Recueil du Symposium 2001 de Statistique Canada
La qualité des données d'un organisme statistique : une perspective méthodologique

IMPUTATION HOT-DECK POUR LE MODÈLE DE RÉPONSE

Wayne A. Fuller[1] et Jae Kwang Kim[2]

RÉSUMÉ

L'échantillonnage d'enquête a souvent recours à l'imputation hot-deck, méthode qui consiste à remplacer les valeurs manquantes par des valeurs fournies par les répondants. Un modèle propice à ce genre de méthode est celui selon lequel on présume que les probabilités de réponse sont égales à l'intérieur des ccegalitéà lr'i.2(r)10ti0.9(é)17.9(é)1.

2. MODÈLE DE RÉPONSE POUR IMPUTATION

Prenons une population de N éléments identifiés par un ensemble d'indices $U = \{1, 2, \ldots, N\}$. À chaque unité i de la population correspond une variable d'étude y_i et un vecteur $\mathbf{x}_i$ de renseignements auxiliaires. L'ensemble de vecteurs, $(y_i, \mathbf{x}_i)$, $i = 1, 2, \ldots, N$, est représenté par F.

Supposons que A représente les indices des éléments d'un échantillon, choisis selon un ensemble de règles probabilistes appelé *mécanisme d'échantillonnage*. La fonction de l'indicateur d'échantillonnage pour l'élément j est la suivante :

$$I_j = \begin{cases} 1 & \text{si } j \in A \\ 0 & \text{si } j \notin A \end{cases}, \qquad j = 1, ..., N \tag{1}$$

et le vecteur lié aux indicateurs est $\mathbf{B} = (I_1, \ldots, I_N)$. Supposons que la quantité dans la population qui nous intéresse est $\theta_N = \theta\ (y_1, y_2, \ldots, y_N)$ et que $\hat{\theta}$ est un estimateur linéaire de θ_N fondé sur l'ensemble de l'échantillon,

$$\hat{\theta} = \sum_{i \in A} w_i y_i \,. \tag{2}$$

Si w_i est l'inverse de la probabilité de sélection, alors (2) est sans biais pour la population totale.

Supposons que A_R et A_M représentent respectivement l'ensemble d'indices des répondants et celui des non-répondants de l'échantillon. Définissons la fonction de l'indicateur de réponse

$$R_i = \begin{cases} 1 & \text{si } i \in A_R \\ 0 & \text{si } i \in A_M \end{cases} \tag{3}$$

et supposons que $\mathbf{R} = \{(i, R_i);\ i \in A\}$. La distribution de $\mathbf{R}$ est appelée *mécanisme de réponse*.

Supposons que la population finie U se compose de G cellules d'imputation. L'ensemble d'éléments compris dans la cellule g est U_g. Supposons que n_g est le nombre d'éléments de l'échantillon compris dans la cellule d'imputation g et que r_g, $r_g > 0$ est le nombre de répondants compris dans cette même cellule. Supposons maintenant le modèle de réponse uniforme intra-cellule selon lequel les réponses r_g comprises dans une cellule équivalent à un échantillon de Poisson choisi avec des probabilités égales parmi les n_g éléments.

Selon la méthode d'imputation, supposons que d_{ij} est le nombre de fois que Y_i sert de donateur pour les valeurs manquantes de Y_j et définissons $\mathbf{d} = \{ d_{ij};\ i \in A_R,\ j \in A_M \}$. La distribution de $\mathbf{d}$ est appelée *mécanisme d'imputation*. Supposons que w_{ij}^* est le facteur appliqué au poids initial pour l'élément j lorsque Y_i sert de donateur pour l'élément j. Pour l'élément j, $j \in A_M$,

$$Y_{Ij} = \sum_{i \in A_R} d_{ij} w_{ij}^* y_i \tag{4}$$

est la moyenne pondérée des valeurs imputées. Le facteur w_{ij}^* est appelé *fraction d'imputation*. Il s'agit de la fraction que le donateur i donne pour la valeur manquante Y_j. Il convient de noter que $w_{ii}^* = 1$ pour $i \in A_R$ et $w_{ii}^* = 0$ pour $i \in A_M$. Les valeurs de d_{ij} sont des entiers naturels et la somme des fractions d'imputation des donateurs pour une valeur manquante est limitée à un, soit :

$$\sum_{i \in A_R} d_{ij} w_{ij}^* = 1, \qquad \forall j \in A\,. \tag{5}$$

Un estimateur calculé selon l'équation (4) et ayant certaines valeurs de $w_{ij}^* < 1$ est appelé *estimateur obtenu par imputation fractionnelle*.

