ESTIMATION DE VARIANCE EN PRÉSENCE D'IMPUTATION : OÙ EN SOMMES-NOUS ?

Eric Rancourt[1]

1. Introduction

Peu importe le temps et l'effort mis à la préparation d'une enquête, il y a toujours une partie des unités de l'échantillon pour laquelle l'information désirée n'est pas obtenue. Ce problème de non-réponse est souvent traité par l'imputation car elle permet d'utiliser à profit l'information partielle recueillie et de créer un ensemble de données complet. Plusieurs arguments peuvent être invoqués pour ou contre l'imputation mais le contexte du présent document est celui où l'imputation est considérée comme étant un fait accompli. Le lecteur trouvera une excellente discussion sur l'imputation dans Kovar et Whitridge (1995). Puisque l'on travaille à partir de données d'enquête, il s'ensuit que le processus d'estimation est entaché d'erreurs dont, entre autres, l'erreur échantillonnale. Afin de connaître la précision des estimations, on utilise habituellement le calcul de l'estimation de la variance. Si l'imputation est utilisée pour pallier la non-réponse, les estimations et leur précision seront affectées et ce, même avec la meilleure méthode d'imputation. On se doit donc de quantifier l'impact de l'imputation. Ce faisant, on pourra non seulement mieux informer les utilisateurs de la qualité réelle des estimations, mais on pourra également tenter de contrôler l'impact de l'imputation et de le réduire.

1. Eric Rancourt, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa (Ontario) Canada, K1A 0T6.

Si on utilise les méthodes pour ensembles de données complets pour estimer la variance, on ne tient pas compte du fait que certaines données ont été imputées. C'est-à-dire que l'on fait l'hypothèse implicite que les données imputées se comportent comme des observations réelles. Dans ce cas, la variance totale sera sous-estimée. Plusieurs solutions à ce problème ont été développées et cet article décrit les suivantes :

1) l'imputation multiple, Rubin (1978, 1987)

2) l'approche assistée d'un modèle, Särndal (1990, 1992) et Deville et Särndal (1991) ;

3) l'approche à deux phases, Rao (1990) et Rao et Sitter (1995) ;

4) la technique du jackknife, Rao (1991) et Rao et Shao (1992) ;

6) la méthode pour imputation hot-deck, Provost (1995) ;

7) le bootstrap, Shao et Sitter (1996) ;

8) la méthode d'imputation de tous les cas, Montaquila et Jernigan (1997) ;

9) la méthode des échantillons balancés répétés (ou BRR), Shao, Chen et Chen (1998).

Cet article est divisé comme suit. À la section 2 on discute de l'importance de tenir compte de l'imputation dans le calcul de la variance. Par la suite, la section 3 contient une description des composantes de la variance totale. Dans la section 4 se trouvent les descriptions de ces diverses méthodes permettant le calcul correct de la variance, suivies d'une comparaison des méthodes à la section 5. On poursuit avec les points à considérer lors de l'application des méthodes à la section 6, et une présentation de l'approche adoptée à Statistique Canada à la section 7

2. Importance d'estimer correctement la variance

Lors de la production d'estimations à partir de données d'enquêtes, l'intérêt principal est évidemment pour les estimations ponctuelles. C'est pourquoi une importance mitigée est parfois accordée au calcul de la variance. Cependant, tel que mentionné dans Gagnon, Lee, Provost, Rancourt et Särndal (1997), l'estimation de la variance est très importante car elle permet :

1) de fournir une mesure de qualité des estimations ;

2) d'aider à tirer des conclusions correctes ;

3) aux agences statistiques d'informer les utilisateurs de la qualité des données.

Dans le cas de données imputées, il est encore plus important de mesurer la précision des estimations. En effet, l'imputation ajoute un processus supplémentaire pouvant être source d'erreurs. Ainsi, l'estimation de la variance qui tient compte de l'imputation permet, en plus des trois points cités plus haut :

4) de mieux connaître l'impact de l'imputation ;

5) d'améliorer l'estimation de la variance totale ;

6) d'effectuer une meilleure répartition des ressources entre un échantillon plus grand et un meilleur processus de vérification et imputation (selon la taille relative de la variance due à l'échantillonnage et la variance due à l'imputation).

