot)$ de l'équation (13) est supposée être monotone et dérivable deux fois. Sa forme fonctionnelle est connue, mais la valeur du paramètre qui la régit, $\boldsymbol{\phi}$, ne l'est pas. Lorsqu'elle est introduite dans l'équation des poids de calage, $w_k = a_k f(\mathbf{h}_k\mathbf{q})$, de sorte que l'équation de calage proprement dite, $\sum_s w_k\mathbf{x}_k = T_{\mathbf{x}}$ soit vérifiée, $f(\mathbf{h}_k\mathbf{q})$ estime implicitement l'inverse de la probabilité de réponse de l'élément. Contrairement à la situation où le calage est utilisé pour la correction de l'écart de $\sum_s a_k\mathbf{x}$ par rapport à $T_{\mathbf{x}}$ dû purement à l'erreur d'échantillonnage, $f(0)$ et $f'(0)$ n'ont pas à valoir 1 et $\mathbf{h}_k\boldsymbol{\phi}$ n'a pas à valoir zéro.

Le choix le plus évident pour $\mathbf{h}_k$ lorsqu'on postule le modèle de réponse donné par l'équation (13) est $\mathbf{x}_k$ proprement dit. Dans un exemple courant de pondération par calage pour corriger pour la non-réponse, les composantes de $\mathbf{x}_k$ sont des variables indicatrices : $x_{gk} = 1$ quand k est dans le groupe g et zéro sinon. Si les groupes sont mutuellement exclusifs, la pondération par calage équivaut à une repondération dans les classes de poststratification. Voir, par exemple, Särndal, Swensson et Wretman (1992, page 585). Le modèle prédictif qui sous-tend habituellement le calage (le qualificatif « prédictif » est nécessaire pour distinguer ce modèle du modèle de quasi-randomisation) suppose que chaque élément k du groupe g, qu'il réponde ou non, a une moyenne courante : $E_\varepsilon(y_k) = \beta_g$. Le modèle de réponse quasi aléatoire est analogue : $p_k = 1/\phi_g$. Les deux modèles sont toutefois conceptuellement très différents.

Si les groupes ne sont pas mutuellement exclusifs, l'ajustement proportionnel itératif (API) est une méthode de détermination des poids de calage. Il en existe d'autres qui dépendent de la forme exacte de la fonction de réponse hypothétique $f(\cdot)$. Le modèle prédictif demeure linéaire, $E_\varepsilon(y_k) = \mathbf{x}_k\boldsymbol{\beta}$, tandis que le modèle de réponse qui donne lieu à l'API, $p_k = \exp\{-\mathbf{x}_k\boldsymbol{\phi}\}$, ne l'est pas. Berry, Flatt et Pierce (1996) donnent un exemple d'utilisation de l'API pour ajuster pour la non-réponse.

Dans de nombreuses applications de la pondération par calage, les composantes de $\mathbf{x}_k$ sont continues ou semi-continues, plutôt que dichotomiques. Dans une enquête annuelle sur les récoltes, par exemple, soit x_{1k} la quantité de maïs récoltée lors du recensement de l'agriculture précédent par l'exploitation agricole k, x_{2k}, la quantité de blé récoltée par cette exploitation, x_{3k}, la quantité de pommes de terre récoltées, et ainsi de suite. L'enquête annuelle sur les récoltes possède un modèle prédictif hypothétique pour la superficie consacrée à la culture du maïs par l'exploitation agricole k, y_{1k}, de la forme $y_{1k} = \mathbf{x}_k\beta_{1k} + \varepsilon_{1k}$. L'indice 1 désigne le maïs. Il existe d'autres valeurs d'enquête d'intérêt, comme la superficie consacrée à la culture du blé, et éventuellement des modèles prédictifs hypothétiques pour chacune.

Le modèle de réponse quasi aléatoire pour l'enquête sur les récoltes dépend des hypothèses émises au sujet de $f(\cdot)$ et de $\mathbf{h}_k$ dans l'équation (13) avec $\mathbf{h}_k$ éventuellement égal

à $\mathbf{x}_k$. Contrairement au modèle prédictif, le même modèle de quasi-randomisation hypothétique s'applique à toutes les variables de l'enquête.

Des choix prometteurs pour $f(\cdot)$ sont $\exp(\cdot)$ et $1+\exp(\cdot)$, ce dernier correspondant à un modèle de probabilité de réponse ajusté au moyen d'une fonction logistique de $\mathbf{h}_k\boldsymbol{\phi}$. Il pourrait également être raisonnable de supposer que $h_{gk} = x_{gk}^{\lambda}$ pour $\lambda < 1$. En particulier, fixer $\lambda = 0$ signifie que la probabilité que l'exploitation agricole k réponde à l'enquête annuelle sur les récoltes dépend uniquement du fait qu'elle ait déclaré du maïs, du blé ou des pommes de terre lors du recensement de l'agriculture précédent, plutôt que du volume déclaré de ces récoltes.

Dans l'exemple de l'enquête sur les récoltes, les composantes de $\mathbf{x}_k$ provenant du recensement précédent étaient les meilleurs prédicteurs disponibles des valeurs correspondantes pour l'enquête annuelle *avant* l'échantillonnage. Le fait que l'entreprise agricole k réponde à l'enquête est toutefois davantage une fonction de la superficie courante consacrée à la culture du maïs, si tant est qu'il y en ait, que d'une approximation prédéterminée de cette valeur. Par conséquent, il est tentant d'introduire les valeurs d'enquête dans $\mathbf{h}_k$, plutôt que les valeurs de recensement correspondantes. Comme nous allons voir, cette procédure pose un problème théorique.