Un estimateur linéaire obtenu par imputation hot-deck peut être formulé comme suit :

$$\hat{\theta}_I = \sum_{i \in A_R} \left(\sum_{i \in A} d_{ij} w_j w_{ij}^* \right) \tag{6}$$

$$=: \sum_{i \in A} \alpha_i Y_i \,, \tag{7}$$

où la notation $A =: B$ signifie que B est défini comme étant égal à A. La somme de $w_{ij}^* w_j$ pour tous les receveurs dont i est un donateur (y compris lorsqu'il sert de donateur pour lui-même), représentée par α_i, correspond au poids total du donateur i. Si une unité répondante i ne sert pas de donateur, sauf pour elle-même, alors $\alpha_i = w_i$.

3. IMPUTATION FRACTIONNELLE PLEINEMENT EFFICACE

Supposons que tous les éléments d'une cellule d'imputation présentent la même probabilité de réponse et que les réponses sont indépendantes. On peut alors obtenir la distribution globale d'un estimateur imputé selon le modèle de réponse en utilisant la structure probabiliste de l'échantillonnage à phases multiples, où le modèle de réponse est considéré comme le mécanisme d'échantillonnage de deuxième phase.

Un estimateur imputé $\hat{\theta}_I$ pour l'estimateur de l'échantillon complet $\hat{\theta}$ est appelé conditionnellement *sans biais conditionnel* pour $\hat{\theta}$ selon le modèle de réponse s'il satisfait l'équation

$$E(\hat{\theta}_I - \hat{\theta} \,|\, \mathsf{F}, A) = 0 \,, \tag{8}$$

où l'espérance de (8) est calculée à partir de la jointe définie par le modèle de réponse et le mécanisme d'imputation, étant donné l'échantillon réalisé A et la population finie fixe F. Pour qu'un estimateur imputé selon l'équation (7) soit conditionnellement sans biais selon le modèle de réponse présumé, une condition nécessaire et suffisante est:

$$E(R_i - \alpha_i \,|\, \mathsf{F}, A) = w_i \,, \tag{9}$$

où α_i est le poids total du donateur i défini en (7).

Supposons que $\pi_{i2} = \Pr(i \in A_R \,|\, i \in A)$. Si π_{i2} est connu et supérieur à 0, alors une méthode d'imputation hot-deck satisfaisant l'équation

$$E(\alpha_i \,|\, A, A_R) = w_i \pi_{i2}^{-1} \tag{10}$$

produit un estimateur sans biais pour la population totale. Plus précisément, l'imputation hot-deck où $\alpha_i = w_i \pi_{i2}^{-1}$ pour toutes les valeurs de $i \in A_R$ produit l'estimateur imputé le plus efficace lorsqu'il s'agit de minimiser la variance due à l'imputation pour les estimateurs qui sont sans biais selon le modèle de réponse.

Si π_{i2} est inconnu et que les probabilités de réponse dans une cellule sont uniformes, alors un estimateur raisonnable du total est la somme pondérée des estimateurs obtenus par la méthode des ratios

$$\hat{\theta}_{FE} = \sum_{g=1}^{G} \left(\sum_{i \in A \cap U_g} w_i \right) \frac{\sum_{i \in A_R \cap U_g} w_i y_i}{\sum_{i \in A \cap U_g} w_i} \,, \tag{11}$$

