Techniques d'enquête, décembre 1992
Vol. 18, n° 2, pp. 257–268
Statistique Canada

Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation

CARL-ERIK SÄRNDAL[1]

RÉSUMÉ

Dans presque toutes les grandes enquêtes, l'imputation est utilisée sous une forme ou sous une autre. Le présent article expose une méthode d'estimation de la variance lorsque l'imputation simple (plutôt que multiple) est utilisée pour créer un ensemble complet de données. L'imputation ne reproduit jamais les vraies valeurs (sauf dans des cas tout à fait exceptionnels). L'erreur totale de l'estimation de l'enquête est considérée, dans cet article, comme la somme de l'erreur d'échantillonnage et de l'erreur d'imputation. Par conséquent, on en arrive à une variance globale qui est la somme d'une variance d'échantillonnage et d'une variance d'imputation. L'article s'attarde principalement à l'estimation de ces deux composantes, en utilisant les données après imputation, c'est-à-dire les valeurs réellement observées et les valeurs imputées. La méthode est fondée sur un modèle, en ce sens que le modèle sous-jacent à la méthode d'imputation, ainsi que la distribution de randomisation ayant servi à la sélection de l'échantillon, détermineront ensemble la forme que prendront les estimateurs de la variance. Les résultats théoriques sont confirmés par une simulation de Monte Carlo.

MOTS CLÉS: Imputation simple; estimation de la variance; modèle d'imputation; inférence fondée sur un modèle.

1. DIFFÉRENTS TYPES D'IMPUTATION

Le présent article fait état de travaux réalisés dans le cadre de l'élaboration du Système généralisé d'estimation (SGE) de Statistique Canada. Des estimations de la variance seront régulièrement effectuées à l'aide des différents modules d'estimation définissant le SGE. Il est apparu nécessaire de mettre au point des méthodes qui puissent servir à l'estimation de la variance dans les cas où l'ensemble de données contient des valeurs imputées, comme c'est le cas dans la plupart des enquêtes.

La pondération et l'imputation sont deux des principales méthodes permettant de traiter les données manquantes à des fins d'estimation. Dans les publications récentes traitant de la pondération, on considère habituellement que les poids servant à compenser la non-réponse devraient être l'inverse des probabilités de réponse propres à un mécanisme de réponse dont on fait l'hypothèse. Puisque les probabilités de réponse, en général, sont inconnues, elles doivent être estimées à partir des données disponibles. L'imputation, par contre, a l'avantage de fournir un tableau de données complet. Un tel tableau simplifie le traitement des données, mais cela ne signifie pas pour autant que les "méthodes d'estimation ordinaires" peuvent être utilisées directement. Les valeurs imputées se fondent sur des échantillons, et ont par conséquent leurs paramètres statistiques propres, comme une moyenne et une variance.

De nos jours, l'imputation est une technique abondamment utilisée. Il est intéressant de signaler ce que Pritzker, Ogus et Hansen (1965) disent de la ligne de conduite adoptée en matière d'imputation par le US Bureau of the Census: "Essentiellement, notre approche en ce qui concerne le problème de ... l'imputation est que nous devrions recueillir de l'information par mesure directe pour une proportion très élevée des agrégats à totaliser, avec un contrôle suffisant

[1] Carl-Erik Särndal, Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, succursale A, Montréal (Québec) H3C 3J7.

de la qualité, de telle sorte que pratiquement toute règle raisonnable ... d'imputation donnera grosso modo les mêmes résultats ... Pour ce qui est de l'imputation dans les recensements et les enquêtes par sondage, nous avons adopté une norme selon laquelle un faible niveau d'imputation, de l'ordre de 1 ou 2 %, est ce que nous visons."

Idéalement, un niveau d'imputation de 1 ou 2 % demeure un objectif que nous devrions chercher à atteindre. On note toutefois un taux d'imputation beaucoup plus élevé dans la plupart des enquêtes réalisées de nos jours par les grands organismes d'enquête. Il est clair que si 30 % des valeurs sont imputées, les effets de l'imputation ne peuvent être passés sous silence. L'imputation peut engendrer une erreur systématique (biais) de l'estimation ponctuelle; c'est peut-être là son plus grand danger. Toutefois, même si l'on peut trouver une méthode d'imputation ne produisant pas d'erreur systématique appréciable, on ne peut omettre l'effet souvent considérable qu'exerce l'imputation sur la précision (la variance) de l'estimation ponctuelle. Il importe de se doter de méthodes simples, mais valables, d'estimation de la variance pour les données d'enquête contenant des imputations, afin de pouvoir évaluer adéquatement les coefficients de variation des estimations de l'enquête.

Plusieurs méthodes d'imputation différentes ont été proposées. Celles-ci peuvent être classées de diverses façons. L'un des classements possibles fait intervenir le nombre d'imputations effectuées. Dans les méthodes d'**imputation simple**, une seule valeur est imputée en remplacement d'une donnée manquante. On obtient ainsi un tableau de données complet, dans lequel les valeurs imputées sont marquées. Les estimations sont calculées au moyen de l'ensemble de données complet. Dans le cas de l'**imputation multiple**, deux valeurs ou plus sont imputées à chaque valeur manquante. Plusieurs ensembles de données complets sont ainsi obtenus. Les estimations sont calculées au moyen des ensembles de données complets.

Les méthodes d'imputation diffèrent également selon le modèle qui sous-tend l'imputation. Certaines méthodes se fondent sur un modèle **explicite**, par exemple lorsque la valeur imputée est obtenue par un ajustement de régression, un quotient ou une moyenne. Dans d'autres méthodes, le modèle est seulement **implicite**, par exemple dans l'imputation hot deck et l'imputation du plus proche voisin. Les distinctions qui précèdent sont importantes dans le cadre du présent article.

Statistique Canada utilise actuellement diverses méthodes d'imputation: plus proche voisin, quotient courant, moyenne courante, valeur précédente, moyenne précédente, tendance auxiliaire. Ce sont toutes des méthodes d'imputation simple. Les valeurs imputées sont produites dans le Système généralisé de vérification et d'imputation (SGVI); de là elles entrent dans le Système généralisé d'estimation (SGE), où les estimations ponctuelles et les estimations de la variance sont calculées dans plusieurs modules d'estimation différents. Le présent article traite en particulier de la méthode d'imputation par le quotient courant, l'une de celles fondées sur un modèle explicite.

