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

Article

Estimation de la variance sous imputation composite : méthodologie programmée dans le SEVANI

par Jean-François Beaumont et Joël Bissonnette

Décembre 2011

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2011
Vol. 37, N° 2, pp. 183-192
Statistique Canada, N° 12-001-X au catalogue

Estimation de la variance sous imputation composite : méthodologie programmée dans le SEVANI

Jean-François Beaumont et Joël Bissonnette [1]

Résumé

L'imputation composite est fréquemment employée dans les enquêtes auprès des entreprises. Le terme « composite » signifie que l'on utilise plus d'une méthode d'imputation pour remplacer les valeurs manquantes d'une variable d'intérêt. La littérature consacrée à l'estimation de la variance sous imputation composite est peu abondante. Afin de surmonter ce problème, nous examinons une extension de la méthodologie élaborée par Särndal (1992). Cette extension est de nature assez générale et est facile à mettre en oeuvre, à condition d'utiliser des méthodes d'imputation linéaires pour remplacer les valeurs manquantes. Cette catégorie de méthodes comprend l'imputation par régression linéaire, l'imputation par donneur et l'imputation par valeur auxiliaire, parfois appelée imputation « cold-deck » ou imputation par substitution. Elle englobe donc les méthodes les plus couramment utilisées par les organismes statistiques nationaux pour imputer les valeurs manquantes. Notre méthodologie a été intégrée au Système d'estimation de la variance due à la non-réponse et à l'imputation (SEVANI), mis au point à Statistique Canada. Une étude par simulation est effectuée pour en évaluer les propriétés.

Mots clés : Imputation par valeur auxiliaire ; imputation composite ; imputation par donneur ; modèle d'imputation ; imputation linéaire ; imputation par régression ; SEVANI.

1. Introduction

L'imputation composite est fréquemment employée dans les enquêtes auprès des entreprises. Le terme « composite » signifie que l'on utilise plus d'une méthode d'imputation pour remplacer les valeurs manquantes d'une variable d'intérêt. Le choix de la méthode dépendra de la disponibilité de variables auxiliaires. Par exemple, on pourra utiliser l'imputation par le ratio pour imputer une valeur manquante si l'on dispose d'une variable auxiliaire ; sinon, on pourra opter pour l'imputation par la moyenne.

Il y a une abondante littérature consacrée au problème que pose l'estimation de la variance lorsque l'on utilise une méthode d'imputation unique, dont les excellents comptes rendus de Lee, Rancourt et Särndal (2001) ainsi que de Haziza (2009). Par contre, on recense peu de travaux sur l'estimation de la variance lorsque l'on a recours à l'imputation composite, même si ce type d'imputation est souvent utilisé dans la pratique. Mentionnons notamment Rancourt, Lee et Särndal (1993), qui ont proposé et évalué empiriquement un estimateur de variance jackknife. Sitter et Rao (1997) ont poussé plus loin l'étude théorique et ont obtenu des estimateurs de variance jackknife et par linéarisation convergents par rapport au plan de sondage. Dans les deux articles, les auteurs ont considéré deux méthodes d'imputation, dont l'imputation par le ratio, dans des conditions d'échantillonnage aléatoire simple où l'on supposait que la non-réponse était uniforme. Puis, Felx et Rancourt (2001) ont étendu la méthode générale proposée par Särndal (1992) et par Deville et Särndal (1994) à l'imputation composite au moyen d'hypothèses simplificatrices. Enfin, pour tenir compte de l'imputation composite, Shao et Steel (1999) ont élaboré une approche inverse générale d'estimation de la variance qui est fort intéressante (voir également Kim et Rao 2009). Shao et Steel (1999) ont fait valoir que leur approche inverse donnait lieu à des calculs moins compliqués que celle de Deville et Särndal (1994). Nous ne partageons pas entièrement ce point de vue. Nos résultats montrent que, de façon générale, l'application que nous faisons de l'approche de Särndal donne en fait lieu à des calculs plus simples que ce n'est le cas avec celle de Shao et Steel. Cela dit, l'approche inverse pourrait devenir beaucoup plus attrayante lorsque la fraction de sondage est négligeable et que l'on opte pour une technique d'estimation de la variance par réplication (on trouvera à la section 7 des commentaires plus détaillés à ce sujet).

Nous utilisons comme point de départ la méthode proposée par Särndal (1992). Elle requiert un modèle d'imputation valide, c'est-à-dire un modèle pour la variable qui est imputée. À première vue, l'extension de cette méthode à l'imputation composite semble assez fastidieuse, comme l'ont souligné Shao et Steel (1999), jusqu'à ce que l'on remarque que la plupart des méthodes d'imputation utilisées dans la pratique donnent des estimateurs imputés qui sont des fonctions linéaires des valeurs observées de la variable d'intérêt. Cela simplifie considérablement le calcul d'un estimateur de variance, même si l'on n'utilise qu'une seule méthode d'imputation. Pour estimer la part de la variance

1. Jean-François Beaumont, Statistique Canada, Division de la recherche et de l'innovation en statistique, pré Tunney, Ottawa (Ontario), Canada, K1A 0T6. Courriel : jean-francois.beaumont@statcan.gc.ca ; Joël Bissonnette, Statistique Canada, Division des méthodes d'enquêtes auprès des entreprises, pré Tunney, Ottawa (Ontario), Canada, K1A 0T6. Courriel : joel.bissonnette@statcan.gc.ca.

globale qui est attribuable à l'échantillonnage, nous utilisons une méthode (voir Beaumont et Bocci 2009) légèrement différente de celle proposée par Särndal (1992), ce qui permet de simplifier encore plus les calculs. Les résultats de nos travaux ont été mis en œuvre dans la version 2 du Système d'estimation de la variance due à la non-réponse et à l'imputation (SEVANI), qui a été mis au point à Statistique Canada (voir Beaumont, Bissonnette et Bocci 2010).

Le plan de l'article est le suivant. À la section 2, nous présentons la notation et expliquons ce qu'est l'imputation composite. L'imputation linéaire est définie à la section 3. À la section 4, nous décrivons notre approche d'inférence ainsi que nos principales hypothèses. Différents résultats sont exposés à la section 5 concernant l'estimation de la variance sous imputation composite. La section 6 présente les résultats d'une étude par simulation visant à évaluer l'efficacité de notre estimateur de la variance. À la section 7, nous commentons brièvement l'approche inverse afin de mettre en lumière les différences par rapport à notre approche. La section 8 est une courte conclusion.

2. Qu'est-ce que l'imputation composite ?

Supposons que nous voulons estimer le total d'un domaine de population $\theta = \sum_{k\in U} d_k y_k$, où U est la population finie de taille N, y est la variable d'intérêt et d est une variable indicatrice de domaine précisant si l'unité de population k fait partie du domaine d'intérêt ($d_k = 1$) ou non ($d_k = 0$). Un échantillon s de taille n est choisi à partir de la population finie U selon un plan d'échantillonnage probabiliste $p(s)$. S'il n'y a pas de valeurs manquantes, θ peut être estimé au moyen de l'estimateur de Horvitz-Thompson $\hat{\theta} = \sum_{k\in s} w_k d_k y_k$, où $w_k = 1/\pi_k$ est le poids de sondage et π_k est la probabilité de sélection de l'unité k. Il serait possible d'étendre nos résultats aux estimateurs de calage, mais, par souci de simplicité, nous ne le faisons pas ici.

La variable y peut être manquante pour certaines des unités échantillonnées, mais nous faisons l'hypothèse que la variable indicatrice de domaine d est toujours observée pour ces unités. L'ensemble des unités échantillonnées qui sont assorties d'une valeur de y observée – les répondants – est désigné au moyen de s_r. On suppose que cet ensemble a été généré conformément à un mécanisme de non-réponse $q(s_r \mid s)$. L'ensemble des non-répondants est désigné par $s_m = s - s_r$. Il est subdivisé en J sous-ensembles mutuellement exclusifs, $s_m^{(j)}$, $j = 1, ..., J$, de sorte que $s_m = \bigcup_{j=1}^{J} s_m^{(j)}$, si l'on a recours à l'imputation composite au moyen de $J > 1$ méthodes d'imputation. Toutes les valeurs de y manquantes dans un sous-ensemble donné $s_m^{(j)}$ sont imputées au moyen de la même méthode, j. Toutefois, différentes méthodes d'imputation sont utilisées pour imputer les valeurs manquantes dans différents sous-ensembles. Nous pouvons exprimer de la façon suivante l'estimateur imputé ainsi obtenu :

$$\hat{\theta}_I = \sum_{k\in s_r} w_k d_k y_k + \sum_{k\in s_m} w_k d_k y_k^* = \sum_{k\in s_r} w_k d_k y_k + \sum_{j=1}^{J} \sum_{k\in s_m^{(j)}} w_k d_k y_k^*, \qquad (2.1)$$

où y_k^* est la valeur imputée de y pour l'unité k.