où w_i est proportionnel à l'inverse de la probabilité de sélection. L'estimateur (11) est dit pleinement efficace parce qu'il ne présente aucune variabilité due à la sélection aléatoire des donateurs. Si les valeurs de w_i sont les mêmes pour tous les éléments d'une cellule, le ratio

$$\left(\sum_{i \in A_R \cap U_g} w_i \right)^{-1} \sum_{i \in A_R \cap U_g} w_i y_i \tag{12}$$

est une simple moyenne et, par conséquent, sans biais pour la moyenne de la cellule à condition qu'il y ait au moins un répondant dans la cellule. Si les valeurs de w_i dans une cellule ne sont pas égales, alors (12)

risque de présenter un biais de ratio. Il est possible que le nombre d'éléments d'une cellule, n_g, soit positif et que le nombre de répondants, r_g, soit nul. Si tel est le cas dans la pratique, les cellules sont alors groupées. Pour définir systématiquement un estimateur ayant des moments finis, nous ramenons le ratio énoncé en (12) à zéro et nous ramenons $(\sum_{i\in A_R\cap U_g} w_i)^{-1} \sum_{i\in A\cap U_g} w_i$ à un lorsque $r_g = 0$.

La variance approximative de l'estimateur (11), donnée en (19) du théorème 3.1, est la variance de l'estimateur de l'échantillon complet, plus un terme qui dépend des probabilités de réponse de la cellule et des variances intra-cellule.

Théorème 3.1

Supposons une suite de populations finies à indice υ, de taille N_υ, où $N_\upsilon > N_{\upsilon-1}$. Supposons ensuite que y est une caractéristique de la population ayant des quatrièmes moments bornés et qu'on choisit un échantillon de taille $n_\upsilon \geq n_{\upsilon-1}$ parmi la υ^e population ayant des probabilités de sélection connues. Supposons aussi que la population est décomposée en cellules G_υ, $G_\upsilon \geq G_{\upsilon-1}$, mutuellement exclusives et exhaustives, que la taille de la population de la cellule g est $N_{g\upsilon}$, que celle de l'échantillon compris dans la cellule g est $n_{g\upsilon}$ et que le nombre de répondants compris dans la cellule g est $r_{g\upsilon}$. Supposons que

$$K_{SL} G_\upsilon^{-1} \leq N_\upsilon^{-1} N_{g\upsilon} \leq K_{SU} G_\upsilon^{-1} \quad \textit{pour toutes les valeurs de } \upsilon, \tag{13}$$

$$G_\upsilon < K_G n_\upsilon^{\lambda} \quad \textit{pour toutes les valeurs de } \upsilon, \tag{14}$$

$$K_{wL} \leq n_\upsilon w_i \leq K_{wU} \quad \textit{pour toutes les valeurs de } \upsilon, \tag{15}$$

$$K_\pi \leq \pi_{g\upsilon} \quad \textit{pour toutes les valeurs de } g \textit{ et de } \upsilon, \tag{16}$$

où K_π, K_{SL}, K_{SU}, K_G, K_{wL} et K_{wU} sont des constantes positives fixes, $0 \leq \lambda < 0.5$, $\pi_{g\upsilon}$ est la probabilité de réponse commune dans la cellule g de la population υ, et les valeurs de w_i sont les poids de (2) pour l'estimateur de la moyenne. Supposons que l'estimateur basé sur l'ensemble de l'échantillonnage $\hat{\theta}_\upsilon$ est sans biais pour la moyenne de la population finie.

Supposons que

$$V\{\hat{\theta}_\upsilon \mid \mathsf{F}_\upsilon\} < K_M V\{\hat{\theta}_{SRS,\upsilon} \mid \mathsf{F}\} \tag{17}$$

pour une valeur fixe de K_M pour n'importe quelle valeur de y ayant des quatrièmes moments bornés et que $\hat{\theta}_{SRS,\upsilon}$ est l'estimateur de θ en fonction d'un échantillon aléatoire simple de taille n_υ. Supposons que pour chaque valeur de $i \neq j = 1, 2, \ldots, N_\upsilon$,

$$P(R_{i\upsilon} = 1, R_{j\upsilon} = 1) = P(R_{i\upsilon} = 1)\, P(R_{j\upsilon} = 1). \tag{18}$$