Il est à noter que l'importance de tenir compte de l'imputation dans le calcul de la variance varie selon les conditions de l'enquête. Selon ces conditions (le taux de réponse, la méthode d'imputation, la fraction de sondage, la méthode d'estimation et la qualité des variables auxiliaires utilisées), la variance due à l'imputation pourrait atteindre un très grand pourcentage de la variance totale. Et même, dans le cas d'un recensement, il serait possible de calculer la variance due à l'imputation qui serait, par définition, 100% de la variance totale.

3. Imputation et estimation : Cadre théorique

À partir d'une population $U = \{1,...,k,...,N\}$, on tire un échantillon s. Le poids de sondage pour l'unité k est $a_k = 1/\pi_k$, où π_k est la probabilité de sélection. On désire estimer le total de la variable d'intérêt y, $Y_U = \sum_U y_k$. En l'absence de non-réponse, on utiliserait

$$\hat{Y}_s = \sum_s a_k g_k y_k$$

où, par exemple dans le cas de l'estimateur généralisé de régression (GREG) on a $g_k = 1 + (\sum_U \mathbf{x}_k - \sum_s \mathbf{x}_k/\pi_k)'(\sum_s \mathbf{x}_k \mathbf{x}_k'/\sigma_k^2 \pi_k)^{-1} \mathbf{x}_k/\sigma_k^2$ qui est le facteur d'ajustement au total de données auxiliaires, $\mathbf{X} = \sum_U \mathbf{x}_k$.

En présence de non-réponse, l'échantillon se retrouve scindé en deux parties : l'ensemble r des répondants et l'ensemble o des non-répondants. On a alors $o = s - r$. Pour traiter la non-réponse de l'unité k, $k \in o$, on impute une valeur

$\hat{y}_k$. Si cette imputation utilise une variable auxiliaire, elle sera dénotée par z_k. L'ensemble de données après imputation est $\{y_{\bullet k} : k \in s\}$, où

$$y_{\bullet k} = \begin{cases} y_k & \text{si } k \in r \\ \hat{y}_k & \text{si } k \in o. \end{cases}$$

Ainsi, en présence d'imputation, si la pondération demeure inchangée, l'estimation de Y_U devient

$$\hat{Y}_{\bullet s} = \sum_s a_k g_k y_{\bullet k}.$$

La quantité que l'on désire estimer étant Y_U, on peut décomposer l'erreur totale de la façon suivante :

$$\hat{Y}_{\bullet s} - Y_U = (\hat{Y}_s - Y_U) + (\hat{Y}_{\bullet s} - \hat{Y}_s),$$

où $\hat{Y}_s - Y_U$ est l'erreur d'échantillonnage et

$$\hat{Y}_{\bullet s} - \hat{Y}_s = \sum_o a_k g_k (\hat{y}_k - y_k)$$

est l'erreur d'imputation.

À partir des deux termes d'erreur ci-haut mentionnés, on pourra évaluer l'écart quadratique moyen. En supposant que l'imputation permette de faire une estimation sans biais, on aura plutôt l'expression de la variance. Voir par exemple Särndal (1992) pour la dérivation. L'expression de la variance totale que l'on obtient se compose de la variance due à l'échantillonnage, de la variance due à l'imputation et d'un terme mixte. Un estimateur de variance pour $\hat{Y}_{\bullet s}$ est donc

$$\hat{V}_{\text{TOT}} = \hat{V}_{\text{ÉCH}} + \hat{V}_{\text{IMP}} + \hat{V}_{\text{MIX}}$$

où $\hat{V}_{\text{ÉCH}}$ est l'estimateur de la variance échantillonnale, $\hat{V}_{\text{IMP}}$ est l'estimateur de la variance due à l'imputation et $\hat{V}_{\text{MIX}}$ correspond à deux fois la covariance entre les deux erreurs. Ce terme mixte est, dans plusieurs cas, relativement petit par rapport aux deux autres.