2. QUELQUES RÉFLEXIONS SUR L'INPUTATION MULTIPLE

La méthode de l'imputation multiple a été proposée par D.B. Rubin vers 1977. Ses idées sont traitées dans plusieurs articles, en particulier dans Herzog et Rubin (1983) et Rubin (1986) qui en exposent la teneur, ainsi que dans un livre, Rubin (1987). L'imputation multiple présente à la fois des avantages et des inconvénients; il en va de même de l'imputation simple.

Rubin (1986) voit dans l'imputation simple l'inconvénient suivant: "... la valeur imputée exclut toute incertitude sur la valeur à imputer. Si une valeur était vraiment appropriée, elle ne serait pas manquante. Ainsi, lorsqu'on assimile les valeurs imputées à des valeurs observées, on sous-estime systématiquement l'élément d'incertitude même en supposant que l'on connaisse les motifs exacts de la non-réponse."

L'imputation multiple est attrayante du fait qu'elle incorpore l'idée d'une variabilité dans l'imputation. C'est précisément cette variabilité – la variabilité à l'intérieur des ensembles complets de données et entre ces ensembles – qui est exploitée dans les méthodes d'estimation de la variance liées à l'imputation multiple. Ces méthodes font un usage très fructueux des notions statistiques de base. (On pourrait faire valoir, toutefois, que la sélection d'un échantillon comporte elle aussi une variabilité, mais que la plupart des enquêtes doivent se limiter au prélèvement d'un seul échantillon et que l'estimation doit se faire uniquement à partir de ce dernier.)

On peut constater, à l'aide d'exemples simples, que le fait de traiter les valeurs imputées comme des valeurs observées peut entraîner une grave sous-estimation de l'incertitude réelle; les responsables de l'échantillonnage sont depuis longtemps sensibles à ce danger. Et il est vrai que parfois, les utilisateurs traitent les valeurs imputées exactement comme les valeurs observées, avec comme résultat des énoncés de précision erronés. Avec les moyens informatiques d'aujourd'hui, il est facile de faire une imputation selon une règle ou une autre, mais il est plus difficile d'obtenir des estimations valides de la variance.

La citation ci-dessus laisse entendre que l'imputation d'une valeur unique, parce qu'aucune variation ne lui est associée, ne permet pas d'en arriver à des estimations raisonnables de la variance, et conduit nécessairement à une sous-estimation. Je ne partage pas cet avis. Les méthodes que j'examine montrent que des estimations valides de la variance sont bel et bien possibles avec l'imputation simple.

Une méthode d'estimation de la variance en présence de valeurs imputées devrait avoir les propriétés suivantes: a) reposer sur un fondement théorique solide; b) être robuste à l'égard des hypothèses régissant l'imputation; c) être pratique, simple à réaliser et facilement acceptée par les utilisateurs.

Bien que l'imputation multiple ait les qualités énoncées en a) et b), il est clair qu'elle est dépourvue de la propriété c), à tout le moins dans certaines applications. Dans l'élaboration du SGE, nous devons miser sur des méthodes d'application facile, aptes à recueillir facilement l'adhésion de l'utilisateur. L'utilisateur de l'ensemble de données (quelqu'un qui n'est pas avant tout un statisticien) peut comprendre facilement que le statisticien fait une imputation unique, avec l'objectif d'attribuer la meilleure valeur possible à l'élément manquant. Certes, il peut être intéressant à certaines fins, par exemple pour des analyses secondaires, de disposer de plusieurs ensembles complets de données, mais les coûts de stockage de tels ensembles obligeront souvent à renoncer à une telle option.

Il est fort possible que l'imputation multiple se révèle utile dans des contextes différents, et pour des raisons autres que celles qui sont essentielles à l'élaboration du SGE. L'imputation multiple est l'un des moyens qui existent pour traiter le problème de la sous-estimation de la variance, du moins dans certaines situations. Cette méthode, par ailleurs récemment critiquée par Fay (1991), ne constitue pas la seule solution possible. Voyons les possibilités offertes par les méthodes d'imputation simple. Le méthode décrite ci-dessous se fonde sur Särndal (1990).

3. VARIANCE D'IMPUTATION ET VARIANCE D'ÉCHANTILLONNAGE

Une règle d'imputation correspond à un modèle (explicite ou implicite) de la relation entre les variables d'intérêt de l'enquête. Autrement dit, dès que l'analyste a établi une règle d'imputation, il ou elle a, en fait, choisi un modèle. Le principe du développement qui suit est que si une telle règle est considérée suffisamment bonne pour les estimations ponctuelles (absence d'erreur systématique), elle est également suffisamment bonne pour les estimations connexes de la variance. En d'autres termes, le modèle devrait permettre à la fois le contrôle du biais et la validité de l'estimation de la variance.

Soit $U = \{1, \ldots, k, \ldots, N\}$ une population finie; supposons que y désigne une des variables étudiées dans l'enquête. L'objectif est d'estimer le total de y pour l'ensemble de la population, c.-à-d. $t = \sum_U y_k$. (Pour n'importe quel ensemble C d'unités de la population, où $C \subseteq U$, $\sum_C$, $\sum_{k \in C}$ est une désignation abrégée de $t = \sum_U y_k$; par exemple, signifie $\sum_{k \in U} y_k$.) Un échantillon probabiliste s est prélevé selon un plan d'échantillonnage déterminé. Les probabilités d'inclusion sont connues, et l'on obtiendrait des estimations ordinaires, fondées sur le plan, de la variance si toutes les unités $k \in s$ étaient observées. Toutefois, il y a des données manquantes. Soit r le sous-ensemble de s pour lequel les valeurs y_k sont réellement observées. Pour le complément, $s - r$, des valeurs imputées sont calculées. Les **données après imputation** sont l'ensemble des valeurs dénotées $y_{\blacklozenge k}$, $k \in s$, telles que:

$$y_{\blacklozenge k} = \begin{cases} y_k & \text{si } k \in r \\ y_{\text{imp},k} & \text{si } k \in s - r \end{cases},$$

où y_k est une valeur réellement observée, et $y_{\text{imp},k}$ est la valeur imputée pour l'unité k. Si $r = s$, il n'y a aucune imputation; toutes les données sont des observations réelles.