Sachant une $f(\cdot)$, la méthode itérative décrite à la section 4 permettra souvent de découvrir un vecteur ligne $\mathbf{q}$ tel que $T_{\mathbf{x}} = \sum_s a_k f(\mathbf{h}_k\mathbf{q})\mathbf{x}_k$. Le cas échéant, l'estimation de T_y au moyen de $t_{y\_\mathrm{CAL}} = \sum_s w_k y_k$, où $w_k = a_k f(\mathbf{h}_k\mathbf{q})$, aura de bonnes propriétés sous le modèle prédictif linéaire $y_k = \mathbf{x}_k\boldsymbol{\beta} + \varepsilon_k$, où $E(\varepsilon_k|\{\mathbf{x}_g, \mathbf{h}_g, I_g | g \in U\}) = 0$ pour tout $k \in U$, $I_k = 1$ si l'élément k est présent dans l'échantillon original et qu'il répond, et 0 autrement.

L'absence de biais dans le modèle prédictif est simplement due au fait que les poids satisfont l'équation de calage. Notons toutefois que, si des composantes de $\mathbf{h}_k$ proviennent de l'enquête plutôt que de $\mathbf{x}_k$, l'hypothèse du modèle prédictif voulant que $E(\varepsilon_k|\mathbf{h}_k) = 0$ peut être problématique. Dans les conditions extrêmes, considérons le cas où une telle composante est y_k proprement dit. Habituellement, $E(\varepsilon_k|y_k)$ n'est pas nulle. Dans l'exemple de l'enquête sur les récoltes décrit plus haut, y_k peut être la superficie annuelle consacrée à la culture du maïs de l'exploitation agricole k. L'introduction de cette valeur dans $\mathbf{h}_k$ rend biaisé l'estimateur par calage connexe pour le modèle prédictif pour le maïs.

Cependant, lorsque le modèle prédictif est correct (en traitant $E(\varepsilon_k|\{\mathbf{x}_g, \mathbf{h}_g, I_g | g \in U\}) = 0$ comme une partie intégrante du modèle), la pondération par calage fondée sur tout choix de $f(\cdot)$ produira des estimateurs ayant de bonnes propriétés fondées sur le modèle prédictif. Ces estimateurs auront aussi de bonnes propriétés fondées sur le modèle de quasi-randomisation lorsque le modèle de réponse de l'équation (13) est correct pour ce choix de $f(\cdot)$. Dans un certain sens, un modèle protège contre l'échec de l'autre. Voir Kott (1994).

Comme nous l'avons souligné, le modèle prédictif a plus de chance de tenir lorsque $\mathbf{h}_g = \mathbf{x}_g$. Même ainsi, il arrive que les ε_k du modèle donné par l'équation (2) satisfassent $E(\varepsilon_k|\{\mathbf{x}_g | g \in U\}) = 0$, mais non $E(\varepsilon_k|\{\mathbf{x}_g I_g | g \in U\}) = 0$; autrement dit, le mécanisme d'échantillonnage, y compris la réponse, n'est pas ignorable en ce qui concerne le modèle prédictif.

Nous pouvons décomposer I_k en $I_{k1}I_{k2}$, où $I_{k1} = 1$ si, et uniquement si, k est présent dans l'échantillon original, et $I_{k2} = 1$ si, et uniquement si, k répond s'il est échantillonné. Le lecteur que cela intéresse peut confirmer que la pondération par calage offre une certaine protection contre le biais si le modèle prédictif de l'équation (2) est vérifié quand $E(\varepsilon_k|\{\mathbf{x}_g, \mathbf{h}_g, I_{g2} | g \in U\}) = 0$; c'est-à-dire quand le mécanisme de réponse est ignorable en ce qui concerne le modèle prédictif, mais pas nécessairement le mécanisme d'échantillonnage original.

6.2 Estimation de l'erreur quadratique moyenne sous quasi-randomisation

Que l'on puisse déclarer raisonnablement ou non que t_{y_CAL} est sans biais par rapport au modèle prédictif n'a aucun effet sur ses propriétés sous quasi-randomisation. Notons que $\mathbf{h}_k\boldsymbol{\phi}$ et $\mathbf{h}_k\mathbf{q}$ sont des valeurs scalaires et non des vecteurs. Puisque $T_{\mathbf{x}} = \sum_s a_k f(\mathbf{h}_k\mathbf{q})\mathbf{x}_i$, nos hypothèses et le théorème de la valeur moyenne ($f(\mathbf{h}_k\boldsymbol{\phi}) = f(\mathbf{h}_k\mathbf{q}) + f'(\theta_k)(\mathbf{h}_k\boldsymbol{\phi} - \mathbf{h}_k\mathbf{q})$) révèlent

$$T_{\mathbf{x}} - \sum_{k\in S} a_k f(\mathbf{h}_k\boldsymbol{\phi})\mathbf{x}_k = \sum_{k\in S} a_k [f'(\theta_k)\mathbf{h}_k(\mathbf{q} - \boldsymbol{\phi})]\mathbf{x}_k$$
$$= \mathbf{O}_P(N/\sqrt{n})$$

pour une grandeur scalaire θ_k comprise entre chaque $\mathbf{h}_k\mathbf{q}$ et $\mathbf{h}_k\boldsymbol{\phi}$. Il découle de cela que, si $\sum_s a_j f'(\mathbf{h}_j\boldsymbol{\phi})_j \mathbf{h}'_j\mathbf{x}_j / N$ est inversible à la fois pour la population réalisée N et la limite de probabilité (rappelons que f est monotone, donc que f' n'est jamais nulle), alors

$$\mathbf{q} - \boldsymbol{\phi} = \left\{\left[\sum_{j\in S} a_j f'(\mathbf{h}_j\mathbf{q})_j \mathbf{h}'_j\mathbf{x}_j\right]^{-1}\right\}' \left[T_{\mathbf{x}} - \sum_{i\in S} a_i f(\mathbf{h}_i\boldsymbol{\phi})\mathbf{x}_i\right]'$$
$$= \mathbf{O}_P(1/\sqrt{n})$$
$$= \left\{\left[\sum a_j f'(\mathbf{h}_j\boldsymbol{\phi})_j \mathbf{h}'_j\mathbf{x}_j\right]^{-1}\right\}' \left[T_{\mathbf{x}} - \sum a_i f(\mathbf{h}_i\boldsymbol{\phi})\mathbf{x}_i\right]'$$
$$+ \mathbf{O}_P(1/n).$$