Alors,

$$\hat{\theta}_{FE\upsilon} - \hat{\theta}_\upsilon = \sum_{g_\upsilon=1}^{G_\upsilon} \sum_{i\in A_{g\upsilon}} w_{i\upsilon} (\pi_{g\upsilon}^{-1} R_{i\upsilon} - 1) e_{i\upsilon} + o_p(n_\upsilon^{-1/2}) .$$

De plus,

$$V(\tilde{\theta}_{FE\upsilon} \mid \mathsf{F}_\upsilon) = V(\hat{\theta}_\upsilon \mid \mathsf{F}_\upsilon) + E\left\{ \sum_{g_\upsilon=1}^{G_\upsilon} \pi_{g\upsilon}^{-1}(1-\pi_{g\upsilon}) \sum_{i\in A_{g\upsilon}} w_{i\upsilon}^2 e_{i\upsilon}^2 \mid \mathsf{F}_\upsilon \right\}, \tag{19}$$

où

$$\tilde{\theta}_{FE\upsilon} = \hat{\theta}_\upsilon + \sum_{g_\upsilon=1}^{G_\upsilon} \sum_{i\in A_{g\upsilon}} w_{i\upsilon} (\pi_{g\upsilon}^{-1} R_{i\upsilon} - 1) e_{i\upsilon} ,$$

F_υ est la υ^e population, $A_{g\upsilon}$ est l'ensemble d'indices d'échantillonnage dans la g^e cellule pour le υ^e échantillon, $e_{i\upsilon} = y_{i\upsilon} - \bar{Y}_{g\upsilon}$, et $\bar{Y}_{g\upsilon}$ est la moyenne de la population de la variable y comprise dans la cellule g_υ de la population F_υ.

On peut mettre en œuvre l'estimateur (11) au moyen d'une imputation fractionnelle dans laquelle chaque unité répondante d'une cellule d'imputation sert de donateur pour chaque non-répondant compris dans la cellule, et le poids d'imputation est proportionnel au poids d'échantillonnage. Alors, l'estimateur (11) peut être formulé comme suit : l'estimateur obtenu par imputation fractionnelle

$$\hat{\theta}_{FEFE} = \sum_{g=1}^{G} \sum_{j \in A \cap U_g} \sum_{i \in A_R \cap U_g} w_j w_{ij}^* y_i \ , \tag{20}$$

où $w_j w_{ij}^*$ est le poids du donateur *i* pour le receveur *j*, w_{ij}^* est la fraction d'imputation du donateur *i* pour le receveur *j* défini en (4), et

$$w_{ij}^* = \begin{cases} (\sum_{s \in A_R \cap U_g} w_s)^{-1} w_i R_i & \text{si } R_j = 0 \\ 1 & \text{si } R_j = 1 \text{ et } i = j \ . \end{cases} \tag{21}$$

L'estimateur (20) avec w_{ij}^* de (21), équivalent algébrique de (11), est appelé estimateur *pleinement efficace obtenu par imputation fractionnelle* (en anglais : FEFI).

Considérons maintenant l'estimation de la variance. En fonction d'une réponse complète, supposons qu'un estimateur de la variance par répétition est

$$\hat{V}(\hat{\theta}) = \sum_{k=1}^{L} c_k (\hat{\theta}^{(k)} - \hat{\theta})^2 \ , \tag{22}$$

où $\hat{\theta}^{(k)}$ est la k^e estimation de θ_N en fonction des observations comprises dans la k^e répétition, *L* est le nombre de répétitions et c_k est un facteur lié à la répétition *k* déterminé selon la méthode de répétition. Lorsque l'estimateur initial $\hat{\theta}$ est un estimateur linéaire selon l'équation (2), la k^e répétition de $\hat{\theta}$ peut être formulée comme suit :

$$\hat{\theta}^{(k)} = \sum_{i \in A} w_i^{(k)} y_i \ , \tag{23}$$

où $w_i^{(k)}$ représente le poids de répétition pour la i^e unité de la k^e répétition.