L'imputation composite est couramment utilisée dans les enquêtes auprès des entreprises. La raison en est qu'il y a des valeurs manquantes dans les variables auxiliaires dont on se sert pour l'imputation. De manière à préciser les idées, représentons au moyen de $\mathbf{x}_k$ le vecteur complet de variables auxiliaires pour l'unité k. Idéalement, toutes les valeurs de y qui sont manquantes seraient imputées au moyen d'une seule méthode d'imputation à partir du vecteur complet $\mathbf{x}_k$. Malheureusement, certaines valeurs peuvent manquer dans les variables auxiliaires, de sorte que nous ne pouvons nous servir de $\mathbf{x}_k$ pour imputer les valeurs de y manquantes pour certains non-répondants ; nous pouvons uniquement utiliser un sous-ensemble de $\mathbf{x}_k$. Le vecteur des variables auxiliaires observées pour l'unité k est désigné par $\mathbf{x}_k^{\text{obs}}$. Ce vecteur ne contient pas nécessairement les mêmes variables observées d'une unité à l'autre. Aux fins d'imputer les valeurs de y manquantes pour une unité k donnée, une méthode d'imputation est choisie en fonction des variables auxiliaires disponibles. Puisqu'il peut exister un certain nombre de profils de non-réponse à l'intérieur du vecteur complet de variables auxiliaires, la stratégie d'imputation peut comporter un certain nombre de méthodes d'imputation.

Exemple

L'exemple qui suit pourra aider à mieux saisir les questions soulevées par l'estimation de la variance en cas d'imputation composite. Supposons que le vecteur complet de variables auxiliaires pour l'unité k est $\mathbf{x}_k = (x_{1k}, x_{2k})$, où x_{1k} est fortement corrélé à y_k mais peut présenter des valeurs manquantes, et x_{2k} est une constante pour toutes les unités échantillonnées ($x_{2k} = 1, k \in s$). Idéalement, on utilisera x_{1k} pour imputer y_k si cette valeur est manquante. Si l'on ne dispose pas de x_{1k}, on peut uniquement utiliser x_{2k}. Le tableau 1 résume l'information disponible pour les divers sous-ensembles de l'échantillon s.

Tableau 1
Information disponible quand il existe une variable auxiliaire x_1 et une constante x_2

Sous-ensembles		y	x_1	x_2	$\mathbf{x}^{\text{obs}}$
s_r	$s_r^{(1)}$	O	O	O	(x_1, x_2)
	$s_r^{(2)}$	O	M	O	(M, x_2)
s_m	$s_m^{(1)}$	M	O	O	(x_1, x_2)
	$s_m^{(2)}$	M	M	O	(M, x_2)

O : valeur observée ; M : valeur manquante.

L'ensemble de non-répondants s_m est subdivisé en deux sous-ensembles, $s_m^{(1)}$ et $s_m^{(2)}$, d'après la disponibilité de x_1. De même, l'ensemble de répondants est subdivisé entre $s_r^{(1)}$ et $s_r^{(2)}$. Dans cet exemple, nous pourrions utiliser l'imputation par le ratio pour imputer les valeurs manquantes de y dans $s_m^{(1)}$ et l'imputation par la moyenne pour les imputer dans $s_m^{(2)}$. Il faut mentionner que l'on pourrait opter pour l'imputation par régression linéaire simple plutôt que pour l'imputation par le ratio (si elle est mieux ajustée aux données). Nous avons opté ici pour l'imputation par le ratio parce qu'il s'agit d'une méthode simple et qui est souvent utilisée dans les enquêtes auprès des entreprises.

Seuls les répondants du sous-ensemble $s_r^{(1)}$ peuvent être utilisés pour imputer les valeurs manquantes de y dans $s_m^{(1)}$ avec l'imputation par le ratio. La valeur imputée pour une unité k dans $s_m^{(1)}$ est $y_k^* = x_{1k} \sum_{l \in s_r^{(1)}} \omega_l^{(1)} y_l / \sum_{l \in s_r^{(1)}} \omega_l^{(1)} x_{1l}$, où $\omega_l^{(1)}$ est un poids utilisé aux fins de l'imputation par le ratio (méthode 1). Les choix typiques seront $\omega_l^{(1)} = w_l$ (imputation pondérée par les poids de sondage) ou $\omega_l^{(1)} = 1$ (imputation non pondérée). Dans le cas de l'imputation par la moyenne, on peut utiliser les répondants du sous-ensemble $s_r^{(2)}$ ainsi que ceux du sous-ensemble $s_r^{(1)}$ pour imputer les valeurs manquantes de y dans $s_m^{(2)}$. Dans la pratique, il est courant d'utiliser les deux ensembles de répondants afin d'accroître la stabilité de la moyenne imputée. La valeur imputée pour une unité k dans $s_m^{(2)}$ est

$$y_k^* = \sum_{l \in s_r} \omega_l^{(2)} y_l \Big/ \sum_{l \in s_r} \omega_l^{(2)},$$

où $\omega_l^{(2)}$ est un poids utilisé aux fins de l'imputation par la moyenne (méthode 2) (les choix typiques pour $\omega_l^{(2)}$ seront les mêmes que pour $\omega_l^{(1)}$, soit $\omega_l^{(2)} = w_l$ ou $\omega_l^{(2)} = 1$). Cela signifie que les unités faisant partie de $s_r^{(1)}$ peuvent être utilisées dans les deux méthodes. Cette situation soulève des problèmes lorsque l'on veut estimer la variance associée à l'estimateur par imputation composite ainsi obtenu. Ces problèmes sont commentés à la section 5.

3. Qu'est-ce que l'imputation linéaire ?

La méthode d'imputation j est dite linéaire si la valeur imputée y_k^* pour une unité échantillonnée $k \in s_m^{(j)}$ peut s'écrire sous la forme linéaire :

$$y_k^* = \varphi_{0k}^{(j)} + \sum_{l \in s_r} \varphi_{lk}^{(j)} y_l. \qquad (3.1)$$

Les quantités $\varphi_{0k}^{(j)}$ et $\varphi_{lk}^{(j)}$, pour $l \in s_r$, sont obtenues sans que l'on utilise les valeurs de y, mais elles peuvent dépendre de s et s_r. Dans la pratique, plusieurs des méthodes d'imputation les plus courantes satisfont à la forme linéaire (3.1), par exemple l'imputation par régression linéaire (pondérée ou non), l'imputation par donneur et l'imputation par valeur auxiliaire. On trouvera un bon examen de ces méthodes dans Haziza (2009). Mentionnons que l'imputation par valeur auxiliaire ne fait pas appel aux valeurs de y des répondants, c'est-à-dire, $y_k^* = \varphi_{0k}^{(j)}$ (voir Beaumont, Haziza et Bocci 2011). Dans le cas de l'imputation par donneur, la valeur imputée y_k^* est égale à la valeur de y d'un répondant choisi de façon appropriée (donneur), de sorte que $\varphi_{0k}^{(j)} = 0$ et $\varphi_{lk}^{(j)} = 0$ pour tous les répondants $l \in s_r$ sauf un. On trouvera des expressions détaillées de $\varphi_{0k}^{(j)}$ and $\varphi_{lk}^{(j)}$ dans le guide méthodologique du SEVANI (Beaumont, Bissonnette et Bocci 2010), que l'on peut obtenir sur demande auprès des auteurs.