Écrivons l'estimateur de t qui serait utilisé dans le cas d'une réponse de 100 % (c.-à-d. $r = s$) de la façon suivante: $\hat{t} = \sum_{k \in s} w_k y_k = \sum_s w_k y_k$, où w_k est le poids attribué à l'observation y_k. Par exemple, pour un échantillonnage aléatoire simple sans remise (EASSR) de n unités parmi N, $w_k = N/n$ pour tous les $k \in s$ lorsque la moyenne de l'échantillon étendue à la population est utilisée pour estimer t, et $w_k = (\bar{z}_U/\bar{z}_s)(N/n) = (\sum_U z_k)/(\sum_s z_k)$ pour tous les $k \in s$ lorsqu'on utilise l'estimateur par quotient avec z comme variable auxiliaire.

Lorsque les données contiennent des valeurs imputées, l'estimateur de t est $\hat{t}_{\blacklozenge} = \sum_s w_k y_{\blacklozenge k}$. Autrement dit, nous supposons que les poids w_k sont identiques à ceux qui sont utilisés lorsque toutes les données sont des observations réelles. Ce principe est utilisé dans les modules d'estimation du SGE. Il s'appuie sur l'hypothèse que l'imputation selon la règle choisie cause peu ou pas d'erreur systématique dans les estimations.

La variance de l'estimation d'un total est accrue par l'imputation, parce que l'imputation (sauf dans des circonstances tout à fait exceptionnelles) ne reproduit pas la vraie valeur y_k. On le constate concrètement quand on applique la règle d'imputation à des unités de l'échantillon réellement observées; des erreurs se produisent toujours. Si la règle n'est pas exempte d'erreur pour les unités répondantes, elle n'est pas non plus exempte d'erreur pour les unités non répondantes. À la section 4, nous exprimons la variance de $\hat{t}_{\blacklozenge}$ comme la somme de deux composantes, une variance d'échantillonnage et une variance attribuable à l'imputation:

$$V_{\text{tot}} = V_{\text{sam}} + V_{\text{imp}}.$$

La variance d'imputation V_{imp} est nulle si toutes les données sont réellement des valeurs observées, ou si la méthode d'imputation est apte à reproduire exactement la vraie valeur y_k pour chaque unité exigeant une imputation. (Aucun de ces deux cas n'est susceptible de survenir en pratique.) Le processus décrit à la section 4 utilise les données après imputation, $y_{\blacklozenge k}$, $k \in s$, pour obtenir des estimations de chacune des deux composantes, ce qui donne:

$$\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}}.$$

La composante $\hat{V}_{\text{sam}}$ est déterminée en deux étapes:

1) Calculer l'estimation ordinaire, fondée sur le plan, de la variance au moyen des données après imputation. (Par exemple, si l'EASSR est utilité et que $r = s$, l'estimation non biaisée ordinaire de la variance de $N\bar{y}_s$ est $N^2(1/n - 1/N)\sum_s (y_k - \bar{y}_s)^2/(n-1)$. Cette formule, appliquée aux données après imputation, donne $N^2(1/n - 1/N)$ $\sum_s (y_{\blacklozenge k} - \bar{y}_{\blacklozenge s})^2/(n-1)$, où $\bar{y}_{\blacklozenge s}$ est la moyenne des n valeurs $y_{\blacklozenge s}$.)

2) Ajouter un terme de correction tenant compte du fait que de nombreuses règles d'imputation produisent des données ayant une variabilité "moindre que naturelle", phénomène propre à entraîner une sous-estimation de la variance d'échantillonnage à moins que des mesures correctives soient prises. Enfin, la composante $\hat{V}_{\text{imp}}$ est facilement calculée à partir des données après imputation. L'utilisateur acceptera facilement l'argument selon lequel la variance fournie par la formule ordinaire n'est pas suffisante à elle seule; quelque chose doit être ajouté pour tenir compte de l'imperfection de la règle d'imputation.

Une propriété intéressante de la méthode est que si aucune imputation n'est nécessaire, c.-à-d. si $r = s$, il en résulte que $\hat{V}_{\text{imp}} = 0$ et que $\hat{V}_{\text{sam}}$ égale l'"estimateur ordinaire de la variance" qui aurait été utilisé si 100 % des valeurs avaient été réellement observées.

4. EXPOSÉ THÉORIQUE

L'erreur totale de $\hat{t}_\blacklozenge$ est ainsi décomposée:

$$\hat{t}_\blacklozenge - t = (\hat{t} - t) + (\hat{t}_\blacklozenge - \hat{t}) = \text{erreur d'échantillonnage} + \text{erreur d'imputation.}$$

L'erreur d'imputation est la différence entre l'estimation inconnue qui aurait été calculée si les données avaient été formées entièrement d'observations réelles, et l'estimation qui peut être calculée à partir des données après imputation. L'erreur d'imputation est

$$\hat{t}_\blacklozenge - \hat{t} = -\sum_{s-r} w_k e_k,$$

où

$$e_k = y_k - y_{\text{imp},k}$$

est un **résidu d'imputation** qui ne peut être observé pour une unité $k \in s-r$. La grandeur de e_k dépend de la qualité d'ajustement offerte par le modèle d'imputation. Les résidus sont faibles si la méthode d'imputation donne des valeurs substituts presque parfaites. L'analyse, à partir d'ici, peut prendre diverses directions. Nous utiliserons une méthode **fondée sur un modèle** dans laquelle trois différentes distributions de probabilité sont examinées. Les symboles d'espérance correspondants sont E_ξ, E_s et E_r. L'indice ξ signifie "par rapport au modèle d'imputation", l'indice s signifie "par rapport au plan d'échantillonnage" et l'indice r signifie "par rapport au mécanisme de réponse, s étant donné". Le modèle découle de la règle d'imputation, et est donc connu; le plan d'échantillonnage est la distribution d'échantillonnage probabiliste donnée, et est donc aussi connu; le mécanisme de réponse est une distribution habituellement inconnue régissant la réponse, l'échantillon s étant donné.