L'estimateur t_{y_CAL} a une erreur de

$$t_{y\_\mathrm{CAL}} - T_y = \sum_{k\in S} a_k f(\mathbf{h}_k\mathbf{q}) y_k - \sum_{k\in U} y_k$$
$$= \sum a_k f(\mathbf{h}_k\mathbf{q}) e_k - \sum e_k,$$

où

$$e_k = y_k - \mathbf{x}_k \left(\sum\nolimits_U f'(\mathbf{h}_j\boldsymbol{\phi}) p_j \mathbf{h}_j' \mathbf{x}_j\right)^{-1} \sum\nolimits_U f'(\mathbf{h}_j\boldsymbol{\phi}) p_j \mathbf{h}_j' y_j,$$

et $p_j = 1/f(\mathbf{h}_j\boldsymbol{\phi})$. Les termes e_k sont de nouveau inconnus. Ils ont été conçus de sorte que $\sum_S a_k f'(\mathbf{h}_k\boldsymbol{\phi})\mathbf{h}_k' e_k = \mathbf{O}_P(N/\sqrt{n})$. En poursuivant :

$$\begin{aligned} & t_{y\_\mathrm{CAL}} - T_y \\ &= \sum_{k\in S} a_k f(\mathbf{h}_k\boldsymbol{\phi}) e_k - \sum_{k\in U} e_k + \sum_{k\in S} a_k \{f(\mathbf{h}_k\mathbf{q}) - f(\mathbf{h}_k\boldsymbol{\phi})\} e_k \\ &= \sum a_k f(\mathbf{h}_k\boldsymbol{\phi}) e_k - \sum e_k + \sum a_k f'(\mathbf{h}_k\boldsymbol{\phi})\mathbf{h}_k(\mathbf{q}-\boldsymbol{\phi}) e_k \\ &\quad + O_P(N/n) \\ &= \sum a_k f(\mathbf{h}_k\boldsymbol{\phi}) e_k - \sum e_k + (\mathbf{q}-\boldsymbol{\phi})' \sum a_k f'(\mathbf{h}_k\boldsymbol{\phi})\mathbf{h}_k' e_k \\ &\quad + O_P(N/n) \\ &= \sum a_k f(\mathbf{h}_k\boldsymbol{\phi}) e_k - \sum e_k + O_P(N/n). \end{aligned} \tag{14}$$

Donc, t_{y_CAL} est convergent sous quasi-randomisation sous des contraintes faibles quand $t = \sum_S a_k f(\mathbf{h}_k\boldsymbol{\phi}) y_k$ l'est.

Pour estimer l'erreur quadratique moyenne sous quasi-randomisation de t_{y_CAL} (c'est-à-dire, l'erreur quadratique moyenne de l'estimateur dans les conditions de randomisation sous le modèle de réponse), nous commençons par noter que la probabilité que les éléments k et $j, k \neq j$, soient tous deux compris dans le sous-échantillon de répondants est $\pi_{kj}^* = \pi_{kj} p_k p_j$. Soit $\pi_k^* = \pi_k p_k$, et rappelons que $a_k = 1/\pi_k$ et $1/p_k = f(\mathbf{h}_k\boldsymbol{\phi})$. Partant de l'équation (14), nous voyons que l'erreur quadratique moyenne sous quasi-randomisation de t_{y_CAL} est approximativement

$$\begin{aligned} & E_1[(t_{y\_\mathrm{CAL}} - T_y)^2] \\ &\approx \sum_{k\in U} \sum_{j\in U} (\pi_{kj}^* - \pi_k^*\pi_j^*)(e_k/\pi_k^*)(e_j/\pi_j^*) \\ &= \sum_{k\in U} (1-\pi_k^*) e_k^2/\pi_k^* \\ &\quad + \sum_{\substack{k\in U \\ k\neq j}} \sum_{j\in U} (\pi_{kj} - \pi_k\pi_j)(e_k/\pi_k)(e_j/\pi_j). \end{aligned} \tag{15}$$

Si l'échantillon original est de type Poisson, alors $v_m = \sum_s (w_k^2 - w_k) r_k^2$ avec

$$r_k = y_k - \mathbf{x}_k \left[\sum_{j\in S} a_j f'(\mathbf{h}_j\mathbf{q})\mathbf{h}_j'\mathbf{x}_j\right]^{-1} \sum_{j\in S} a_j f'(\mathbf{h}_j\mathbf{q})\mathbf{h}_j' y_j, \tag{16}$$

sert à la fois d'estimateur raisonnable de la variance fondée sur le modèle prédictif et de l'erreur quadratique moyenne fondée sur le modèle de quasi-randomisation sous des contraintes faibles, puisque $w_k \approx 1/\pi_k^*$ et $r_k \approx e_k$. Un proche parent du résidu d'échantillon non intuitif dans l'équation (16) est donné dans Folsom et Singh (2000). Voir Kott (2004a) pour une discussion plus approfondie de v_m dans un contexte d'échantillonnage pur.