Supposons que $\Omega_I^{(j)} = \sum_{k \in s_m^{(j)}} w_k d_k y_k^*$ est la contribution de la méthode d'imputation j à l'estimateur $\hat{\theta}_I$. La forme (3.1) permet de décomposer $\Omega_I^{(j)}$ de la façon suivante :

$$\begin{aligned} \Omega_I^{(j)} &= \sum_{k \in s_m^{(j)}} w_k d_k y_k^* \\ &= \sum_{k \in s_m^{(j)}} w_k d_k \varphi_{0k}^{(j)} + \sum_{l \in s_r} y_l \sum_{k \in s_m^{(j)}} w_k d_k \varphi_{lk}^{(j)} \\ &= W_{0d}^{(j)} + \sum_{l \in s_r} W_{dl}^{(j)} y_l, \end{aligned} \qquad (3.2)$$

où $W_{0d}^{(j)} = \sum_{k \in s_m^{(j)}} w_k d_k \varphi_{0k}^{(j)}$ et $W_{dl}^{(j)} = \sum_{k \in s_m^{(j)}} w_k d_k \varphi_{lk}^{(j)}$. À partir de (3.2), l'estimateur imputé (2.1) peut être exprimé sous la forme linéaire suivante :

$$\begin{aligned} \hat{\theta}_I &= \sum_{k \in s_r} w_k d_k y_k + \sum_{j=1}^{J} \Omega_I^{(j)} \\ &= W_{0d}^{(+)} + \sum_{k \in s_r} \left(w_k d_k + W_{dk}^{(+)} \right) y_k, \end{aligned} \qquad (3.3)$$

où $W_{0d}^{(+)} = \sum_{j=1}^{J} W_{0d}^{(j)}$ et $W_{dk}^{(+)} = \sum_{j=1}^{J} W_{dk}^{(j)}$.

Si l'on reprend l'exemple présenté à la fin de la section 2, on constate que, aux fins d'imputation par le ratio, $\varphi_{0k}^{(1)} = 0$ et $\varphi_{lk}^{(1)} = \omega_l^{(1)} x_{1k} / \sum_{l \in s_r} \omega_l^{(1)} x_{1l}$, pour $l \in s_r$, où $\omega_l^{(1)} = 0$, pour $l \in s_r^{(2)}$. Dans le cas de l'imputation par la moyenne, $\varphi_{0k}^{(2)} = 0$ et $\varphi_{lk}^{(2)} = \omega_l^{(2)} / \sum_{l \in s_r} \omega_l^{(2)}$, pour $l \in s_r$. Par conséquent, $W_{0d}^{(1)} = 0$, $W_{0d}^{(2)} = 0$,

$$W_{dl}^{(1)} = \omega_l^{(1)} \sum_{k \in s_m^{(1)}} w_k d_k x_{1k} \Big/ \sum_{k \in s_r} \omega_k^{(1)} x_{1k}$$

et $W_{dl}^{(2)} = \omega_l^{(2)} \sum_{k \in s_m^{(2)}} w_k d_k / \sum_{k \in s_r} \omega_k^{(2)}$. Cela signifie que $W_{0d}^{(+)} = 0$ et que $W_{dk}^{(+)} = W_{dk}^{(1)} + W_{dk}^{(2)}$.

4. Approche d'inférence et principales hypothèses

Nous considérons trois sources de variabilité quand nous évaluons les espérances et les variances de l'estimateur imputé : la variabilité due au modèle d'imputation, au plan d'échantillonnage et au mécanisme de non-réponse. Il est à noter que l'utilisation d'un modèle d'imputation pour faire une inférence lorsqu'il y a imputation est mentionnée par Rubin (1987), Hidiroglou (1989) et Särndal (1992). Dans la

suite de l'article, nous utiliserons les indices *m*, *p* et *q* pour dénoter les espérances, les variances et les covariances évaluées par rapport au modèle d'imputation, au plan d'échantillonnage et au mécanisme de non-réponse, respectivement.

Nous prenons le modèle d'imputation suivant pour décrire la relation entre la variable y et le vecteur de variables auxiliaires observées $\mathbf{x}^{\text{obs}}$:

$$\begin{aligned} E_m(y_k \mid \mathbf{X}^{\text{obs}}) &= \mu_k \\ V_m(y_k \mid \mathbf{X}^{\text{obs}}) &= \sigma_k^2 \\ \text{cov}_m(y_k, y_l \mid \mathbf{X}^{\text{obs}}) &= 0, \end{aligned} \qquad (4.1)$$

pour $k \neq l$ et $k, l \in U$. La matrice de population, $\mathbf{X}^{\text{obs}}$, contient les vecteurs de variables auxiliaires observées, $\mathbf{x}_k^{\text{obs}}$, pour $k \in U$, tandis que μ_k et σ_k^2 sont des fonctions de $\mathbf{x}_k^{\text{obs}}$. Les estimateurs – asymptotiquement sans biais par rapport à m et convergents sous m – de μ_k et de σ_k^2 sont désignés par $\hat{\mu}_k$ et $\hat{\sigma}_k^2$, respectivement. Puisque nous conditionnerons systématiquement sur $\mathbf{X}^{\text{obs}}$, nous excluons cette notation pour simplifier. Par exemple, $E_m(y_k \mid \mathbf{X}^{\text{obs}})$ sera écrit sous la forme $E_m(y_k)$.

Dans le modèle (4.1), nous conditionnons sur les variables auxiliaires observées. Étant donné que le profil de non-réponse du vecteur $\mathbf{x}$ n'est pas le même pour tous les non-répondants, nous devons valider et ajuster un modèle conditionnel distinct pour chaque profil de non-réponse. En principe, ces modèles conditionnels devraient être utilisés pour déterminer quelles méthodes d'imputation il convient de choisir. Notons que le modèle (4.1) se réduit au modèle conditionnel standard (par exemple, Särndal 1992) quand le vecteur $\mathbf{x}$ de variables auxiliaires ne comprend pas de valeurs manquantes.

Remarque : Pour que la méthode d'estimation de la variance à la section 5 soit valide, il faut spécifier correctement μ_k et σ_k^2. Bien qu'une forme paramétrique de μ_k soit souvent acceptable, il pourrait être plus difficile de déterminer une forme paramétrique appropriée de σ_k^2. Pour éviter ce problème et pour obtenir une certaine robustesse aux erreurs de spécification de la variance du modèle, σ_k^2 peut être estimé de façon non paramétrique ; on trouvera dans l'étude empirique de Beaumont, Haziza et Bocci (2011) une illustration de cette propriété dans le contexte d'une imputation par valeur auxiliaire. Relativement à l'imputation par donneur, Beaumont et Bocci (2009) ont montré empiriquement que l'estimation non paramétrique de μ_k et σ_k^2, par voie de lissage par splines pénalisées, réduit de façon significative la vulnérabilité de notre estimateur de la variance aux erreurs de spécification de la moyenne et de la variance du modèle.

En complément du modèle d'imputation (4.1), nous supposons que :

$$F(\mathbf{Y} \mid s, s_r, \mathbf{X}^{\text{obs}}, \mathbf{Z}, \mathbf{D}) = F(\mathbf{Y} \mid \mathbf{X}^{\text{obs}}), \qquad (4.2)$$

où $F(\cdot)$ dénote la fonction de répartition, $\mathbf{Y}$ et $\mathbf{D}$ sont des vecteurs à N éléments contenant respectivement y_k et d_k comme k^{e} élément, et $\mathbf{Z}$ est une matrice de N lignes d'information sur le plan de sondage, qui contient implicitement ou explicitement l'information sur les probabilités de sélection π_k et les probabilités de sélection conjointe π_{kl}, pour $k, l \in U$. Cette hypothèse, souvent implicite dans d'autres articles, nous permet de traiter les indicateurs de réponse, les indicateurs de domaine et l'information sur le plan de sondage comme étant fixes lorsque nous considérons les espérances et les variances sous le modèle. Il faut choisir soigneusement les variables auxiliaires pour satisfaire à cette hypothèse. Par exemple, l'information sur le plan de sondage et les indicateurs de domaine devraient être envisagés à titre d'éventuelles variables auxiliaires.