L'estimateur $\hat{t}_\blacklozenge$ est globalement non biaisé, au sens où $E_\xi E_s E_r(\hat{t}_\blacklozenge - t) = 0$, si deux conditions sont remplies:

a) l'ordre des opérateurs d'espérance peut être modifié, de sorte que $E_\xi E_s E_r(\cdot)$ peut être évalué sous la forme $E_s E_r\{E_\xi(\cdot \mid s, r)\}$, et

b) le résidu d'imputation $e_k = y_k - y_{\text{imp},k}$ a une espérance (selon le modèle) nulle pour chaque $k \in r$, c.-à-d. $E_\xi(e_k) = 0$, d'où il résulte que $E_\xi(\hat{t}_\blacklozenge - \hat{t}) = 0$.

La condition (a) est remplie si la réponse est régie par un mécanisme qui peut dépendre de s et de données auxiliaires, mais non des y_k, $k \in s$. Autrement dit, la probabilité $q(r)$ de réaliser l'ensemble de réponses r est de la forme $q(r) = q(r \mid s, \{x_k : k \in s\})$, où $\{x_k : k \in s\}$ désigne les données auxiliaires. Le mécanisme de réponse peut alors être qualifié de neutre (c.-à-d. qu'on peut ne pas en tenir compte).

Examinons maintenant la variance globale donnée par

$$V_{\text{tot}} = E_\xi E_s E_r\{(\hat{t}_\blacklozenge - t)^2\},$$

qui peut être aussi appelée variance prévisible en vertu du modèle d'imputation ξ. Nous obtenons:

$$\begin{aligned} V_{\text{tot}} &= E_{\xi sr}(\hat{t}_\blacklozenge) = E_\xi E_s E_r\{(\hat{t}_\blacklozenge - t)^2\} \\ &= E_\xi E_s E_r\{(\hat{t} - t) + (\hat{t}_\blacklozenge - \hat{t})\}^2 \\ &= E_\xi V_p + E_s E_r V_{\xi c}, \end{aligned} \tag{4.1}$$

où $V_p = E_s\{(\hat{t} - t)\}^2$ est la variance fondée sur le plan de $\hat{t}$, en supposant que $\hat{t}$ est non biaisé selon le plan pour le total t. (Pour un estimateur légèrement biaisé selon le plan, V_p est l'erreur quadratique moyenne selon le plan de $\hat{t}$.) Notons que $(\hat{t} - t)$ dépend de s seulement, et non de r. De plus,

$$V_{\xi c} = E_\xi\{(\hat{t}_\blacklozenge - \hat{t})^2 \mid s, r\}$$

est la variance (selon le modèle) de l'erreur d'imputation, étant donné des ensembles s et r. L'indice c signifie "conditionnelle". Pour en arriver à (4.1), on suppose que la condition (a) est vérifiée, c.-à-d. que l'espérance E_ξ peut être déplacée à l'intérieur de $E_s E_r$, et que le terme mixte

$$2E_\xi E_s[(\hat{t} - t)\{E_r(\hat{t}_\blacklozenge - \hat{t}) \mid s\}] \tag{4.2}$$

disparaît, ou est suffisamment proche de zéro pour être omis. Ce serait le cas si l'erreur d'imputation probable était nulle ou négligeable en vertu du mécanisme de réponse, étant donné l'échantillon probabiliste s réalisé. Même si (4.2) n'égale pas exactement zéro pour le mécanisme déterminant la réponse, on peut souvent assimiler (4.2) à zéro et utiliser quand même la méthode ci-dessous, laquelle permet d'obtenir une bien meilleure estimation de la variance que si l'on prétendait naïvement que les données imputées sont aussi valables que les données réellement observées. Pour une imputation par le quotient et un EASSR, cas que nous examinons à la section 5, le terme (4.2) égale exactement zéro.

En utilisant les désignations $V_{\text{sam}} = E_\xi V_p$ et $V_{\text{imp}} = E_s E_r V_{\xi c}$ dans (4.1), on obtient

$$V_{\text{tot}} = V_{\text{sam}} + V_{\text{imp}}$$

ou

$$\text{variance totale} = \text{variance d'échantillonnage} + \text{variance d'imputation}.$$

L'objectif est d'estimer la variance totale, de façon à déterminer un intervalle de confiance valide pour la valeur inconnue t. Notre méthode consiste à obtenir des estimations distinctes, $\hat{V}_{\text{sam}}$ et $\hat{V}_{\text{imp}}$, des deux composantes $V_{\text{sam}} = E_\xi V_p$ et $V_{\text{imp}} = E_s E_r V_{\xi c}$. Les données disponibles pour cette estimation sont les $y_{\blacklozenge k}$, $k \in s$. La façon d'obtenir $\hat{V}_{\text{sam}}$ et $\hat{V}_{\text{imp}}$ est la suivante:

(i) Estimation de la composante variance d'échantillonnage. Soit $\hat{V}_p$ l'estimateur ordinaire (non biaisé selon le plan ou quasi non biaisé selon le plan) de la variance selon le plan V_s. Appelons $\hat{V}_{\blacklozenge p}$ la quantité résultant du calcul de $\hat{V}_p$ à partir des valeurs après imputation $y_{\blacklozenge k}$, $k \in s$. Pour de nombreuses règles d'imputation, $\hat{V}_{\blacklozenge p}$ sous-estime V_{sam}. La sous-estimation est compensée de la façon suivante. Évaluons l'espérance conditionnelle

$$E_\xi(\hat{V}_p - \hat{V}_{\blacklozenge p} \mid s, r) = V_{\text{dif}}.$$

Alors, pour s et r donnés, trouvons un estimateur non biaisé selon le modèle, désigné $\hat{V}_{\text{dif}}$, de V_{dif}. Cela exigera normalement l'estimation de certains paramètres du modèle ξ. Par conséquent,

$$E_\xi(\hat{V}_{\text{dif}} \mid s, r) = E_\xi(\hat{V}_p - \hat{V}_{\blacklozenge p} \mid s, r).$$

Alors

$$\hat{V}_{\text{sam}} = \hat{V}_{\blacklozenge p} + \hat{V}_{\text{dif}}$$

est globalement non biaisé pour la composante $V_{\text{sam}} = E_\xi V_p$, comme le montre le calcul qui suit:

$$\begin{aligned} E_\xi E_s E_r(\hat{V}_{\text{sam}}) &= E_s E_r\{E_\xi(\hat{V}_{\blacklozenge p}) + E_\xi(\hat{V}_{\text{dif}})\} \\ &= E_s E_r\{E_\xi(\hat{V}_p)\} = E_\xi E_s(\hat{V}_p) \\ &= E_\xi V_p = V_{\text{sam}}. \end{aligned}$$