Nous proposons, pour l'estimateur *FEFI* $\hat{\theta}_{FEFI}$, la répétition suivante :

$$\begin{aligned} \hat{\theta}_{FEFI}^{(k)} &= \sum_{g=1}^{G} \left(\sum_{i \in A \cap U_g} w_i^{(k)} \right) \frac{\sum_{i \in A_R \cap U_g} w_i^{(k)} y_i}{\sum_{i \in A_R \cap U_g} w_i^{(k)}} \\ &=: \sum_{g=1}^{G} \sum_{i \in A \cap U_g} \sum_{i \in A_R \cap U_g} w_j^{(k)} w_{ij}^{*(k)} y_i \end{aligned} \tag{24}$$

où $w_i^{(k)}$ est le poids de répétition de l'échantillon complet de l'unité *i* et $w_{ij}^{*(k)} = (\sum_{s \in A_R \cap U_g} w_s^{(k)})^{-1} w_i^{(k)}$.

À partir des répétitions (24), l'estimateur de la variance par répétition peut être formulé comme suit :

$$\hat{V}_{FEFI} = \sum_{k=1}^{L} c_k (\hat{\theta}_{FEFI}^{(k)} - \hat{\theta}_{FEFI})^2 \ . \tag{25}$$

Il convient de noter que la somme des poids de répétition des dossiers de chaque receveur est la même que le poids de répétition pour cette unité comprise dans un échantillon complet.

La méthode proposée s'apparente à celle de l'estimateur de la variance pour imputation simple de Rao et Shao (1992). Voir également Yung et Rao (2000). Comme l'équation (24) utilise des répétitions

fractionnelles, l'estimateur (25) est pertinent pour un vecteur de variables *y*. Une fois calculés, les poids de répétition sont pertinents pour toute fonction continue du vecteur *y*. L'estimateur de la variance est convergent.

Théorème 3.2 *Supposons que les hypothèses du théorème 3.1 sont valables. Supposons ensuite que l'estimateur de la variance par répétition pour l'échantillon complet est calculé selon l'équation (22) et que les répétitions satisfont l'équation*

$$\left| \hat{\gamma}_\upsilon^{(k)} - \hat{\gamma}_\upsilon \right|^2 \; < \; \zeta_k^2 n_\upsilon^{-1} V\{\hat{\gamma}_\upsilon\} \tag{26}$$

pour toutes les valeurs de k, où ζ_k *sont des variables aléatoires ayant des quatrièmes moments bornés. Supposons enfin que* $\hat{\theta}_\upsilon$ *est l'estimateur, basé sur de l'échantillon complet, de la moyenne et que l'estimateur, basé sur de l'échantillon complet, de la variance de* $\hat{\theta}_\upsilon$*, représenté par* $\hat{V}(\hat{\theta}_\upsilon)$*, satisfait l'équation*

$$E\left\{ \left[\hat{V}(\hat{\theta}_\upsilon) - V(\hat{\theta}_\upsilon \,|\, \mathsf{F}_\upsilon) \right]^2 \,|\, \mathsf{F}_\upsilon \right\} \; = \; o(n_\upsilon^{-2}) \tag{27}$$

pour toute variable de y ayant des quatrièmes moments bornés. Alors, pour une variable y ayant des quatrièmes moments bornés, l'estimateur de la variance défini en (25) pour une moyenne satisfait l'équation

$$\hat{V}_{FEFI} \; = \; V\,(\,\tilde{\theta}_{FE\upsilon} \,|\, \mathsf{F}_\upsilon) - N_\upsilon^{-2} \sum_{g_\upsilon=1}^{G_\upsilon} \sum_{i \in U_{g\upsilon}} \pi_{g\upsilon}^{-1} \,(1-\pi_{g\upsilon})\, e_{i\upsilon}^2 + o_p\,(\,n_\upsilon^{-1})\,, \tag{28}$$

où $\tilde{\theta}_{FE\upsilon}$ *est défini dans le théorème 3.1, et la répartition est établie à l'égard des mécanismes d'échantillonnage et de réponse.*