La stratégie d'imputation exposée dans l'exemple amorcé à la section 2 pourrait être justifiée au moyen d'un modèle avec $\mu_k = \beta_1 x_{1k}$ et $\sigma_k^2 = \sigma_1^2 x_{1k}$, pour $k \in s_r^{(1)}$ ou $k \in s_m^{(1)}$, et $\mu_k = \beta_2$ et $\sigma_k^2 = \sigma_2^2$, pour $k \in s_r^{(2)}$ ou $k \in s_m^{(2)}$. Les paramètres β_1, β_2, σ_1^2 et σ_2^2 du modèle sont inconnus. Il est à noter que, si l'on suppose que les variables x_{1k} sont des variables aléatoires identiquement distribuées de moyenne μ_x et de variance σ_x^2, $\beta_2 = \beta_1 \mu_x$ et $\sigma_2^2 = \beta_1^2 \sigma_x^2 + \sigma_1^2 \mu_x$. On obtient les valeurs imputées $y_k^* = \hat{\mu}_k$, pour $k \in s_m$, en estimant les paramètres β_1 et β_2 du modèle à partir des données observées. Ainsi, les estimateurs sans biais par rapport à m de β_1 et β_2 pourraient être

$$\hat{\beta}_1 = \sum_{k \in s_r^{(1)}} \omega_k^{(1)} y_k \Big/ \sum_{k \in s_r^{(1)}} \omega_k^{(1)} x_{1k}$$

et

$$\hat{\beta}_2 = \sum_{k \in s_r^{(2)}} \omega_k^{(2)} y_k \Big/ \sum_{k \in s_r^{(2)}} \omega_k^{(2)}$$

respectivement. Dès lors, $\hat{\mu}_k = \hat{\beta}_1 x_{1k}$, pour $k \in s_r^{(1)}$ ou $k \in s_m^{(1)}$, et $\hat{\mu}_k = \hat{\beta}_2$, pour $k \in s_r^{(2)}$ ou $k \in s_m^{(2)}$. Tout comme à la section 2, on peut envisager l'utilisation de l'estimateur – potentiellement plus efficace – $\hat{\beta}_2^* = \sum_{k \in s_r} \omega_k^{(2)} y_k / \sum_{k \in s_r} \omega_k^{(2)}$ à la place de $\hat{\beta}_2$. Malheureusement, $\hat{\beta}_2^*$ est biaisé sous le modèle, puisque :

$$E_m(\hat{\beta}_2^* \mid s, s_r) = \beta_2 + \frac{\sum_{k \in s_r^{(1)}} \omega_k^{(2)} (x_{1k} \beta_1 - \beta_2)}{\sum_{k \in s_r} \omega_k^{(2)}}. \qquad (4.3)$$

Tel que mentionné plus haut, si l'on suppose que les variables x_{1k} sont des variables aléatoires identiquement distribuées de moyenne μ_x et de variance σ_x^2, $\beta_2 = \beta_1 \mu_x$ et l'équation (4.3) peut être reformulée ainsi :

$$\begin{aligned} E_m(\hat{\beta}_2^* \mid s, s_r) &= \beta_2 \\ &+ \beta_1 \frac{\sum_{k \in s_r^{(1)}} \omega_k^{(2)}}{\sum_{k \in s_r} \omega_k^{(2)}} \frac{\sum_{k \in s_r^{(1)}} \omega_k^{(2)} (x_{1k} - \mu_x)}{\sum_{k \in s_r^{(1)}} \omega_k^{(2)}}. \end{aligned} \qquad (4.4)$$

On peut montrer que, dans des conditions faibles, $E_m(\hat{\beta}_2^* \mid s, s_r) = \beta_2 + O_p(1/\sqrt{n})$, de sorte que le biais de $\hat{\beta}_2^*$ sous le modèle est asymptotiquement négligeable. Toutefois, étant donné que $\text{var}_m(\hat{\beta}_2^* \mid s, s_r) = O_p(1/n)$, le biais quadratique du modèle n'est pas nécessairement asymptotiquement négligeable par rapport à la variance de $\hat{\beta}_2^*$ sous le modèle. Au moins, $\hat{\beta}_2^*$ est convergent sous m pour β_2. On peut voir à partir de l'équation (4.3) ou (4.4) qu'il est possible de contrôler le biais de $\hat{\beta}_2^*$ sous le modèle en appliquant un poids $\omega_k^{(2)}$ plus faible aux unités $k \in s_r^{(1)}$ qu'aux unités $k \in s_r^{(2)}$. Par exemple, on pourrait envisager d'utiliser $\omega_k^{(2)} = w_k / n^\alpha$, pour $k \in s_r^{(1)}$ et une certaine valeur de $\alpha > 0$, et $\omega_k^{(2)} = w_k$, pour $k \in s_r^{(2)}$. Dans le cas extrême où $\omega_k^{(2)} = 0$, pour $k \in s_r^{(1)}$, $\hat{\beta}_2^*$ est sans biais sous le modèle, car il est égal à $\hat{\beta}_2$. Précisons que le biais de $\hat{\beta}_2^*$ sous le modèle pourrait être supérieur à $O_p(1/\sqrt{n})$ si x_{1k}, $k \in s_r^{(1)}$, présentent une moyenne différente de x_{1k}, $k \in s_r^{(2)}$. Il pourrait alors être plus important de contrôler le biais de $\hat{\beta}_2^*$ sous le modèle.

Dans le cas de l'imputation par donneur, il faut tenir compte d'une quatrième source de variabilité quand les donneurs sont sélectionnés aléatoirement parmi les répondants pour imputer les données des non-répondants. Dans cet article, l'indice q indiquera implicitement que les moments sont évalués par rapport à la distribution conjointe induite par le mécanisme de non-réponse et le mécanisme de sélection aléatoire des donneurs. Par conséquent, lorsque nous procédons au conditionnement sur s_r, comme dans l'équation (4.2), il faut se rappeler que le conditionnement est effectué non seulement sur l'ensemble de répondants, mais aussi sur l'ensemble de donneurs sélectionnés.

5. Estimation de la variance

Särndal (1992) exprime l'erreur totale de l'estimateur imputé sous la forme suivante :

$$\hat{\theta}_I - \theta = (\hat{\theta} - \theta) + (\hat{\theta}_I - \hat{\theta}), \qquad (5.1)$$

où le premier terme du membre droit de (5.1) est appelé erreur d'échantillonnage et le deuxième, erreur due à la non-réponse. Si nous reprenons les hypothèses de la section 4, et à partir du moment où $E_p(\hat{\theta} - \theta) = 0$, le biais global de l'estimateur imputé se réduit à $E_{mpq}(\hat{\theta}_I - \theta) = E_{pq}B_m$, où $B_m = E_m(\hat{\theta}_I - \hat{\theta} \mid s, s_r)$ est le biais (conditionnel) de l'estimateur imputé par rapport au modèle. À partir de (2.1), le biais par rapport au modèle peut s'exprimer sous la forme suivante :

$$B_m = \sum_{j=1}^{J} \sum_{k \in s_m^{(j)}} w_k d_k E_m(y_k^* - y_k \mid s, s_r). \qquad (5.2)$$

Cela signifie que le biais par rapport au modèle et le biais global disparaissent si l'espérance sous le modèle de l'erreur d'imputation, $y_k^* - y_k$, est nulle, pour $k \in s_m^{(j)}$ et $j = 1, ..., J$. En principe, la stratégie d'imputation doit être choisie de sorte que cette condition soit remplie, au moins approximativement. C'est une hypothèse courante dans la littérature (par exemple, Särndal 1992 ; Shao et Steel 1999).

Dans l'exemple amorcé à la section 2, le biais par rapport au modèle (5.2) se réduit à :

$$B_m = \left(\sum_{k \in s_m^{(2)}} w_k d_k \right) E_m(\hat{\beta}_2^* - \beta_2 \mid s, s_r).$$

L'expression de $E_m(\hat{\beta}_2^* - \beta_2 \mid s, s_r)$ peut être donnée par (4.3) ou (4.4). Tel que mentionné dans le paragraphe qui suit l'équation (4.4), on peut contrôler le biais par rapport au modèle, B_m, en appliquant un poids $\omega_k^{(2)}$ plus faible aux unités $k \in s_r^{(1)}$ qu'aux unités $k \in s_r^{(2)}$. Le biais sera également peu marqué si le nombre de non-répondants à l'égard desquels on a procédé à l'imputation au moyen de la méthode 2 est peu élevé. Il est à noter que notre approche d'estimation de la variance (ou de l'erreur quadratique moyenne, ou EQM) requiert l'hypothèse un peu plus faible selon laquelle $E_q(B_m \mid s)$ est négligeable (se reporter à la section 5.3).