(ii) Estimation de la composante variance d'imputation. Il suffit de trouver un estimateur, $\hat{V}_{\xi c}$, qui soit non biaisé selon le modèle pour $V_{\xi c}$. Autrement dit, $E_\xi(\hat{V}_{\xi c}) = V_{\xi c}$. Encore ici, l'estimation de paramètres inconnus du modèle ξ peut être nécessaire. Il en résulte que $\hat{V}_{\xi c}$ est globalement non biaisé pour la composante variance d'imputation V_{imp}, puisque

$$E_s E_r E_\xi(\hat{V}_{\xi c}) = E_s E_r V_{\xi c} = V_{\text{imp}}.$$

Enfin, un estimateur globalement non biaisé de V_{tot} est donné par

$$\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}},$$

où $\hat{V}_{\text{sam}} = \hat{V}_{\blacklozenge p} + \hat{V}_{\text{dif}}$ et $\hat{V}_{\text{imp}} = \hat{V}_{\xi c}$. Notons que le rôle de $\hat{V}_{\text{dif}}$ est d'apporter une correction tenant compte du fait que les données après imputation peuvent afficher une variation "moindre que naturelle". Cela se produit souvent lorsque $y_{\text{imp},k}$ égale la valeur prévue provenant d'un ajustement de régression, c.-à-d. "la valeur située sur la ligne". La variation autour de la ligne n'est pas reflétée par la valeur prévue.

Pour que l'estimateur $\hat{V}_{\text{tot}}$ construit ci-dessus soit globalement non biaisé, la condition (a) doit être remplie, le terme énoncé en (4.2) doit être égal à zéro, et le modèle d'imputation doit être approprié, de façon que $\hat{V}_{\text{dif}}$ et $\hat{V}_{\xi c}$ soient non biaisés selon le modèle pour V_{dif} et $V_{\xi c}$, respectivement. Il se peut que de légers écarts par rapport au modèle d'imputation supposé n'aient pas de conséquences graves, mais si la définition du modèle d'imputation est grossièrement déficiente, il est clair que $\hat{V}_{\text{tot}}$ pourra être gravement biaisé, en raison du biais selon le modèle de $\hat{V}_{\text{dif}}$ et $\hat{V}_{\xi c}$. Des simulations de Monte Carlo décrites dans Lee, Rancourt et Särndal (1992) montrent que l'estimateur de la variance $\hat{V}_{\text{tot}}$ est passablement robuste à l'égard des déficiences du modèle d'imputation. Quoi qu'il en soit, le fait d'ajouter les termes $\hat{V}_{\text{dif}}$ et $\hat{V}_{\xi c}$ constitue une importante amélioration par rapport à l'utilisation de l'estimateur de variance non corrigé élémentaire $\hat{V}_{\blacklozenge p}$.

Notons que si le modèle d'imputation demeure valable, une estimation non biaisée de la variance est fournie par la méthode même si les probabilités de réponse diffèrent entre les unités, pourvu qu'elles dépendent des valeurs x_k seulement. Autrement dit, nous pouvons appliquer

un profil de réponse systématique en vertu duquel les unités ayant des valeurs x_k élevées sont moins susceptibles de répondre que les unités ayant des valeurs x_k faibles. Si les probabilités de réponse dépendent explicitement des valeurs y_k, la situation est différente; le mécanisme de réponse n'est pas neutre et la condition (a) n'est pas remplie. Il y aura alors un biais dans $\hat{V}_{\text{tot}}$ attribuable à cette non-neutralité; les simulations décrites dans Lee, Rancourt et Särndal (1992) donnent une idée de l'ampleur de ce biais.

Exemple. On utilise l'EASSR pour prélever l'échantillon s, formé de n unités parmi N. Désignons par m la taille de l'ensemble de réponses r. Supposons qu'on impute la moyenne des répondants aux unités exigeant une imputation. Le modèle d'imputation correspondant ξ prévoit que $y_k = \beta + \epsilon_k$, où les ϵ_k sont des termes d'erreur non corrélés avec $E_\xi(\epsilon_k) = 0$, $V_\xi(\epsilon_k) = \sigma^2$. Autrement dit, $y_{\blacklozenge k} = y_k$ si $k \in r$ et $y_{\blacklozenge k} = \hat{\beta} = \bar{y}_r$ si $k \in s - r$, et nous obtenons l'estimateur $\hat{t}_\blacklozenge = (N/n)\sum_s y_{\blacklozenge k} = N\bar{y}_r$. Ici, l'estimateur ordinaire, fondé sur le plan, de la variance pour une réponse de 100 % est $\hat{V}_p = N^2(1/n - 1/N)\sum_s (y_k - \bar{y}_s)^2/(n-1)$; lorsque cette formule est appliquée aux données après imputation, on obtient $\hat{V}_{\blacklozenge p} = N^2(1/n - 1/N)\{(m-1)/(n-1)\}S^2_{yr}$, où $S^2_{yr} = \sum_r (y_k - \bar{y}_r)^2/(m-1)$. D'autres calculs donnent $\hat{V}_{\text{dif}} = N^2(1/n - 1/N)\{(n-m)/(n-1)\}S^2_{yr}$ et $\hat{V}_{\text{imp}} = N^2(1/m - 1/n)S^2_{yr}$. Ainsi, $\hat{V}_{\text{sam}} = \hat{V}_{\blacklozenge p} + \hat{V}_{\text{dif}} = N^2(1/n - 1/N)S^2_{yr}$ et $\hat{V}_{\text{tot}} = N^2(1/m - 1/N)S^2_{yr}$, lequel est facile à accepter comme "bon" estimateur de la variance pour cette règle d'imputation simple. Le tableau qui suit montre la contribution de chacun des trois termes à l'estimateur de la variance totale $\hat{V}_{\text{tot}}$, pour différents taux d'imputation, en supposant que N est grand comparativement à m et à n, et que $(m-1)/m \approx (n-1)/n \approx 1$.