Si l'on peut faire abstraction de la correction pour la population finie, l'estimateur (25) est convergent pour $V\{\hat{\theta}\}$. Si la taille de l'échantillon est importante par rapport à *N*, alors un estimateur de

$$N_\upsilon^{-2} \sum_{i=g_\upsilon}^{G_\upsilon} \sum_{i \in U_{g\upsilon}} \pi_{g\upsilon}^{-1} \,(1-\pi_{g\upsilon})\, e_{i\upsilon}^2$$

doit être ajouté à (25). Un estimateur est

$$N^{-1} \sum_{g=1}^{G} \left(\sum_{i \in A_g} w_i \right) \hat{\pi}_g^{-1} (1-\hat{\pi}_g)\,(\,r_g - 1)^{-1} \sum_{i \in A_R \cap U_g} (y_i - \bar{y}_{sg})^2\,, \tag{29}$$

où $\hat{\pi}_g = n_g^{-1} r_g$ et $\bar{y}_{sg} = r_g^{-1} \sum_{i \in A_R \cap U_g} y_i$.

On peut construire l'estimateur (29) directement ou au moyen de répétitions.

4. LE MODÈLE DE LA MOYENNE DES CELLULES

Les méthodes d'imputation et d'estimation de la variance énoncées pour le modèle de réponse produisent également des estimateurs convergants pour le modèle de la moyenne des cellules. Selon ce modèle, les éléments compris dans une cellule de la population finie constituent la réalisation de variables aléatoires distribuées de manière indépendante et identique avec une moyenne μ_g et une variance σ_g^2. Ainsi, pour le modèle de la moyenne des cellules

$$Y_i \overset{i.i.}{\sim} (\mu_g, \sigma_g^2), \qquad i \in U_g\,, \tag{30}$$

où U_g représente l'ensemble d'indices pour la g^{e} cellule d'imputation et $\overset{i.i.}{\sim}$ est l'abréviation de *distribué de manière indépendante et identique*. La méthode d'imputation fondée sur le modèle de réponse n'est pas

nécessairement pleinement efficace pour la moyenne de la population selon le modèle de la moyenne des cellules, mais l'estimateur de la moyenne et celui de la variance de la moyenne estimée sont convergents.

La distribution de Y dans l'échantillon est déterminée par le mécanisme d'échantillonnage et par la distribution du vecteur Y. Si la distribution de Y est indépendante du mécanisme d'échantillonnage, on dit que le mécanisme d'échantillonnage est *négligeable*. Le mécanisme de réponse est *négligeable* si la distribution conditionnelle de Y en fonction du résultat A et de l'ensemble de répondants A_R, est la même que la distribution conditionnelle de Y en fonction de A.

Si le mécanisme d'échantillonnage et le mécanisme de réponse sont négligeables, alors le modèle de la moyenne des cellules est valable tant pour les unités répondantes que pour les non-répondants. Ainsi,

$$Y_i \mid (A, A_R) \overset{i.i.}{\sim} (\mu_g, \sigma_g^2) \qquad i \in U_g \,. \tag{31}$$

On peut considérer les conclusions des théorèmes 3.1 et 3.2 comme des résultats conditionnels pour une population finie donnée. Si le modèle de la moyenne des cellules est valable, on peut assouplir l'hypothèse d'une probabilité de réponse commune.

Théorème 4.1 *Supposons que les hypothèses du théorème 3.1 sont valables à l'exception de (16) et qu'il existe quelques valeurs de $K_\pi > 0$ en vertu desquelles*

$$K_\pi \; < \; \pi_{g\upsilon i} \tag{32}$$

pour toutes les valeurs de g et de υ, où $\pi_{g\upsilon i}$ est la probabilité de réponse de l'élément i dans la population υ. Supposons que le modèle de la moyenne des cellules (31) est valable. Il existe alors une suite de variables aléatoires $\tilde{\theta}_{FE\upsilon}$ avec