Dans le cas de l'échantillonnage aléatoire simple sans remise on pourrait, pour l'estimateur de la variance échantillonnale, $\hat{V}_{\text{ÉCH}}$, utiliser la formule habituelle d'estimation de variance,

$$\hat{V}_{\text{ORD}} = N^2 \frac{(1-f)}{n} \sum_s \frac{(y_{\bullet k} - \bar{y}_{\bullet s})^2}{n-1},$$

mais elle sous-estimerait la variance échantillonnale puisque les calculs se font sur l'ensemble de données après imputation. Il faudrait plutôt utiliser

$$\hat{V}_{\text{ÉCH}} = N^2 \frac{(1-f)}{n} \sum_s \frac{(y_k - \bar{y}_s)^2}{n-1},$$

mais y_k n'est pas disponible pour les non-répondants. Il faut donc estimer $\hat{V}_{\text{ÉCH}}$ par

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

où $\hat{V}_{\text{DIF}}$ sera obtenu à l'aide d'un estimateur de $\hat{V}_{\text{ÉCH}} - \hat{V}_{\text{ORD}}$. Par exemple, dans Lee, Rancourt et Särndal (1995) on présente $\hat{V}_{\text{DIF}}$ pour l'imputation par la moyenne, par quotient, hot-deck et plus proche voisin. Pour certaines méthodes d'imputation, l'ensemble de données après imputation contient des valeurs ayant une variabilité suffisante pour que $\hat{V}_{\text{ORD}}$ soit assez près de $\hat{V}_{\text{ÉCH}}$. Il n'est donc pas nécessaire d'utiliser $\hat{V}_{\text{DIF}}$. Sous plusieurs conditions, c'est le cas de l'imputation par plus proche voisin et de l'imputation hot-deck. Dans Rao et Sitter (1997), on trouve une méthode d'imputation hot-deck construite spécifiquement pour éviter l'utilisation de $\hat{V}_{\text{DIF}}$.

On constate maintenant que, pour estimer correctement la variance totale, il ne suffit pas d'estimer la variance due à l'imputation, mais il faut également estimer correctement la variance due à l'échantillonnage.

4. Méthodes d'estimation de variance en présence d'imputation

Cette section décrit les méthodes d'estimation de variance en présence d'imputation mentionnées à la section 1. Pour simplifier la notation, le cas de l'échantillonnage aléatoire simple sans remise est traité.

4.1 Imputation multiple

La méthode de l'imputation multiple a été développée par Rubin (1978, 1987). Elle consiste à imputer, selon une méthode donnée, plusieurs valeurs pour chaque donnée manquante créant ainsi plusieurs ensembles de données après imputation $j = 1,...,J$. On peut donc utiliser les méthodes d'estimation habituelles sur chaque ensemble de données. Une fois que l'on a estimé la variabilité dans chaque ensemble de données, et entre les ensembles de données après imputation, on peut combiner les résultats et on a

$$\hat{V}_{\text{IM}} = \hat{V}_{\text{INTERNE}} + \hat{V}_{\text{ENTRE}}$$

qui ressemblent à $\hat{V}_{\text{ÉCH}}$ et $\hat{V}_{\text{IMP}}$. Plus spécifiquement, l'expression se lit comme suit:

$$\hat{V}_{\text{IM}} = \frac{1}{M}\sum_{j=1}^{M} N^2 \frac{1-f}{n} S^2_{y\bullet js} + \left(1+\frac{1}{M}\right)\frac{N^2}{M-1}\sum_{j=1}^{M}(\bar{y}_{\bullet js} - \bar{y}_{\bullet\bullet s})^2$$

où M est le nombre d'ensembles de données après imputation, $f = n/N$,

$$S^2_{y\bullet js} = \frac{1}{n-1}\sum_s (y_{\bullet jk} - \bar{y}_{\bullet js})^2 \text{ et } \bar{y}_{\bullet\bullet s} = \frac{1}{M}\sum_{j=1}^{M} \bar{y}_{\bullet js}.$$

4.2 Approche assistée d'un modèle

L'approche assistée d'un modèle a été développée par Särndal (1990, 1992) et Deville et Särndal (1991, 1994). C'est une méthode pour l'imputation simple qui consiste à utiliser un modèle pour estimer la variance due à l'imputation en plus de celle due à l'échantillonnage. L'objectif est d'obtenir un estimateur de V_{TOT} pour

$\hat{Y}_{\bullet s}$ en construisant des estimateurs des composantes $V_{\text{ÉCH}}, V_{\text{IMP}}$, et V_{MIX} en utilisant un modèle de la forme

$$\xi : y_k = \beta z_k + \varepsilon_k \ ; \ E_\xi(\varepsilon_k) = 0; \ E_\xi(\varepsilon_k^2) = \sigma^2 z_k; \text{ et } E_\xi(\varepsilon_k \varepsilon_{k'}) = 0 \text{ pour } k \neq k',$$

où z est la variable d'imputation dont la valeur z_k est disponible au moins pour $k \in s$.