taux d'imputation en %	contribution en % à $\hat{V}_{\text{tot}}$		
$100\ (1 - m/n)$	$\hat{V}_{\blacklozenge p}$	$\hat{V}_{\text{dif}}$	$\hat{V}_{\text{imp}}$
10	81	9	10
20	64	16	20
30	49	21	30

Le tableau illustre les dangers de considérer les valeurs imputées comme des données réelles. Avec 30 % de valeurs imputées, l'estimateur de la variance ordinaire $\hat{V}_{\blacklozenge p}$ couvre, dans cet exemple, moins de la moitié de la variance totale correctement estimée. L'imputation par la moyenne des répondants est un exemple utile, du fait de la relative simplicité des résultats. Mais habituellement, en situation réelle, une telle imputation n'est ni justifiée ni efficace. Le modèle sous-jacent n'est pas assez raffiné pour éviter une erreur systématique dans les estimations ponctuelles, et les résidus $e_k = y_k - \bar{y}_r$ peuvent varier considérablement.

5. APPLICATION À L'IMPUTATION PAR LE QUOTIENT COURANT

La méthode se fonde sur l'hypothèse qu'une valeur auxiliaire positive x_k est connue pour chaque unité $k \in s$. Si $k \in s - r$, nous imputons $y_{\text{imp},k} = \hat{B}x_k$ où $\hat{B} = (\sum_r y_k)/(\sum_r x_k)$. Les données après imputation sont les suivantes:

$$y_{\blacklozenge k} = \begin{cases} y_k & \text{si } k \in r \\ \hat{B}x_k & \text{si } k \in s-r. \end{cases}$$

L'imputation par le quotient courant se fonde sur le modèle

$$y_k = \beta x_k + \epsilon_k, \tag{5.1}$$

où les ϵ_k sont des erreurs du modèle non corrélées telles que

$$E_\xi(\epsilon_k) = 0, \quad V_\xi(\epsilon_k) = \sigma^2 x_k. \tag{5.2}$$

Supposons que l'échantillon s soit choisi par EASSR. Désignons les tailles respectives de s, r et $s - r$ par n, m et $n - m$. Si aucune imputation n'était nécessaire, l'estimateur de $t = \sum_U y_k$ serait $\hat{t} = N\bar{y}_s$. En utilisant les données après imputation, nous obtenons

$$\hat{t}_\blacklozenge = (N/n) \sum_s y_{\blacklozenge k} = N\bar{x}_s\bar{y}_r/\bar{x}_r. \tag{5.3}$$

(Le trait supérieur et l'indice s, r ou $s - r$ désignent la "moyenne ordinaire", par exemple $\bar{y}_r = \sum_r y_k/m$, $\bar{x}_{s-r} = \sum_{s-r} x_k/(n - m)$, *etc.*) D'après les résultats de la section précédente, nous avons $V_{\text{tot}} = V_{\text{sam}} + V_{\text{imp}}$, avec $V_{\text{sam}} = E_\xi\{N^2(1/n - 1/N)S^2_{yU}\}$ et $V_{\text{imp}} = E_s E_r\{N^2(1/m - 1/n)C_1\sigma^2\}$, où $S^2_{yU} = \sum_U (y_k - \bar{y}_U)^2/(N - 1)$ et $C_1 = \bar{x}_s\bar{x}_{s-r}/\bar{x}_r$, une constante connue. Le terme mixte (4.2) égale exactement zéro dans ce cas. Notre méthode d'estimation de la variance donne $\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}}$ avec

$$\hat{V}_{\text{sam}} = N^2(1/n - 1/N)\{S^2_{y\blacklozenge s} + C_0\hat{\sigma}^2\}, \tag{5.4}$$

$$\hat{V}_{\text{imp}} = N^2(1/m - 1/n)C_1\hat{\sigma}^2, \tag{5.5}$$

où $S^2_{y\blacklozenge s} = \sum_s (y_{\blacklozenge k} - \bar{y}_{\blacklozenge s})^2/(n - 1)$ est la variance calculée à partir des données après imputation, et nous avons choisi d'estimer σ^2 par la formule non biaisée selon le modèle

$$\sigma^2 = \frac{1}{\bar{x}_r\{1 - (1/m)(cv_{xr})^2\}} \frac{\sum_r (y_k - \hat{B}x_k)^2}{m - 1},$$

où $cv_{xr} = S_{xr}/\bar{x}_r$ est le coefficient de variation de x dans l'ensemble de réponses r. La constante C_0 est donnée par

$$C_0 = \frac{1}{\sigma^2} E_\xi(S^2_{ys} - S^2_{y\blacklozenge s}),$$

où

$$S^2_{ys} = \frac{1}{n - 1} \sum_s (y_k - \bar{y}_s)^2$$

est la variance de l'échantillon (inconnue) selon des données formées de 100 % d'observations réelles. Après évaluation, on a

$$C_0 = \frac{1}{n - 1}\left\{ \sum_{s-r} x_k - \frac{\sum_{s-r} x_k^2}{\sum_r x_k} + \frac{1}{n} \frac{\sum_{s-r} x_k \sum_s x_k}{\sum_r x_k} \right\}.$$

Si m n'est pas trop petit, les approximations $\hat{\sigma}^2 \approx (\sum_r e_k^2)/(\sum_r x_k)$ où $e_k = y_k - \hat{B}x_k$ et $C_0 \approx (1 - m/n)\bar{x}_{s-r}$ sont suffisamment bonnes pour la plupart des applications.

La composante variance d'imputation peut être ainsi exprimée

$$\hat{V}_{\text{imp}} = N^2(1/m - 1/n)A\bar{x}_s\hat{\sigma}^2,$$

où $A = \bar{x}_{s-r}/\bar{x}_r$. La constante A reflète l'effet de sélection dû à la non-réponse. Si les unités ayant une valeur élevée sont moins enclines à répondre que les unités ayant une valeur faible, A peut être beaucoup plus grand que un et, pour un échantillon donné s et un nombre de répondants donné m, la composante $\hat{V}_{\text{imp}}$ tend à être élevée, par rapport au cas où, disons, toutes les unités sont également enclines à répondre. Cette tendance apparaît raisonnable intuitivement.