$$p \lim_{\upsilon \to \infty} n_\upsilon^{1/2} (\hat{\theta}_{FE\upsilon} - \tilde{\theta}_{FE\upsilon}) \; = \; 0$$

et

$$V(\tilde{\theta}_{FE\upsilon}) \; = \; V(\hat{\theta}_\upsilon) + E\left\{ \sum_{g_\upsilon=1}^{G_\upsilon} \sum_{i \in A_{g\upsilon}} \tilde{\pi}_{g\upsilon}^{-2} \pi_{g\upsilon i} (1 - \pi_{g\upsilon i}) w_{i\upsilon}^2 \sigma_{g\upsilon}^2 \right\}, \tag{33}$$

où

$$\tilde{\pi}_{g\upsilon} \; = \left(\sum_{i \in A_{g\upsilon}} w_{i\upsilon} \right)^{-1} \sum_{i \in A_{g\upsilon}} w_{i\upsilon} \pi_{g\upsilon i} \,.$$

Théorème 4.2 *Supposons que les hypothèses du théorème 4.1 sont valables. Supposons ensuite que l'estimateur de la variance par répétition pour l'échantillon complet est calculé selon l'équation (22) et qu'il satisfait les hypothèses du théorème 3.2. Alors, l'estimateur de la variance pour une moyenne définie selon l'équation (25) satisfait l'équation*

$$\hat{V}_{FEFI} \; = \; V(\tilde{\theta}_{FE\upsilon} \mid \mathsf{F}_\upsilon) - N_\upsilon^2 \sum_{g_\upsilon=1}^{G_\upsilon} \sum_{i \in U_{g\upsilon}} \tilde{\pi}_{g\upsilon}^{-2} \pi_{g\upsilon i} (1 - \pi_{g\upsilon i}) \, \sigma_{g\upsilon}^2 + o_p(n_\upsilon^{-1}) \,.$$

5. APPROXIMATIONS SELON LA MÉTHODE PLEINEMENT EFFICACE

Dans les sections précédentes, l'estimateur $\hat{\theta}_{FEFI}$ était construit de manière à ne produire aucune variance due à l'imputation. La mise en œuvre de la méthode d'imputation fractionnelle décrite en (20) pourrait nécessiter le recours à un grand nombre de donateurs pour chaque receveur. Nous énonçons donc une méthode prévoyant un nombre fixe de donateurs par receveur, qui est pleinement efficace pour le grand total, mais pas nécessairement pour les sous-populations. La méthode attribue des donateurs pour produire

une faible variance entre receveurs des valeurs imputées et modifie les poids des donateurs pour atteindre la pleine efficacité pour le total.

Supposons que w_i représente les poids initiaux des répondants. Définissons

$$S_{gw} = \sum_{i \in A_{Rg}} w_i , \tag{34}$$

soit la somme des poids des répondants compris dans la cellule g. Nous supposons qu'il faut attribuer M donateurs à chaque receveur. Les donateurs sont attribués de telle sorte que leur distribution est une approximation de celle des répondants. Une méthode de sélection possible est la stratification, où les donateurs sont choisis dans une strate présentant la probabilité

$$P_{ijgh} = S_{gwh}^{-1} w_i . \tag{35}$$

et

$$S_{gwh} = \sum_{i \in A_{Rgh}} w_i$$

est la somme des poids des éléments donateurs attribués à la strate h dans la cellule g. Étant donné l'ensemble de donateurs pour le receveur j, la valeur initiale de w_{ij}^* est

$$w_{ij0}^* = S_{gw}^{-1} S_{gwh} ,$$

où l'élément i appartient à la strate gh.

Une autre méthode de sélection possible consiste à procéder à un échantillonnage systématique avec une probabilité proportionnelle aux poids pour choisir les donateurs pour chaque receveur. Selon cette méthode, les valeurs initiales de w_{ij0}^* sont M^{-1}.