Pour un plan de sondage général, nous pouvons nous approcher d'un bon estimateur de la variance/erreur quadratique moyenne avec

$$\begin{aligned} v_{\mathrm{com}} &= \sum_{k\in S} (w_k^2 - w_k) r_k^2 \\ &\quad + \sum_{\substack{k\in S \\ k\neq j}} \sum_{j\in S} [(\pi_{kj} - \pi_k\pi_j)/\pi_{kj}](w_k r_k)(w_j r_j). \end{aligned} \tag{17}$$

Le deuxième membre de l'équation (17) estime le deuxième membre de l'équation (15) avec r_k remplaçant e_k. Notons que $\sum_U (1-\pi_k^*) e_k^2/\pi_k^*$ dans l'équation (15) est estimé par $\sum_S (w_k^2 - w_k) r_k^2$ plutôt que par $\sum_S w_k^2 (1-\pi_k^*) r_k^2$, ce qui rendrait v_{com} plus convergent avec v_{SSW} de l'équation (8). Cette substitution donne un estimateur de variance ayant de bonnes propriétés basées sur le modèle prédictif quand les ε_k sont non corrélées, et $\sigma_k^2 = \mathbf{x}_k\boldsymbol{\zeta}$, pour un certain $\boldsymbol{\zeta}$. Elle peut être faite même en l'absence de non-réponse.

Lorsque l'échantillon réel comprend plusieurs degrés et que les probabilités de sélection de premier degré sont suffisamment faibles pour être ignorées, v_{ST2} de l'équation (10) peut être utilisée comme estimateur de la variance/erreur quadratique moyenne avec r_k de nouveau défini par l'équation (16).

Quand f est linéaire, $f'(\theta) = 1$, et les r_k de l'équation (16) sont calculés comme s'il n'y avait aucune non-réponse. Cela est également vrai pour l'estimateur de la variance/erreur quadratique moyenne v_{ST2}. Malheureusement, cette f correspond à une fonction de probabilité de réponse de forme peu maniable : $p_k = 1/\mathbf{h}_k\boldsymbol{\phi}$. Fuller, Loughin et Baker (1994) ont fait ces constatations pour le cas où $\mathbf{h}_k = c_k\mathbf{x}_k$.

Dans l'équation (11), le jackknife, v_J, peut être calculé à l'aide des poids de rééchantillonnage jackknife :

$$\begin{aligned} w_{k(\alpha j)} &= w_k a_{k(\alpha j)}/a_k + \left(\sum_{m\in U} \mathbf{x}_m - \sum_{m\in S} w_m [a_{m(\alpha j)}/a_m]\mathbf{x}_m\right) \\ &\quad \times \left(\sum_{m\in S} a_{m(\alpha j)} f'(\mathbf{h}_m\mathbf{q})\mathbf{h}_m'\mathbf{x}_m\right)^{-1} a_{k(\alpha j)} f'(\mathbf{h}_k\mathbf{q})\mathbf{h}_k', \end{aligned} \tag{18}$$

ce qui est une généralisation évidente des poids de rééchantillonnage jackknife de l'équation (12). De nouveau, si $f'(\theta) = 1$, v_J peut être calculé comme s'il n'y avait pas de non-réponse.

7. Modélisation de la couverture

Folsom et Singh (2000) ont fait remarquer que le traitement de la non-réponse au moyen de la pondération par calage peut aussi être appliqué pour corriger pour le sous-dénombrement. Dans le contexte, la phase quasi

aléatoire, de l'échantillonnage a lieu conceptuellement avant que l'échantillon réel soit tiré. Nous supposons que la population associée à la base de sondage est un échantillon de Poisson provenant d'une population complète hypothétique pour laquelle le vecteur $T_{\mathbf{x}}$ doit être connu. La population de la base de sondage devient F, tandis que la population complète hypothétique est U. Nous supposons que la probabilité que l'élément $k \in U$ soit dans F est modélisé correctement par l'équation (13). Si la première (de U à F) et la deuxième (de F à S) phases d'échantillonnage sont indépendantes, alors toute la théorie élaborée pour l'utilisation de la pondération par calage en vue de traiter la non-réponse peut être transposée au traitement du sous-dénombrement.

Il convient de souligner que la correction de l'erreur de couverture par calage est une extension de la pratique bien connue de correction par poststratification souvent utilisée dans les enquêtes téléphoniques. Comme dans le cas particulier de la poststratification, il faut utiliser comme cible de calage pour U des quantités que l'on peut supposer dépourvues d'erreur ou ayant une erreur quadratique moyenne très faible comparativement aux estimateurs par calage proprement dit.

Folsom et Singh ont fait remarquer que le surdénombrement (dénombrement multiple) ou une combinaison de sous- et de surdénombrement peuvent être traités en suivant leur méthode. La définition de p_k dans l'équation (13) devient le nombre prévu de fois que k est présent dans la base de sondage, nombre qui peut maintenant être supérieur à 1 à cause du dénombrement multiple éventuel.

Folsom et Singh proposent en outre de donner à $f(\cdot)$ la forme flexible :

$$f(\mathbf{x}_k\boldsymbol{\phi}) = \frac{U(C-L)\exp(\mathbf{x}_k\boldsymbol{\phi}) + L(U-C)}{(U-C) + (C-L)\exp(\mathbf{x}_k\boldsymbol{\phi})}, \qquad (19)$$

où $L \geq 0, 1 < U \leq \infty,$ et $L < C \leq U$ sont des constantes prédéterminées. Ils donnent à cette expression le nom de « modèle exponentiel général » ou « MEG ». Observons que, si $C = 1, U = \infty,$ et $L = 0, p_k = 1/f(\mathbf{x}_k\boldsymbol{\phi}) = \exp(-\mathbf{x}_k\boldsymbol{\phi})$. Similairement, si $C = 2,$ $U = \infty,$ et $L = 1,$ $p_k = [1 + \exp(\mathbf{x}_k\boldsymbol{\phi})]^{-1}$; autrement dit, la probabilité de couverture (ou de réponse) est logistique. Les valeurs L et U servent de bornes sur l'*ajustement par calage*, $f(\cdot),$ tandis que $C = f(0)$ est effectivement son centre.

Les auteurs ont rendu l'ajustement par calage dans le MEG encore plus souple en postulant trois classes d'unités d'échantillonnage, chacune ayant son propre ensemble de valeurs U, C et L. Ils ont proposé son utilisation pour la correction de l'erreur de couverture ainsi que de la non-réponse totale.