Deux cas spéciaux sont à signaler: 1) Si tous les x_k sont égaux à 1, l'estimation de la variance totale devient simplement

$$\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}} = N^2(1/m - 1/N)S_{yr}^2,$$

où S_{yr}^2 est la variance des m observations réelles y_k. Cela correspond à la variance obtenue selon un plan d'échantillonnage à deux phases, chaque phase étant fondée sur un EASSR. 2) Si aucune imputation n'est nécessaire, c.-à-d. si $s = r$, on a $\hat{V}_{\text{imp}} = 0$, et

$$\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} = N^2(1/n - 1/N)S_{ys}^2.$$

Autrement dit, notre méthode donne l'estimateur bien connu de la variance pour un EASSR.

Une étude de Monte Carlo comportant 100,000 ensembles de réponses répétés r a été effectuée afin de confirmer les résultats ci-dessus pour une imputation par le quotient courant. Une population finie de taille $N = 100$ a été créée conformément au modèle décrit par (5.1) et (5.2). L'ensemble de réponses typique r a été obtenu de la façon suivante: tirage par EASSR d'un échantillon s de taille $n = 30$; s étant donné, création de r par un mécanisme de réponse prenant la forme d'essais de Bernoulli indépendants, un pour chaque $k \in s$, avec une probabilité θ_k que le résultat soit "réponse". Trois mécanismes de réponse différents ont été utilisés; mécanisme 1: θ_k s'accroît avec y_k de telle sorte que $\theta_k = 1 - \exp(-a_1 y_k)$; mécanisme 2: θ_k s'accroît à mesure que y_k décroît, de telle sorte que $\theta_k = \exp(-a_2 y_k)$; mécanisme 3: θ_k est constant à 0.7, c.-à-d. que le mécanisme de réponse est uniforme. Les constantes a_1 et a_2 dans les deux premiers mécanismes de réponse (qui peuvent être qualifiés de non neutres) ont été fixées de manière à produire une probabilité de réponse moyenne de 0.7. Les tailles des ensembles de réponses r réalisés ont donc varié autour d'une moyenne de 21 pour les trois mécanismes. Pour chaque r, l'estimation ponctuelle $\hat{t}_\blacklozenge$ donnée par (5.3) a été calculée, ainsi que trois estimateurs de la variance différents, $\hat{V} = \hat{V}(\hat{t}_\blacklozenge)$. Ces estimateurs étaient: 1) l'estimateur de la variance **fondé sur un modèle** $\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}}$, égal au total de (5.4) et de (5.5); 2) l'estimateur de la variance pour un échantillonnage **à deux phases** $N^2(1/n - 1/N)S_{yr}^2 + N^2(1/m - 1/n)\sum_r e_k^2/(m-1)$, qui découle de la théorie ordinaire de l'échantillonnage à deux phases avec l'hypothèse d'un sous-échantillonnage EASSR de m répondants parmi n unités de l'échantillon initial (Rao 1990); et 3) l'estimateur de la variance **non corrigé ordinaire** $N^2(1/n - 1/N)S_{y\blacklozenge s}^2$, obtenu en faisant comme si les valeurs imputées étaient aussi bonnes que les valeurs réelles. Les résultats figurent dans le tableau qui suit:

Estimateur de $\hat{V}$	Biais relatif de $\hat{V}$ en %		
	Mécanisme 1	Mécanisme 2	Mécanisme 3
Fondé sur un modèle	−0.20	−4.64	−3.99
À deux phases	9.95	−12.49	−1.11
Non corrigé ordinaire	−25.73	−37.90	−33.21

Le biais relatif d'un estimateur $\hat{V}$ a été calculé sous la forme $\{\text{moy}(\hat{V}) - \text{var}(\hat{t}_\blacklozenge)\}/\text{var}(\hat{t}_\blacklozenge)$, où $\text{moy}(\hat{V})$ est la moyenne des 100,000 valeurs de $\hat{V}$, et $(\hat{t}_\blacklozenge)$ est la variance des 100,000 valeurs de $\hat{t}_\blacklozenge$. La simulation montre que l'estimateur de la variance fondé sur un modèle $\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}}$ est quasi non biaisé pour les trois mécanismes de réponse. D'une certaine façon, ce n'est pas surprenant parce que la population a été créée de manière à concorder avec le modèle d'imputation par le quotient. Les mécanismes 1 et 2 sont du type non neutre et ne répondent pas à la condition (a) de la section 4, requise pour que $\hat{V}_{\text{tot}}$ soit dépourvu de biais. Fait intéressant, toutefois, le biais de $\hat{V}_{\text{tot}}$ dans cet exemple demeure quand même faible. L'estimateur à deux phases fonctionne bien pour le mécanisme 3 (réponse uniforme), cas pour lequel il a été conçu; dans les autres cas, il est biaisé. Enfin, considérer les valeurs imputées comme des données réelles entraîne, comme prévu, une sous-estimation prononcée de la vraie variance pour les trois mécanismes. Une étude de Monte Carlo plus approfondie de l'estimation par le quotient est décrite dans Lee, Rancourt et Särndal (1992). Cet article donne une idée de l'effet d'une définition déficiente du modèle d'imputation, sujet également abordé dans Rao (1992).

6. VALEURS IMPUTÉES AYANT UN RÉSIDU AJOUTÉ

Nous pouvons distinguer deux types de valeurs imputées: 1) la valeur imputée $y_{\text{imp},k}$ est constituée d'une valeur prévue seulement, $y_{\text{pred},k}$, par exemple quand il s'agit d'une valeur provenant d'une ligne ou d'une surface résultant d'un ajustement de régression. Par exemple, dans la méthode d'imputation par le quotient courant utilisée ci-dessus, $y_{\text{imp},k} = y_{\text{pred},k} = \hat{B}x_k$, où $\hat{B} = (\sum_r y_k)/(\sum_r x_k)$; 2) la valeur imputée $y_{\text{imp},k}$ est constituée d'une valeur prévue et d'un résidu, c.-à-d. que $y_{\text{imp},k} = y_{\text{pred},k} + e_k^*$. Le terme résiduel, dont le rôle est de rendre les valeurs imputées plus conformes aux observations réelles, peut être obtenu par échantillonnage des résidus $e_k = y_k - y_{\text{pred},k}$ calculés pour les unités répondantes $k \in r$. Un moyen de le faire est exposé ci-dessous. Ce type d'imputation est parfois recommandé, dans la littérature, comme moyen de préserver les distributions des valeurs imputées; voir, par exemple, l'analyse faite dans Little (1988). Le processus d'imputation devient alors plus ardu à réaliser et, pour les besoins du SGE (dont le but principal est une estimation valide de la précision des estimations de l'enquête), il n'est pas clair que les avantages réalisés justifient l'effort additionnel.