À partir de (5.1), Särndal (1992) décompose l'EQM globale en trois composantes :

$$E_{mpq}(\hat{\theta}_I - \theta)^2 = E_m \text{var}_p(\hat{\theta}) + E_{pq}E_m\{(\hat{\theta}_I - \hat{\theta})^2 \mid s, s_r\} + 2E_{pq}E_m\{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) \mid s, s_r\}. \qquad (5.3)$$

L'EQM globale (5.3) équivaut dès lors à peu près à la variance globale, $\text{var}_{mpq}(\hat{\theta}_I - \theta)$, lorsque le biais global est négligeable. Les premier, deuxième et troisième termes du membre droit de (5.3) sont appelés variance d'échantillonnage, variance due à la non-réponse et composante mixte, respectivement. La somme des deux derniers termes peut être appelée composante due à la non-réponse, car ces termes disparaissent en l'absence de non-réponse. La composante due à la non-réponse correspond simplement à la différence entre l'EQM – ou la variance – globale et la variance d'échantillonnage. Nous allons ci-après élaborer un estimateur pour chacun de ces trois termes.

5.1 Estimation de la variance d'échantillonnage

Soit $v(y)$, un estimateur de $\text{var}_p(\hat{\theta})$ sans biais par rapport à p que nous utiliserions s'il y avait réponse complète. L'estimateur de Horvitz-Thompson typique est :

$$v(y) = \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} (w_k d_k y_k)(w_l d_l y_l), \qquad (5.4)$$

où π_{kl} est la probabilité de sélection conjointe des unités k et l. En présence de non-réponse, $\hat{V}_{\text{ORD}} = v(y_\bullet)$ est l'estimateur naïf de la variance d'échantillonnage qui traite les valeurs imputées comme des valeurs réelles, où $y_\bullet$ est la variable y imputée, c'est-à-dire $y_{\bullet k} = y_k$, pour $k \in s_r$, et $y_{\bullet k} = y_k^*$, pour $k \in s_m$.

Särndal (1992) a proposé l'estimateur sans biais par rapport à m, p et q suivant de la variance d'échantillonnage $V_{\text{SAM}} = E_m \text{var}_p(\hat{\theta})$:

$$\hat{V}_{\text{SAM}} = \hat{V}_{\text{ORD}} + \hat{V}_{\text{DIF}},$$

où $\hat{V}_{\text{DIF}}$ est un estimateur sans biais par rapport à m de $V_{\text{DIF}} = E_m(v(y) - \hat{V}_{\text{ORD}} \mid s, s_r)$. Malheureusement, l'expression de $\hat{V}_{\text{DIF}}$ est généralement difficile à déterminer, surtout quand on a recours à l'imputation composite.

Beaumont et Bocci (2009) ont simplifié les calculs de Särndal en conditionnant sur $\mathbf{Y}_r$, soit le vecteur contenant les valeurs de y pour les répondants. De façon plus explicite, soit $V_{\text{DIF}}^{\text{C}} = E_m(v(y) - \hat{V}_{\text{ORD}} \mid s, s_r, \mathbf{Y}_r)$, et $\hat{V}_{\text{DIF}}^{\text{C}}$ est un estimateur sans biais par rapport à m de $V_{\text{DIF}}^{\text{C}}$; c'est-à-dire, $E_m(\hat{V}_{\text{DIF}}^{\text{C}} \mid s, s_r, \mathbf{Y}_r) = V_{\text{DIF}}^{\text{C}}$. Notre estimateur de la variance d'échantillonnage sans biais par rapport à m, p, et q est $\hat{V}_{\text{SAM}}^{\text{C}} = \hat{V}_{\text{ORD}} + \hat{V}_{\text{DIF}}^{\text{C}}$. Étant donné que $\hat{V}_{\text{ORD}}$ est une constante lorsque l'on conditionne sur s, s_r et $\mathbf{Y}_r$, on peut obtenir $\hat{V}_{\text{SAM}}^{\text{C}}$ simplement en estimant $E_m(v(y) \mid s, s_r, \mathbf{Y}_r)$. Si l'on utilise (5.4) :

$$E_m(v(y) \mid s, s_r, \mathbf{Y}_r) = v(y_\bullet^\mu) + \sum_{k \in s_m} (1 - \pi_k) w_k^2 d_k \sigma_k^2, \quad (5.5)$$

où $y_{\bullet k}^\mu = y_k$, pour $k \in s_r$, et $y_{\bullet k}^\mu = \mu_k$, pour $k \in s_m$. Nous obtenons un estimateur $\hat{V}_{\text{SAM}}^{\text{C}}$ de (5.5) en remplaçant la moyenne inconnue μ_k et la variance inconnue σ_k^2 dans (5.5) par les estimateurs sans biais par rapport à m (ou à tout le moins convergents sous m) $\hat{\mu}_k$ et $\hat{\sigma}_k^2$. Cet estimateur est facile à calculer, à condition de disposer d'un progiciel qui traite les cas de réponse complète pour obtenir le premier terme du membre droit de (5.5). La formule générale (5.5) peut être utilisée pour chaque stratégie d'imputation. La seule différence entre les diverses stratégies a trait au choix du modèle d'imputation et des estimateurs $\hat{\mu}_k$ et $\hat{\sigma}_k^2$.

5.2 Estimation de la variance due à la non-réponse

Nous obtenons un estimateur sans biais par rapport à m, p et q de la variance due à la non-réponse $V_{\text{NR}} = E_{pq} E_m\{(\hat{\theta}_I - \hat{\theta})^2 \mid s, s_r\}$ en trouvant un estimateur sans biais par rapport à m de :

$$E_m\{(\hat{\theta}_I - \hat{\theta})^2 \mid s, s_r\} = \text{var}_m\{(\hat{\theta}_I - \hat{\theta}) \mid s, s_r\} + B_m^2. \quad (5.6)$$

À partir de $\hat{\theta}_I$ tel que défini dans la première équation de (3.3), nous pouvons décomposer l'erreur due à la non-réponse sous imputation composite en J composantes :

$$\hat{\theta}_I - \hat{\theta} = \sum_{j=1}^{J} (\Omega_I^{(j)} - \Omega^{(j)}),$$

où $\Omega^{(j)} = \sum_{k \in s_m^{(j)}} w_k d_k y_k$. Chacune de ces j composantes, $\Omega_I^{(j)} - \Omega^{(j)}$, est associée à une méthode d'imputation différente. Étant donné que y_k^* ne fait intervenir que les valeurs de y observées, il en est de même de $\Omega_I^{(j)} = \sum_{k \in s_m^{(j)}} w_k d_k y_k^*$ de sorte que $\Omega_I^{(j)}$ et $\Omega^{(j)}$ sont indépendants sous le modèle. Par conséquent, la variance sous le modèle de l'erreur due à la non-réponse peut s'écrire sous la forme suivante :

$$\text{var}_m\{(\hat{\theta}_I - \hat{\theta}) \mid s, s_r\} = \sum_{i=1}^{J} \sum_{j=1}^{J} \text{cov}_m(\Omega_I^{(i)}, \Omega_I^{(j)} \mid s, s_r) + \sum_{j=1}^{J} \text{var}_m(\Omega^{(j)} \mid s, s_r). \quad (5.7)$$

Précisons que les covariances $\text{cov}_m(\Omega_I^{(i)}, \Omega_I^{(j)} \mid s, s_r)$, pour $i \neq j$, ne seront pas forcément négligeables, car certaines des valeurs de y observées peuvent être utilisées dans plus d'une méthode d'imputation.

Les calculs de la variance sous le modèle à partir de (5.7) peuvent être assez compliqués lorsque l'on utilise plusieurs méthodes d'imputation, du fait que les covariances ne sont pas négligeables. Dans le cas des méthodes d'imputation linéaire, le traitement algébrique est grandement simplifié. Si l'on utilise la deuxième équation de (3.3), l'erreur due à la non-réponse s'exprime ainsi :

$$\hat{\theta}_I - \hat{\theta} = W_{0d}^{(+)} + \sum_{k \in s_r} W_{dk}^{(+)} y_k - \sum_{k \in s_m} w_k d_k y_k. \quad (5.8)$$

Puisque l'erreur due à la non-réponse est linéaire dans les valeurs de y, sa variance sous le modèle est donnée par :

$$\text{var}_m\{(\hat{\theta}_I - \hat{\theta}) \mid s, s_r\} = \sum_{k \in s_r} (W_{dk}^{(+)})^2 \sigma_k^2 + \sum_{k \in s_m} w_k^2 d_k \sigma_k^2. \quad (5.9)$$

Si le biais par rapport au modèle, B_m, est négligeable, on obtient un estimateur $\hat{V}_{\text{NR}}$ sans biais par rapport à m, p, et q, de la variance due à la non réponse, V_{NR}, en remplaçant σ_k^2 dans l'équation (5.9) par un estimateur sans biais par rapport à m (et convergent sous m), $\hat{\sigma}_k^2$. Si le biais par rapport au modèle n'est pas négligeable, on pourra l'estimer au moyen d'un estimateur convergent sous m, $\hat{B}_m$; à partir de l'équation (5.6), l'estimateur de la variance due à la non-réponse $\hat{V}_{\text{NR}}$ peut être remplacé par $\hat{V}_{\text{NR}} + \hat{B}_m^2$. Il est à noter que $\hat{B}_m^2$ est convergent sous m pour B_m^2, pourvu que $\hat{B}_m$ soit convergent sous m pour B_m. On peut obtenir l'estimateur $\hat{B}_m$ en utilisant (5.8) et en écrivant le biais par rapport au modèle sous la forme suivante :

$$B_m = E_m(\hat{\theta}_I - \hat{\theta} \mid s, s_r) = W_{0d}^{(+)} + \sum_{k \in s_r} W_{dk}^{(+)} \mu_k - \sum_{k \in s_m} w_k d_k \mu_k. \quad (5.10)$$

Pour obtenir l'estimateur $\hat{B}_m$, nous remplaçons μ_k dans la formule (5.10) par un estimateur convergent sous m, $\hat{\mu}_k$.