8. Un petit exemple empirique

Puisque les poids de rééchantillonnage jackknife exprimés par l'équation (18) sont nouveaux, il est prudent de chercher à savoir s'ils fonctionnent effectivement avec des données réelles. Pour ce faire, nous avons pris les données MU281 de Särndal, Swensson et Wretman (1992) et les avons répétées 20 fois (de sorte que $N = 5\,620$). Par échantillonnage aléatoire simple stratifié, nous avons sélectionné 16 unités dans chacune des huit strates de taille inégale. La variable RMT85 a servi de y_k et la variable P75, de x_k dans $\mathbf{x}_k = (1, x_k)$. À chacune des 128 unités échantillonnées, nous avons attribué une probabilité d'être présente dans le sous-échantillon de répondants, S, qui diminuait avec la taille de x_k; en particulier, $p_k = \exp(-0{,}35\, x_k / M_x),$ où M_x était la moyenne de population de x_k. Dans les 1 600 simulations, la taille de S variait de 78 à 110, avec une moyenne d'environ 93,8.

Le total T_y a été estimé de deux façons, avec $t_{y\_\mathrm{LIN}} = \sum_S a_k(1 + \mathbf{x}_k\mathbf{q})y_k$ et avec $t_{y\_\mathrm{EXP}} = \sum_S a_k \exp(\mathbf{x}_k\mathbf{q}^{(\mathrm{exp})})y_k,$ où $\mathbf{q}$ et $\mathbf{q}^{(\mathrm{exp})}$ étaient, respectivement, sélectionnés de sorte que l'équation de calage soit vérifiée. Le premier était un estimateur GREG, tandis que le second était un estimateur par ajustement proportionnel itératif généralisé. Les deux estimateurs étaient sans biais sous le modèle prédictif sous-entendu $(y_k = \mathbf{x}_k\boldsymbol{\beta} + \varepsilon_k),$ mais seul t_{y_EXP} était convergent dans des conditions de randomisation sous le modèle de réponse correcte. L'estimateur GREG supposait implicitement que $p_k = 1/(\phi_0^{(\mathrm{LIN})} + \phi_1^{(\mathrm{LIN})}\mathbf{x}_k)$ pour $\phi_0^{(\mathrm{LIN})}$ et $\phi_1^{(\mathrm{LIN})}$ inconnus.

La petite taille de l'échantillon comparativement à la population de chaque strate a permis d'ignorer la correction pour population finie dans l'estimation de la variance/erreur quadratique moyenne (appelée dans la suite « estimation de la variance »). Nous avons estimé les variances en utilisant *i*) l'estimateur par linéarisation, $v_{\mathrm{ST2}},$ dans l'équation (10) avec r_k défini par l'équation (16) et *ii*) le jackknife proposé, $v_J,$ dans l'équation (11) avec les poids de rééchantillonnage définis par l'équation (18). Pour rendre le calcul du jackknife plus simple, les 16 sous-échantillons dans chaque strate ont été attribués aléatoirement à l'une de quatre grappes, de sorte que 32 répliques jackknife seulement ont dû être calculées.

Aux fins de comparaison, une meilleure version de l'estimateur de variance par linéarisation, dénotée $v_{\mathrm{ST2}(e)},$ a également été calculée avec r_k remplacé par $e_k = y_k - \mathbf{x}_k(\sum_U f'(\mathbf{x}_j\boldsymbol{\phi})\ p_j\mathbf{x}'_j\mathbf{x}_j)^{-1}\sum_U f'(\mathbf{x}_j\boldsymbol{\phi})p_j\mathbf{x}'_j y_j,$ où $\boldsymbol{\phi}$ et p_j étaient connus. En pratique, e_k est rarement connu, mais le calcul de $v_{\mathrm{ST2}(e)}$ est utile ici pour les comparaisons.

Il convient de souligner que les calculs de r_k et e_k différaient légèrement selon que l'on voulait calculer l'estimateur de la variance pour t_{y_LIN} ou pour t_{y_EXP}.

Pour t_{y_LIN}, $f'(\mathbf{x}_j \boldsymbol{\phi}) = f'(\mathbf{x}_j \mathbf{q}) = 1$; pour t_{y_EXP}, $f'(\mathbf{x}_j \mathbf{q}^{(\exp)}) = \exp(\mathbf{x}_j \mathbf{q}^{(\exp)})$ et $f'(\mathbf{x}_j \boldsymbol{\phi}) = 1/p_j$.

Le tableau 1 donne les moyennes empiriques (la moyenne sur les 1 600 simulations) des deux estimateurs pour T_y normalisés de sorte que $T_y = 100$. Bien que tous deux soient pratiquement sans biais, t_{y_LIN} diffère significativement de 100 au seuil de signification de 0,05; ce n'est pas le cas pour t_{y_EXP}. Cela n'est pas étonnant, parce que seul le dernier est fondé sur le modèle de réponse correcte.

Pour chaque estimateur, les estimateurs de variance et les erreurs quadratiques moyennes empiriques ont été normalisés de sorte que les moyennes empiriques des $v_{\mathrm{ST2}(e)}$ respectifs soient égales à 100. Aucun $v_{\mathrm{ST2}(e)}$ n'avait une moyenne empirique significativement différente de l'erreur quadratique moyenne empirique (EQME) de l'estimateur associé. Ce résultat était un peu décevant. Il semble que, bien que t_{y_LIN} ait un biais empirique significatif, celui-ci était une composante tellement faible de l'erreur quadratique moyenne de l'estimateur que la différence entre l'EQME de cet estimateur et la moyenne empirique de $v_{\mathrm{ST2}(e)}$ n'était pas significative.