Une fois les donateurs attribués, les poids initiaux w_{ij}^* sont rajustés au moyen de méthodes de régression pour que la somme des poids donne les estimateurs pleinement efficaces de la moyenne de y et que la fonction de distribution cumulée estimée fondée sur les poids, soit une approximation de l'estimateur pleinement efficace de la fonction de distribution cumulée. Des études de simulation ont montré que cette méthode était efficace lorsque $M = 5$.

6. RÉSUMÉ

Nous avons décrit les propriétés de l'imputation fractionnelle, qui consiste à prélever plusieurs donateurs pour chaque valeur manquante et à attribuer à chaque donateur une fraction du poids du non-répondant. Nous avons montré que l'imputation fractionnelle au moyen d'un petit nombre de donateurs pour chaque non-répondant pouvait donner un estimateur pleinement efficace de la moyenne. L'imputation fractionnelle permet de construire un seul ensemble de répétitions d'estimation de la variance à utiliser pour estimer la variance des variables imputées, des variables observées à l'égard de tous les répondants et en fonction des hypothèses du modèle, pour des fonctions des deux types de variables. Par exemple, les répétitions donnent des estimations convergents des variances des moyennes de domaine. L'imputation fractionnelle donne des estimations de la variance ayant un biais et une variance plus faibles que les estimateurs à imputations multiples avec le même nombre d'imputations.

REMERCIEMENTS

Cette recherche a été soutenue en partie par Westat et par le département de l'Éducation de l'Université Iowa State en vertu du contrat de sous-traitance n° ED--99--CO—0109. Nous remercions Jean Opsomer et Damiao Da Silva pour leurs observations très utiles.

BIBLIOGRAPHIE

Fay, R. E. (1991), "A Design-Based Perspective on Missing Data Variance", *Proceedings of Bureau of the Census Annual Research Conference, American Statistical Association*, pp. 429–440.

Fay, R. E. (1992), "When are Inferences From Multiple Imputation Valid?" *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 227–232.

Ford, B. M. (1983), *An Overview of Hot Deck Procedures. Incomplete Data in Sample Surveys*, Vol. 2. New York: Academic Press.

Kalton, G. et Kasprzyk, D. (1986), "Le traitement des données d'enquête manquantes", *Techniques d'enquête*, 12, pp. 1–17.

Little, R. J. A. et Rubin, D. B. (1987), "*Statistical Analysis with Missing Data*", New York: Wiley.

Rao, J. N. K. (1996), "On Variance Estimation with Imputed Survey Data", *Journal of the American Statistical Association*, 91, pp. 499–506.

Rao, J. N. K. et Shao, J. (1992), "Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation", *Biometrika*, 79, pp. 811–822.

Rao, J. N. K. et Sitter, R. R. (1995), "Variance Estimation Under Two-Phase Sampling with Applications to Imputation for Missing Data", *Biometrika*, 82, pp. 453–460.

Rubin, D. B. (1976), "Inference and Missing Data", *Biometrika*, 63, pp. 581–590.

Rubin, D. B. (1978), "Multiple Imputations in Sample Surveys: A Phenomenological Bayesian Approach to Nonresponse", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 20–28.

Rubin, D. B. et Schenker (1986), "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse", *Journal of the American Statistical Association*, 81, pp. 366–374.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Sande, I. G. (1983), *Hot Deck Imputation Procedures, Incomplete Data in Sample Surveys*, Vol. 3. New York: Academic Press.

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, pp. 257–268.

Shao, J., Chen, Y. et Chen, Y. (1998), "Balanced Repeated Replication for Stratified Multistage Survey Data Under Imputation", *Journal of the American Statistical Association*, 93, pp. 819–831.

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, pp. 254–265.

Sitter, R. R. (1997), "Variance Estimation for the Regression Estimator in Two-Phase Sampling", *Journal of the American Statistical Association*, 92, pp. 780–787.

Tollefson, M. et Fuller, W. A. (1992), "Variance Estimation for Sampling with Random Imputation", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 140–145.

Yung, W. et Rao, J. N. K. (2000), “Jackknife Variance Estimation Under Imputation for Estimators Using Poststratification Information”, *Journal of the American Statistical Association*, 95, pp. 903–915.