5.3 Estimation de la composante mixte

On obtient un estimateur sans biais par rapport à m, p et q de la composante mixte

$$V_{\text{MIX}} = 2E_{pq}E_m\{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) \mid s, s_r\}$$

en trouvant un estimateur sans biais par rapport à m de :

$$\begin{aligned} 2E_m\{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) \mid s, s_r\} = \\ 2\text{cov}_m\{(\hat{\theta}_I - \hat{\theta}), (\hat{\theta} - \theta) \mid s, s_r\} \\ + 2B_m E_m\{(\hat{\theta} - \theta) \mid s, s_r\}. \end{aligned} \quad (5.11)$$

Puisque l'erreur due à la non-réponse et l'erreur d'échantillonnage sont linéaires dans les valeurs de y, l'utilisation de (5.8) donne :

$$\begin{aligned} 2\text{cov}_m\{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) \mid s, s_r\} = \\ 2\sum_{k \in s_r} W_{dk}^{(+)}(w_k - 1)d_k\sigma_k^2 - 2\sum_{k \in s_m} w_k(w_k - 1)d_k\sigma_k^2. \end{aligned} \quad (5.12)$$

Si le biais par rapport au modèle, B_m, est négligeable, on obtient un estimateur $\hat{V}_{\text{MIX}}$ sans biais par rapport à m, p, et q de la composante mixte, V_{MIX}, en remplaçant σ_k^2 dans l'équation (5.12) par un estimateur sans biais par rapport à m (et convergent sous m), $\hat{\sigma}_k^2$. Il est à noter que la composante mixte ne sera pas forcément négligeable (Brick, Kalton et Kim 2004) et qu'elle est même souvent négative dans la pratique.

Si le biais par rapport au modèle, B_m, n'est pas négligeable, il ne sera peut-être pas possible d'estimer facilement la deuxième composante du membre droit de (5.11), parce que $E_m\{(\hat{\theta} - \theta) \mid s, s_r\}$ requiert que l'on connaisse $\mathbf{x}_k^{\text{obs}}$ ainsi que la variable indicatrice de domaine d pour la partie non échantillonnée de la population, or cette information pourrait ne pas être disponible. Il est possible de surmonter le problème en modifiant le cadre inférenciel. Nous pouvons modéliser la distribution multivariée complète reliant y, $\mathbf{x}$ et d, au lieu de conditionner sur d et $\mathbf{x}^{\text{obs}}$. Nous n'avons pas jouté cette idée dans le SEVANI parce qu'elle entraîne une tâche de modélisation plus complexe et qu'il est difficile d'obtenir une expression générale de la variance facile à mettre en œuvre. En pratique, si le biais par rapport au modèle n'est pas trop grand, le fait d'ignorer la deuxième composante du membre droit de (5.11) ne devrait pas causer trop de souci. À la section 5.4, nous proposons une statistique qui peut aider à déterminer si le biais par rapport au modèle est important ou non.

La composante mixte peut aussi être exprimée sous la forme suivante :

$$\begin{aligned} V_{\text{MIX}} &= 2E_{pq}E_m\{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) \mid s, s_r\} \\ &= 2E_{pq}[\text{cov}_m\{(\hat{\theta}_I - \hat{\theta}), (\hat{\theta} - \theta) \mid s, s_r\}] \\ &\quad + 2E_p[E_q(B_m \mid s)E_m\{(\hat{\theta} - \theta) \mid s\}]. \end{aligned}$$

L'expression (5.12) peut dès lors être utilisée pour obtenir un estimateur de V_{MIX}, à condition que $E_q(B_m \mid s)$ soit négligeable. Cette hypothèse est plus faible que le fait d'exiger que B_m soit négligeable, puisqu'on y satisfait si B_m ou $E_q(\hat{\theta}_I - \hat{\theta} \mid s)$ est négligeable. Par exemple, dans notre exemple précédent, B_m peut ne pas être négligeable mais, si $d_k = 1$ et $\omega_k^{(1)} = \omega_k^{(2)} = w_k$, $E_q(\hat{\theta}_I - \hat{\theta} \mid s) \approx 0$ sous une non-réponse uniforme (se reporter à Sitter et Rao 1997).

5.4 Estimation de l'EQM globale/de la variance globale

L'EQM globale, ou la variance globale si le biais global est négligeable,

$$V_{\text{TOT}} = E_{mpq}(\hat{\theta}_I - \theta)^2 = V_{\text{SAM}} + V_{\text{NR}} + V_{\text{MIX}}$$

peut être estimée au moyen de $\hat{V}_{\text{TOT}} = \hat{V}_{\text{SAM}}^{\text{C}} + \hat{V}_{\text{NR}} + \hat{V}_{\text{MIX}}$ si le biais par rapport au modèle, B_m, est négligeable. L'estimateur de la composante de la non-réponse est $\hat{V}_{\text{NR}} + \hat{V}_{\text{MIX}}$. Du point de vue de l'utilisateur, l'estimateur $\hat{V}_{\text{TOT}}$ présente plus d'intérêt que ses composantes. Cela étant, l'utilisateur pourrait s'intéresser à l'estimateur de la variance d'échantillonnage, $\hat{V}_{\text{SAM}}^{\text{C}}$, ou au ratio $\hat{V}_{\text{SAM}}^{\text{C}}/\hat{V}_{\text{TOT}}$. Ce dernier sert à estimer l'apport de la variance d'échantillonnage à la variance globale.

Tel qu'indiqué à la section 5.2, si le biais par rapport au modèle n'est pas négligeable, la variance de la non-réponse peut être estimée par $\hat{V}_{\text{NR}} + \hat{B}_m^2$ plutôt que par $\hat{V}_{\text{NR}}$. L'estimateur de l'EQM globale est alors $\hat{V}_{\text{TOT, ADJ}} = \hat{V}_{\text{SAM}}^{\text{C}} + (\hat{V}_{\text{NR}} + \hat{B}_m^2) + \hat{V}_{\text{MIX}}$.

Une statistique pouvant être utile à titre de diagnostic pour déterminer l'ampleur du biais sous le modèle est $|\hat{B}_m| / \sqrt{\hat{V}_{\text{TOT}}}$, ou encore $|\hat{B}_m| / \sqrt{\hat{V}_{\text{TOT, ADJ}}}$. Si l'une ou l'autre de ces statistiques présente une valeur élevée, cela peut indiquer que le biais sous le modèle n'est pas négligeable et que la procédure d'imputation composite doit être remise en question. Comparativement à $|\hat{B}_m| / \sqrt{\hat{V}_{\text{TOT}}}$, $|\hat{B}_m| / \sqrt{\hat{V}_{\text{TOT, ADJ}}}$ présente l'avantage d'être bornée, c'est-à-dire :

$$0 \le |\hat{B}_m| / \sqrt{\hat{V}_{\text{TOT, ADJ}}} \le 1.$$

5.5 Imputation par la régression aléatoire

Un résidu de régression aléatoire e_k est parfois ajouté à la valeur imputée par régression y_k^* afin de préserver la variabilité naturelle de la variable y. Nous proposons de générer indépendamment les résidus aléatoires e_k avec $E_*(e_k \mid s, s_r) = 0$ et $\text{var}_*(e_k \mid s, s_r) = \hat{\sigma}_k^2$, où l'indice * indique que l'espérance et la variance sont considérées par rapport au mécanisme d'imputation aléatoire. Nous obtenons ainsi la valeur imputée $y_k^{*R} = y_k^* + r_k e_k$, où $r_k = 1$ si

un résidu aléatoire a été ajouté aux fins d'imputation de l'unité k, à défaut de quoi $r_k = 0$. L'estimateur imputé (2.1), dans lequel y_k^* est remplacé par y_k^{*R}, est désigné par $\hat{\theta}_I^* = \hat{\theta}_I + \sum_{k \in s_m} w_k\, d_k\, r_k\, e_k$. Étant donné que $E_*(e_k \mid s, s_r) = 0$, l'ajout d'un résidu aléatoire n'entraîne aucun biais dans l'estimateur imputé. L'EQM globale de $\hat{\theta}_I^*$ peut être exprimée ainsi :

$$E_{mpq*}(\hat{\theta}_I^* - \theta)^2 = E_{mpq}(\hat{\theta}_I - \theta)^2 + E_{mpq}\,\mathrm{var}_*(\hat{\theta}_I^* \mid s, s_r). \quad (5.13)$$