Les $v_{\mathrm{ST2}(e)}$ ont été choisis comme valeurs de référence pour le tableau plutôt que les erreurs quadratiques moyennes empiriques, parce que chaque $v_{\mathrm{ST2}(e)}$ avait environ la moitié de l'erreur-type empirique de l'EQME correspondante (qui, elle-même, correspondait à la moyenne des 1 600 carrés des écarts) et était corrélée plus fortement avec les estimateurs de variance. Les valeurs t pour cette partie du tableau ont également été calculées par rapport aux $v_{\mathrm{ST2}(e)}$.

Les deux estimateurs de variance par linéarisation avaient un biais par défaut étonnamment grand. Il semble que les w_{k_LIN} et w_{k_EXP} inhabituellement grands avaient tendance à rendre les r_k associés appréciablement plus faibles que les e_k en valeur absolue. Les problèmes associés à des valeurs inhabituellement grandes de w_{k_LIN} et w_{k_EXP} paraissent être plus atténués dans le cas des estimateurs jackknife.

Pour accélérer l'asymptotique des estimateurs de variance par linéarisation (c'est-à-dire, réduire l'écart entre r_k et e_k), nous avons calculé un ajustement ponctuel de v_{ST2} en remplaçant chaque r_k par $r_{k(\mathrm{ajusté})} = r_k / \omega_k$, où $\omega_k^2 = 1 - \mathbf{x}_k (\sum_S a_j f'(\mathbf{x}_j \mathbf{q}) \mathbf{x}_j' \mathbf{x}_j)^{-1} a_k f'(\mathbf{x}_k \mathbf{q}) \mathbf{x}_k' = 1 + O_P(1/n)$. Notons que, sous le modèle prédictif avec les ε_k non corrélés et $E(\varepsilon_k^2) = \sigma_k^2$, $E(r_{k(\mathrm{ajusté})}^2) \approx \sigma_k^2$. La quasi-égalité est exacte lorsque tous les $a_j f'(\mathbf{x}_j \mathbf{q})$ et σ_j, sont, respectivement, égaux.

Le v_{ST2} ajusté pour t_{y_LIN} ainsi que t_{y_EXP} demeurait entaché d'un biais par défaut, tandis que le v_J présentait un biais par excès d'une valeur légèrement plus faible. Bien que ces biais soient significatifs, ils étaient raisonnablement petits (de 4,5 à 11,2 %) et donnent à penser que les estimateurs de variance étaient peut-être effectivement asymptotiquement sans biais, comme nous l'avons démontré théoriquement aux sections précédentes.

Tableau 1
Moyennes empiriques des estimateurs basées sur 1 600 simulations*

Moyenne empirique (erreur-type)		Valeur t (test de signification bilatéral)	
Estimateurs pour $T_y (T_y = 100)$			
t_{y_LIN}	99,84 (0,06)	−2,79 (0,02)	différent de
t_{y_EXP}	100,04 (0,06)	0,58 (0,56)	T_y
Estimateurs de variance pour $t_{y\_\mathrm{LIN}} (E_{\mathrm{EMP}}(v_{\mathrm{ST2}(e)}) = 100)$			
v_{ST2}	83,59 (1,53)	−19,96 (< 0,0001)	différent de
$v_{\mathrm{ST2(ajusté)}}$	95,53 (1,80)	−6,09 (< 0,0001)	$v_{\mathrm{ST2}(e)}$
v_J	104,69 (2,28)	3,60 (0,0003)	
EQME	99,35 –	−0,18 (0,85)	
Estimateurs de variance pour $t_{y\_\mathrm{EXP}} (E_{\mathrm{EMP}}(v_{\mathrm{ST2}(e)}) = 100)$			
v_{ST2}	73,12 (1,54)	−18,22 (< 0,0001)	différent de
$v_{\mathrm{ST2(ajusté)}}$	88,79 (1,98)	−8,57 (< 0,0001)	$v_{\mathrm{ST2}(e)}$
v_J	107,00 (2,73)	4,09 (< 0,0001)	
EQME	101,21 –	0,33 (0,74)	
Autres statistiques			
relvar $(v_{\mathrm{ST2}(e)[\mathrm{LIN}]})$	0,051 –	–	
relvar $(v_{\mathrm{ST2}(e)[\mathrm{EXP}]})$	0,059 –	–	
$\dfrac{(v_{\mathrm{ST2}(e)[\mathrm{LIN}]} - v_{\mathrm{ST2}(e)[\mathrm{EXP}]})}{(E_{\mathrm{EMP}}(v_{\mathrm{ST2}(e)[\mathrm{EXP}]})}$	−0,1340 (0,010)	−13,87 (< 0,0001)	

* Dans quatre simulations supplémentaires, la convergence n'a pas été atteinte en dix itérations pour t_{y_EXP}. Ces simulations ont été exclues de l'analyse.

Lors de l'utilisation de $v_{\mathrm{ST2}(e)}$ comme approximation efficace de l'EQME, l'erreur quadratique moyenne empirique de t_{y_EXP}, qui intégrait le modèle de réponse correcte, était plus de 13 % plus importante que celle de t_{y_LIN}, qui n'intégrait pas ce modèle. Toutefois, il ne convient pas de faire de grandes généralisations à partir d'un ensemble de données comportant deux variables de calage seulement. Voir Crouse et Kott (2004) pour un ensemble différent de résultats.

Qu'il soit préférable ou non d'intégrer le modèle de réponse correcte dans l'estimateur de calage, si on le fait, alors les estimateurs de variance discutés à la section précédente, peut-être avec l'estimateur par linéarisation corrigé comme il est suggéré à la présente section, semblent utilisables.