Nous présentons cependant une marche à suivre pour l'imputation par "valeur prévue plus résidu" dans le cas où le modèle d'imputation par le quotient courant est pris comme point de départ: Pour $k \in r$, calculons $e_k = y_k - \hat{B}x_k$, où $\hat{B} = (\sum_r y_k)/(\sum_r x_k)$, puis $\tilde{e}_k = e_k/\sqrt{x_k}$. On obtient ainsi un ensemble de m "résidus standardisés" $\tilde{e}_k$. Ensuite, pour une unité $k \in s - r$, calculons $e_k^0 = \sqrt{x_k}\tilde{e}_k$, où $\tilde{e}_k$ est tiré par EASAR de l'ensemble, et x_k appartient à l'unité exigeant une imputation. Les unités ayant une valeur x élevée ont alors tendance à obtenir des résidus e_k^0 plus élevés, ce qui est conforme au modèle. Posons ensuite $e_k^* = e_k^0 - (\sum_{s-r} e_k^0)/(n - m)$. Pour $k \in s-r$, imputons $y_{\text{imp},k} = \hat{B}x_k + e_k^*$, $k \in s-r$; pour $k \in r$, nous avons les observations réelles, y_k. Puisque la somme des e_k^* doit par définition égaler zéro pour l'ensemble des unités de $s - r$, l'estimateur ponctuel est donné par $\hat{t}_\blacklozenge = (N/n)\sum_s y_{\blacklozenge k} = N\bar{x}_s\bar{y}_r/\bar{x}_r$ comme à la section 5, mais sa variance est différente. On peut montrer que $E_\xi E_s E_r E_\# (S^2_{y\blacklozenge s} - S^2_{ys}) \approx 0$, où $E_\#$ désigne la moyenne par rapport à la sélection au hasard d'un résidu standardisé. Autrement dit, la différence entre la variance calculée à partir des données après imputation, $S^2_{y\blacklozenge s}$, et la variance inconnue d'un échantillon formé entièrement d'observations réelles, S^2_{ys}, est approximativement égale à zéro en moyenne. Nous pouvons utiliser $\hat{V}_{\text{sam}} = N^2(1/n - 1/N)S^2_{y\blacklozenge s}$ en tant qu'estimateur à peu près globalement non biaisé de la composante variance d'échantillonnage. Il n'est pas nécessaire maintenant d'ajouter une correction $\hat{V}_{\text{dif}}$. Toutefois, il reste toujours à calculer un estimateur de la variance d'imputation $V_{\text{imp}} = N^2(1/m - 1/n)C_1\sigma^2$ et à l'ajouter à $\hat{V}_{\text{sam}}$.

7. CONCLUSIONS

Les travaux continus sur les techniques d'estimation de la variance dont il est fait état dans le présent article visent à étendre l'analyse: 1) aux méthodes d'imputation fondées sur un modèle qui sont seulement implicites, en particulier la méthode d'imputation du plus proche voisin; 2) au cas où plusieurs méthodes d'imputation sont utilisées dans la même étude.

Deville et Särndal (1992) présentent les résultats d'une extension dans laquelle l'estimateur de Horwitz-Thompson $\hat{t} = \sum_s y_k/\pi_k$ sert de prototype. L'estimateur utilisant les données après imputation est alors donné par

$$\hat{t}_\blacklozenge = \sum_r y_k/\pi_k + \left(\sum_{s-r} x_k/\pi_k\right)' \hat{B} = \sum_s y_k/\pi_k - \sum_{s-r} e_k/\pi_k,$$

où $e_k = y_k - x_k' B$ est le résidu d'imputation pour l'unité k, obtenu par régression multiple.

REMERCIEMENTS

J'exprime ma reconnaissance à M. Hidiroglou, P. Lavallée, Y. Leblond, H. Lee et G. Reinhard de Statistique Canada pour leur collaboration aux travaux ayant mené à cet article. Les commentaires de deux arbitres ont permis d'améliorer le manuscrit initial, et ont été vivement appréciés.

BIBLIOGRAPHIE

DEVILLE, J.C., et SÄRNDAL, C.-E. (1992). Variance estimation for survey data with regression imputation. Rapport technique.

HERZOG, T.N., et RUBIN, D.B. (1983). Using multiple imputations to handle nonresponse in surveys. Dans *Incomplete Data in Sample Surveys*, (Éds. W.G. Madow, I. Olkin et D.B. Rubin). New York: Academic Press, 209-245.

FAY, R.E. (1991). A design-based perspective on missing data variance. Proceedings, 1991 Annual Research Conference, U.S. Bureau of the Census, 429-440.

LEE, H., RANCOURT, E., et SÄRNDAL, C.-E. (1992). Experiments with variance estimation from survey data with imputed values. Rapport, Division des méthodes d'enquêtes-entreprises, Statistique Canada, soumis pour publication.

LITTLE, R.J.A. (1988). Missing-data adjustments in large surveys (avec discussion). *Journal of Business and Economic Statistics*, 6, 287-301.

PRITZKER, L., OGUS, J., et HANSEN, M.H. (1965). Computer editing methods: some applications and results. *Bulletin de l'Institut International de Statistique*, 41, 442-466.

RAO, J.N.K. (1990). Variance estimation under imputation for missing data. Ce manuscrit a été lu avec la gracieuse permission de l'auteur.

RAO, J.N.K. (1992). Jackknife variance estimation under imputation for missing survey data. Ce manuscrit a été lu avec la gracieuse permission de l'auteur.

RUBIN, D.B. (1986). Initiation a l'imputation multiple pour les cas de non-réponse. *Techniques d'enquête*, 12, 41-52.

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

SÄRNDAL, C.-E. (1990). Estimation of precision in the generalized estimation system when imputation is used. Rapport, Secteur de l'informatique et de la méthodologie, Statistique Canada, mars 31, 1990.