Nous estimons le premier terme du membre droit de (5.13) comme à la section 5.4. Nous estimons le deuxième terme par :

$$\mathrm{var}_*(\hat{\theta}_I^* \mid s, s_r) = \sum_{k \in s_m} w_k^2\, d_k\, r_k\, \hat{\sigma}_k^2. \quad (5.14)$$

6. Étude par simulation

Nous avons effectué une étude par simulation Monte Carlo pour évaluer la méthodologie décrite à la section 5. Une population bivariée de $N = 400$ unités a été générée contenant une variable auxiliaire x et une variable d'intérêt y. Pour chaque unité que compte la population, la variable auxiliaire a été générée selon une loi gamma de moyenne 48 et de variance 768. La variable d'intérêt a été générée conditionnellement à x selon une loi gamma de moyenne $1,5x$ et de variance $16x$. On a attribué de façon aléatoire une valeur manquante de x à la moitié de la population. Étant donné qu'aucun domaine d'intérêt n'a été généré, θ correspond au total de population de la variable y.

Nous avons sélectionné 10 000 échantillons à partir de cette population par échantillonnage aléatoire simple sans remise. Nous avons considéré deux tailles d'échantillon : $n = 100$, et $n = 250$. Pour chaque échantillon, la non-réponse associée à la variable y a été générée indépendamment d'une unité à l'autre, la probabilité de non-réponse étant de 0,3. Nous avons utilisé la même stratégie d'imputation que dans l'exemple de la section 2, avec $\omega_l^{(1)} = 1$, pour $l \in s_r^{(1)}$, et $\omega_l^{(2)} = 1$, pour $l \in s_r$. Les non-répondants pour la variable y avec valeur de x observée ont été imputés au moyen de la méthode d'imputation par le ratio, l'imputation par la moyenne étant utilisée lorsque la valeur de x était manquante.

Les valeurs de y de la population sont demeurées fixes tout au long des répétitions de l'expérience par simulation ; chaque répétition consistait à sélectionner un échantillon, puis à générer la non-réponse à l'égard de la variable y. Si nous nous en étions tenus strictement au développement théorique exposé à la section 5, nous aurions généré de nouvelles valeurs de y lors de chaque répétition conformément au modèle d'imputation. Il est toutefois plus courant dans la littérature de fixer les valeurs de y de la population lorsque l'on procède à une expérience par simulation. À titre d'exemple, nos conditions de simulation sont pour l'essentiel les mêmes que celles commentées par Rancourt, Lee et Särndal (1993), qui se sont aussi penchés sur l'imputation composite.

Nous avons calculé la variance d'échantillonnage Monte Carlo et l'EQM globale, $V_{\mathrm{SAM}}^{\mathrm{MC}} = \sum_{r=1}^{R}(\hat{\theta}_r - \theta)^2 / R$ et $V_{\mathrm{TOT}}^{\mathrm{MC}} = \sum_{r=1}^{R}(\hat{\theta}_{I,\,r} - \theta)^2 / R$, respectivement, où l'indice r indique que les estimations sont fondées sur la r^{e} répétition, et $R = 10\,000$. Le biais relatif Monte Carlo de tout estimateur de V_{SAM}, par exemple v_{SAM}, est calculé ainsi : $\mathrm{RB}(V_{\mathrm{SAM}}) = \sum_{r=1}^{R}(v_{\mathrm{SAM},r} - V_{\mathrm{SAM}}^{\mathrm{MC}}) / (V_{\mathrm{SAM}}^{\mathrm{MC}} R)$. De même, nous avons calculé le biais relatif Monte Carlo d'un estimateur de V_{TOT}, désigné comme étant $\mathrm{RB}(V_{\mathrm{TOT}})$, et celui d'un estimateur de $V_{\mathrm{SAM}} / V_{\mathrm{TOT}}$, soit $\mathrm{RB}(V_{\mathrm{SAM}} / V_{\mathrm{TOT}})$. Enfin, nous avons calculé les taux de couverture de Monte Carlo des intervalles de confiance à 95 % pour θ, en supposant une distribution normale de $\hat{\theta}_I$.

Les résultats de notre étude par simulation sont présentés au tableau 2. Dans les colonnes identifiées SEVANI, la variance d'échantillonnage, V_{SAM}, et l'EQM globale, V_{TOT}, sont estimées pour chaque échantillon au moyen de $\hat{V}_{\mathrm{SAM}}^{\mathrm{C}}$ et de $\hat{V}_{\mathrm{TOT,\,ADJ}}$, respectivement (se reporter à la section 5.4). Nous avons également obtenu des résultats en remplaçant $\hat{V}_{\mathrm{TOT,\,ADJ}}$ par $\hat{V}_{\mathrm{TOT}}$. Nous ne les présentons toutefois pas dans le tableau 2, car ils étaient très près de ceux obtenus avec $\hat{V}_{\mathrm{TOT,\,ADJ}}$. Cela donne à penser que le biais par rapport au modèle, B_m, n'est pas important ici. Dans les colonnes Naïf, tant la variance d'échantillonnage que l'EQM globale sont estimées au moyen de $\hat{V}_{\mathrm{ORD}}$ (se reporter à la section 5.1).

Tableau 2
Résultats de l'étude par simulation

	$n = 100$		$n = 250$	
	SEVANI	**Naïf**	**SEVANI**	**Naïf**
$\mathrm{RB}(V_{\mathrm{SAM}})$	2,82 %	-17,59 %	3,02 %	-17,68 %
$\mathrm{RB}(V_{\mathrm{SAM}}/V_{\mathrm{TOT}})$	8,30 %	-	5,84 %	-
$\mathrm{RB}(V_{\mathrm{TOT}})$	-5,07 %	-40,68 %	-2,66 %	-52,89 %
Taux de couverture	93,38 %	86,20 %	94,42 %	81,80 %

Ces résultats montrent que la méthodologie décrite à la section 5 et exécutée dans le SEVANI est plus efficace que l'estimateur naïf de la variance aux fins d'estimer les composantes de la variance et d'établir des intervalles de confiance. L'utilisation du SEVANI donne de petits biais relatifs Monte Carlo et des taux de couverture proches du taux nominal cible (95 %). Notre méthodologie est aussi utile aux utilisateurs voulant estimer la part de l'EQM globale attribuable à la variance d'échantillonnage, c'est-à-dire $V_{\mathrm{SAM}} / V_{\mathrm{TOT}}$. Il est à noter que $V_{\mathrm{SAM}}^{\mathrm{MC}} / V_{\mathrm{TOT}}^{\mathrm{MC}}$ est de 71,98 % pour $n = 100$ et de 57,23 % pour $n = 250$. Étant

donné que $V_{\text{SAM}}^{\text{MC}} / V_{\text{TOT}}^{\text{MC}}$ n'est pas proche de 100 %, même pour $n = 100$, les effets de la non-réponse et de l'imputation ne peuvent être systématiquement omis lorsque l'on estime l'EQM globale.