Un deuxième ensemble de 1 600 simulations (non présentées) ont été exécutées en utilisant la même population et un plan d'échantillonnage stratifié, mais en donnant à chaque élément échantillonné 70 % de chances de faire partie de l'échantillon de répondants (la taille moyenne de l'échantillon de répondants était d'environ 89,8). Dans cet ensemble de simulations, les deux estimateurs de T_y sont convergents par rapport au plan d'échantillonnage (randomisation) sous le modèle de réponse. Par conséquent, il n'est pas étonnant que les moyennes empiriques de t_{y_LIN} et de t_{y_EXP} soient presque identiques (écart d'au plus 0,01 %), comme le sont leurs erreurs quadratiques moyennes empiriques (écart d'au plus 1 %). Les moyennes empiriques de chaque paire d'estimateurs de variance (par exemple $\mathrm{var}_{\mathrm{ST2}}$ pour t_{y_LIN} et t_{y_EXP}) étaient aussi très proches (écart d'au plus 1 %). Le biais relatif de l'estimateur v_{ST2} ajusté (comparativement à $\mathrm{var}_{\mathrm{ST2}(e)}$) était de –1,3 % lors de l'estimation de la variance de t_{y_LIN} et de –2,2 % lors de l'estimation de la variance de t_{y_EXP}. Le biais relatif des variances par linéarisation non corrigées était de –9,0 % et de –10,3 %, respectivement. Le biais relatif des deux estimateurs jackknife était de 3,6 %.

9. Discussion

9.1 Estimation explicite d'un modèle de réponse

En présence de non-réponse totale, nombreux sont ceux qui ont essayé d'estimer les probabilités de réponse individuelles, $p_k = 1/f(\mathbf{h}_k\boldsymbol{\phi})$, directement. Cette méthode requiert de l'information sur $\mathbf{h}_k$ pour chaque unité échantillonnée, qu'elle réponde à l'enquête ou non, mais $\mathbf{h}_k$ ne doit pas avoir la même dimension que $\mathbf{x}_k$. La méthode d'ajustement directe n'est généralement pas disponible pour le traitement des erreurs de couverture.

Fuller (2002) a souligné qu'un terme supplémentaire peut figurer dans l'erreur quadratique moyenne sous quasi-randomisation de $t_{y\_\mathrm{GREG}} = \sum_S a_k^* y_k + (T_{\mathbf{x}} - \sum_S a_j^* \mathbf{x}_j)\times (\sum_S c_j a_j^* \mathbf{x}_j' \mathbf{x}_j)^{-1} \sum_S c_k a_k^* \mathbf{x}_k' \mathbf{x}_k$, où S est le sous-échantillon de répondants, $a_k^* = a_k[1 + f(\mathbf{h}_k\mathbf{q})]$, et $\mathbf{q}$ est un estimateur direct convergent pour le paramètre du modèle de quasi-randomisation, $\boldsymbol{\phi}$. Cela ne sous-entend pas que l'estimation directe du modèle de réponse fondée sur une $f(\cdot)$ et un $\mathbf{h}_k$ donnés est moins efficace que le calage analogue lorsque $\mathbf{h}_k$ a la même dimension que $\mathbf{x}_k$. Voir Kim (2004) pour une suggestion du contraire. Néanmoins, la commodité qu'offre l'intégration de la correction pour la non-réponse dans le calage est séduisante, lorsque les estimations de variance doivent être produites.

Un compromis raisonnable consiste à choisir la forme de $f(\cdot)$ et de $\mathbf{h}_k$ en modélisant le comportement de réponse de l'échantillon complet, puis à estimer le paramètre de $f(\cdot)$ implicitement par calage. Ce compromis permet aussi de contourner une faiblesse frappante de l'utilisation de la pondération par calage pour corriger de la non-réponse (ainsi que pour les erreurs de couverture). Les choix de $f(\cdot)$ et $\mathbf{h}_k$ sont motivés principalement par la vraisemblance et la commodité et non par une analyse statistique des données.

9.2 Groupes à homogénéité de réponse

Afin de contrôler l'importance de la repondération due à la non-réponse, Little (1986) a recommandé que l'on estime $\mathbf{q}$ explicitement, puis qu'on divise l'échantillon en C groupes mutuellement exclusifs en se fondant sur les tailles des valeurs ajustées de $f(\mathbf{h}_k\mathbf{q})$. On calcule ensuite le poids corrigé pour chaque élément k dans le groupe c comme dans le cas de la poststratification :

$$w_{k\_\mathrm{ADJ}} = \left(\sum_{F(c)} w_g / \sum_{S(c)} w_g\right) w_k,$$

où $F_{(c)}$ est la partie de l'échantillon original comprise dans le groupe c, $S(c)$ est le sous-échantillon de $F(c)$ qui répond, et w_k est le poids d'échantillonnage attribué à l'élément k après échantillonnage, mais avant sous-échantillonnage quasi aléatoire. Cette approche suppose que chaque élément d'un groupe a (approximativement) la même probabilité de réponse, d'où l'expression « groupe à homogénéité de réponse ».

Un autre moyen d'intégrer les valeurs ajustées de $f(\mathbf{h}_k\mathbf{q})$ dans l'estimation fondé sur la méthodologie développée dans le texte est décrit ci-après. Répartir les valeurs ajustées en P groupes d'après leur taille, où P est de nouveau la dimension de $\mathbf{x}_k$, et soit $\mathbf{d}_k$, un vecteur ligne de variables indicatrices pour les P cellules. En fixant chaque $w_k = a_k[1 + (T_{\mathbf{x}} - \sum_S a_j \mathbf{x}_j)\ (\sum_S a_j \mathbf{d}_j' \mathbf{x}_j)^{-1} \mathbf{d}_k']$, on calcule un ensemble de poids pour le sous-échantillon de répondants qui, contrairement à $\{w_{k\_\mathrm{ADJ}}\}$ ci-dessus, satisfait l'équation de calage pour l'échantillon de répondants. Étant donné la nature de $\mathbf{d}_k$, cette méthode linéaire produit le même ensemble de poids de calage que celui que donnerait l'ajustement de $w_k = a_k \exp(\mathbf{d}_k\mathbf{f})$ – si les deux produisent

un ensemble de poids. Notons que, puisque les poids de calage peuvent être négatifs dans le cas de la méthode linéaire, celle-ci pourrait produire un ensemble que la méthode par ajustement proportionnel itératif généralisé ne pourrait pas produire. La méthode linéaire rééchelonne effectivement les a_k, c'est-à-dire la valeur de chaque élément dans le même groupe, d'une quantité fixe. Donc, elle pourrait ne pas produire des poids étonnamment petits ou étonnamment grands lorsque la dimension de $\mathbf{x}_k$ est faible comparativement à la taille d'échantillon.