7. L'approche inverse

Shao et Steel (1999) ont proposé une approche inverse d'estimation de la variance, élaborée pour les situations où l'on a recours à l'imputation composite. Ils ont fait l'hypothèse que le biais global est négligeable et ont mis de l'avant la décomposition suivante de la variance globale :

$$E_{mpq}(\hat{\theta}_I - \theta)^2 = E_{mq}\text{var}_p(\hat{\theta}_I \mid U_r) + E_{mq}\{E_p(\hat{\theta}_I \mid U_r) - \theta\}^2, \qquad (7.1)$$

où U_r est une population conceptuelle de répondants. Dans le membre droit de l'expression (7.1), l'espérance et la variance intérieures sont déterminées par rapport au plan d'échantillonnage. Malheureusement, l'estimateur imputé $\hat{\theta}_I$ ne sera généralement pas linéaire par rapport à ce plan, même s'il l'est par rapport aux valeurs de *y* observées. Par conséquent, l'estimateur imputé $\hat{\theta}_I$ est généralement linéarisé (se reporter par exemple à Shao et Steel 1999, ainsi qu'à Kim et Rao 2009). De manière plus explicite, les quantités $\varphi_{0k}^{(j)}$ et $\varphi_{lk}^{(j)}$ dépendent souvent de l'échantillon, et ce, de façon non linéaire ; c'est notamment le cas pour l'imputation par la régression linéaire (se reporter à l'exemple présenté à la fin de la section 3) et pour l'imputation par donneur. Il n'est pas toujours facile de prendre en compte la variabilité de $\varphi_{0k}^{(j)}$ et de $\varphi_{lk}^{(j)}$ lorsque l'on utilise (7.1). Par exemple, il n'y a pas d'articles portant sur l'utilisation de l'approche inverse pour estimer la variance lorsque l'on a recours à l'imputation par le plus proche donneur. De plus, étant donné que chaque stratégie d'imputation composite produit son propre estimateur imputé linéarisé, il est ardu de mettre en oeuvre cette méthodologie dans un progiciel généralisé.

Au moyen de notre approche, l'espérance intérieure dans les expressions relatives à la variance due à la non-réponse,

$$V_{\text{NR}} = E_{pq}E_m\{(\hat{\theta}_I - \hat{\theta})^2 \mid s, s_r\},$$

et à la composante mixte,

$$V_{\text{MIX}} = 2E_{pq}E_m\{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) \mid s, s_r\},$$

est calculée par rapport au modèle d'imputation (conditionnellement à *s* et s_r). L'estimateur imputé est linéaire et les calculs sont simples parce que les quantités $\varphi_{0k}^{(j)}$ et $\varphi_{lk}^{(j)}$ sont établies sans faire intervenir les valeurs de *y*. Ces deux quantités n'entrent pas dans l'estimation de la variance d'échantillonnage, $V_{\text{SAM}} = E_m\,\text{var}_p(\hat{\theta})$ (se reporter à l'équation 5.5), de sorte que leur éventuelle non-linéarité par rapport au plan d'échantillonnage ne pose aucun problème. Cela veut dire que l'imputation par le plus proche donneur peut être traitée facilement avec notre approche (se reporter à Beaumont et Bocci 2009).

Ce sont toutes ces raisons qui nous amènent à penser que l'approche inverse risque d'être plus fastidieuse à mettre en œuvre dans un progiciel généralisé que notre approche. Cela ne veut pas dire que l'approche inverse est inutile. Dans les faits, les deux approches aboutissent à des estimateurs de variance identiques lorsqu'un recensement est effectué. Beaumont, Haziza et Bocci (2011) ont montré que l'une et l'autre approches donnent également des estimateurs de variance identiques lorsque l'on utilise l'imputation par valeur auxiliaire (étant donné que $\varphi_{0k}^{(j)}$ et $\varphi_{lk}^{(j)}$ ne dépendent pas de *s* et s_r). Les deux approches dépendent d'une spécification correcte du modèle d'imputation, et aucune des deux ne devrait donner systématiquement de meilleurs résultats que l'autre.

L'approche inverse peut avoir un avantage pratique comparativement à la nôtre quand la fraction de sondage est négligeable. Dans un tel cas, Shao et Steel (1999) montrent que la deuxième composante du membre droit de (7.1) peut être négligée. Pour estimer la première composante, on détermine un estimateur fondé sur le plan de sondage pour $\text{var}_p(\hat{\theta}_I \mid U_r)$. Si l'on choisit une méthode d'estimation de la variance par réplication (comme le jackknife ou le bootstrap) pour estimer $\text{var}_p(\hat{\theta}_I \mid U_r)$, l'approche dans son ensemble devient fort intéressante et pratique. En outre, elle ne dépend pas de la validité du modèle d'imputation, en particulier la spécification correcte de la variance sous le modèle σ_k^2. Les estimateurs de la variance par le jackknife de Rancourt, Lee et Särndal (1993) et de Sitter et Rao (1997) peuvent être justifiés par cette approche.

8. Conclusion

Notre méthodologie relative à l'imputation composite a été mise en oeuvre dans la version 2 du SEVANI en raison de sa facilité d'exécution et de sa généralité. Elle fonctionne pour la plupart des méthodes d'imputation utilisées dans la pratique, car la grande majorité de ces méthodes sont linéaires. Les calculs de variance sont les mêmes pour toutes les stratégies d'imputation composite, une fois que l'on a calculé les quantités $W_{0d}^{(+)}$, $W_{dk}^{(+)}$, $\hat{\mu}_k$ et $\hat{\sigma}_k^2$. Il est ainsi plus facile de procéder au développement d'un système généralisé.

Nous avons mis l'accent sur l'estimation d'un total de domaine au moyen de l'estimateur de Horvitz-Thompson, mais le SEVANI peut aussi être utilisé relativement à des estimateurs de moyennes de domaine et à des estimateurs par calage. Il y a aussi des méthodes paramétriques et non paramétriques d'estimation de μ_k et σ_k^2. On trouvera de plus amples détails dans le guide méthodologique du

SEVANI (Beaumont, Bissonnette et Bocci 2010), que l'on peut obtenir sur demande auprès des auteurs.

Remerciements

Nous tenons à remercier les réviseurs de leurs commentaires. Nous remercions également Mike Hidiroglou, Eric Rancourt et Cynthia Bocci de Statistique Canada de leurs suggestions, sans oublier les discussions tenues sur le sujet. Tous ces commentaires ont servi à peaufiner notre article.

Bibliographie

Beaumont, J.-F., Bissonnette, J. et Bocci, C. (2010). SEVANI, version 2.3, Guide de méthodologie. Rapport interne, Direction de la méthodologie, Statistique Canada.

Beaumont, J.-F., et Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canadian Journal of Statistics*, 37, 400-416.

Beaumont, J.-F., Haziza, D. et Bocci, C. (2011). On variance estimation under auxiliary value imputation in sample surveys. *Statistica Sinica*, 21, 515-537.

Brick, J.M., Kalton, G. et Kim, J.K. (2004). Estimation de variance pour l'imputation hot deck à l'aide d'un modèle. *Techniques d'enquête*, 30, 63-72.

Deville, J.-C., et Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 10, 381-394.

Felx, P., et Rancourt, E. (2001). Applications de la variance due à l'imputation dans l'Enquête sur l'emploi, la rémunération et les heures. Document de travail, Direction de la méthodologie, Statistique Canada, BSMD-2001-009E.

Haziza, D. (2009). Imputation and inference in the presence of missing data. Dans *Handbook of Statistics, Sample Surveys: Theory, Methods and Inference*, (Éds., D. Pfeffermann et C.R. Rao). Amsterdam : Elsevier BV, 29A, 215-246.

Hidiroglou, M.A. (1989). Notes manuscrites non publiées généreusement transmises par l'auteur.

Kim, J.-K., et Rao, J.N.K. (2009). Unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96, 917-932.

Lee, H., Rancourt, E. et Särndal, C.-E. (2001). Variance estimation from survey data under single imputation. Dans *Survey Nonresponse*, (Éds., R.M. Groves, D.A. Dillman, J.L. Eltinge et R.J.A. Little). New-York : John Wiley & Sons, Inc., 315-328.

Rancourt, E., Lee, H. et Särndal, C.-E. (1993). Variance estimation under more than one imputation method. Dans *Proceedings of the International Conference on Establishments Surveys*, juin 1993, Buffalo, American Statistical Association, 374-379.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New-York : John Wiley & Sons, Inc.

Särndal, C.-E. (1992). Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation. *Techniques d'enquête*, 18, 257-268.

Shao, J., et Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

Sitter, R.R., et Rao, J.N.K. (1997). Imputation for missing values and corresponding variance estimation. *Canadian Journal of Statistics*, 25, 61-73.