9.3 Calage de l'échantillon et calage pour la non-réponse

À la section précédente, nous avons indiqué qu'il est possible que les composantes de $\mathbf{h}_k$ dans l'équation (13), c'est-à-dire le modèle de réponse quasi aléatoire, soient inconnues avant le recensement. Lorsqu'on utilise un tel $\mathbf{h}_k$ dans le calage, il n'est peut-être plus raisonnable d'affirmer que l'estimateur t_{y_CAL} résultant est sans biais par rapport au modèle prédictif. Cela est particulièrement ennuyeux lorsque la non-réponse est modérée, comparativement à la taille de l'échantillon. Une idée intéressante consiste à faire le calage en deux phases. La première phase, le calage de l'échantillon, consiste à corriger pour la différence entre $T_{\mathbf{x}}$ et $\sum_F a_k \mathbf{x}_k$, et ne comprendrait aucune composante de $\mathbf{h}_k$ inconnue au moment de l'échantillonnage. La deuxième phase, le calage pour la non-réponse, corrige pour la différence entre $\sum_F a_k \mathbf{x}_k$ et $\sum_S a_k \mathbf{x}_k$ et comprendrait uniquement les variables composantes disponibles après que le sous-échantillon de répondants soit recensé.

Une analyse plus approfondie de cette idée doit être reportée à une autre occasion.

9.4 Travaux avec le NASS

Le National Agricultural Statistics Service (NASS) a utilisé des variantes de l'approche de Fuller et coll. (1994) pour traiter le sous-dénombrement au Recensement de l'agriculture de 2002 (voir Fetter et Kott 2003) et pour la correction d'une enquête sur l'économie agricole avec non-réponse importante, de façon à faire concorder les totaux à ceux d'enquêtes plus fiables (voir Crouse et Kott 2004). Dans cette approche, $f(\cdot)$ est de la forme :

$$f(\mathbf{x}_k\boldsymbol{\phi}) = \begin{cases} L & \text{si} \quad \mathbf{x}_k\boldsymbol{\phi} < L \\ \mathbf{x}_k\boldsymbol{\phi} & \text{si} \quad \mathrm{L} \le \mathbf{x}_k\boldsymbol{\phi} \le U \\ U & \text{si} \quad \mathbf{x}_k\boldsymbol{\phi} > U, \end{cases} \tag{20}$$

qui tronque le calage linéaire à des valeurs préétablies, L et U, pour contrôler l'importance de l'ajustement des poids. Notons que, quand $f(\cdot) = U$ ou L, $f'(\cdot) = 0$. Contrairement à l'ajustement par calage de l'équation (19), $f(\cdot)$ de l'équation (20) n'est pas dérivable deux fois aux valeurs L ou U. Cela ne cause pas de problème en pratique.

La justification originale du calage offerte par l'organisme dans ce contexte était fondée sur la modélisation prédictive. L'équation (20) est simple à appliquer et semble produire des poids qui se situent dans une fourchette acceptable plus fréquemment que d'autres solutions facilement disponibles.

Le NASS étudie les questions suivantes : quelle est la sensibilité de t_{y_CAL} au choix de $f(\cdot)$ en pratique? Un choix différent pour $f(\cdot)$ produirait-il un biais plus faible et, le cas échéant, la réduction du biais absolu se traduirait-elle par une erreur quadratique moyenne plus faible? Quel serait l'effet du remplacement de certaines composantes du vecteur de variables et de calage par un meilleur prédicteur de la non-réponse ou du sous-dénombrement?

Bibliographie

Berry, C.C., Flatt, S.W. et Pierce, J.P. (1996). Correcting unit nonresponse via nonresponse modeling and raking in the California Tobacco Survey. *Journal of Official Statistics*, 12, 349-363.

Crouse, C., et Kott, P.S. (2004). Evaluation alternative calibration schemes for an economic survey with large nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Deming, W.E., et Stephan, F.F. (1940). On a least squares adjustment of a sample frequency table when the expected marginal total are known. *Annals of Mathematical Statistics*, 11, 427-444.

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

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

Fetter, M.J., et Kott., P.S. (2003). Developing a coverage adjustment strategy for the 2002 Census of Agriculture. Présenté à 2003 Federal Committee on Statistical Methodology Research Conference, http://www.fcsm.gov/03papers/fetter_kott.pdf.

Folsom., R.E., et Singh, A.C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. *Proceedings of the 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 1987–88. *Techniques d'enquête*, 20, 79-89.

Kim, J.K. (2004). Efficient nonresponse weighting adjustment using estimated response probability. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Kott, P.S. (1990). The design consistent regression estimator and its conditional variance. *Journal of Statistical Planning and Inference*, 24, 287-296.

Kott, P.S. (1994). A note on handling nonresponse in surveys. *Journal of the American Statistical Association*, 89, 693-696.

Kott, P.S. (2004a). Randomization-assisted model-based survey sampling. *Journal of Statistical Planning and Inference*, 48, 263-277.

Kott, P.S. (2004b). Commentaire. *Techniques d'enquête*, 30, 28-29.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de Statistique*, 54, 139-157.

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

Särndal, C.-E., Swensson, B. et Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of a finite population total. *Biometrika*, 76, 527-537.

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