Les composantes $\hat{V}_{\text{ÉCH}}$, $\hat{V}_{\text{IMP}}$, et $\hat{V}_{\text{MIX}}$ doivent satisfaire $E_\xi(\hat{V}_{\text{ÉCH}} - V_{\text{ÉCH}}) = 0$, $E_\xi(\hat{V}_{\text{IMP}} - V_{\text{IMP}}) = 0$ et $E_\xi(\hat{V}_{\text{MIX}} - V_{\text{MIX}}) = 0$. En particulier, $\hat{V}_{\text{ÉCH}}$ est construit à l'aide de deux termes. Le premier consiste en la "formule ordinaire" de variance $\hat{V}_{\text{ORD}} = N^2 \frac{1-f}{n} S^2_{y\bullet s}$ calculée sur l'ensemble de données après imputation, avec $S^2_{y\bullet s} = \frac{1}{n-1} \sum_s \{y_{\bullet k} - (\sum_s y_{\bullet k}/n)\}^2$. On y ajoute le terme, $\hat{V}_{\text{DIF}}$, construit de façon à satisfaire $E_\xi\{\hat{V}_{\text{DIF}}\} = \frac{N^2(1-f)}{n} E_\xi\{S^2_{ys} - S^2_{y\bullet s}\}$. On obtient $\hat{V}_{\text{ÉCH}} = \hat{V}_{\text{ORD}} + \hat{V}_{\text{DIF}}$, et donc

$$\hat{V}_{\text{AM}} = \hat{V}_{\text{ORD}} + \hat{V}_{\text{DIF}} + \hat{V}_{\text{IMP}} + \hat{V}_{\text{MIX}}.$$

Par exemple, dans le cas de l'imputation par le quotient (ou ratio), où $\hat{y}_k = \hat{B} z_k$ avec $\hat{B} = \sum_r y_k / z_k$, on aura :

$$\hat{V}_{\text{ORD}} = N^2 \frac{1-f}{n} S^2_{y\bullet s} \qquad \hat{V}_{\text{DIF}} = N^2 \frac{1-f}{n^2} \sum_o z_k \hat{\sigma}^2$$

$$\hat{V}_{\text{IMP}} = \frac{N^2}{n^2} \sum_o z_k \left\{ \frac{\sum_o z_k}{\sum_r z_k} + 1 \right\} \hat{\sigma}^2 \qquad \hat{V}_{\text{MIX}} = N^2 \frac{(1-f)}{n^2} \sum_o z_k \left\{ \frac{\sum_o z_k}{\sum_r z_k} - 1 \right\} \hat{\sigma}^2,$$

avec $\hat{\sigma}^2 = \sum_r e_k^2 / \sum_r z_k$ et $e_k = y_k - \hat{B} z_k$.

4.3 Approche en deux phases

Développée par Rao (1990) et Rao et Sitter (1995) à partir de l'idée de l'échantillonnage à deux phases, cette méthode suppose que l'échantillon est la première phase d'un plan de sondage, et que les répondants constituent l'échantillon de deuxième phase. De façon implicite, on suppose donc que les répondants forment un échantillon aléatoire des unités de l'échantillon. L'idée de base est d'obtenir un estimateur de variance formé d'un terme de variance due à la première phase et d'un terme issu de la deuxième phase. On a donc, dans le cas de l'imputation par le ratio

$$\hat{V}_{\mathrm{DP1}} = N^2\left(\frac{1}{n}-\frac{1}{N}\right)S_{yr}^2 + N^2\left(\frac{1}{m}-\frac{1}{N}\right)S_{er}^2.$$

où $S_{yr}^2 = \sum_r (y_k - \bar{y}_r)^2/(m-1)$.

Pour utiliser l'information auxiliaire, Rao (1990) et Rao et Sitter (1995) suggèrent plutôt l'expression suivante basée sur la même approche :

$$\hat{V}_{\mathrm{DP}} = N^2\left(\frac{1}{n}-\frac{1}{N}\right)\hat{B}^2 S_{zs}^2 + 2N^2\left(\frac{1}{n}-\frac{1}{N}\right)\hat{B}S_{zer} + N^2\left(\frac{1}{m}-\frac{1}{N}\right)S_{er}^2.$$

où $S_{zer} = \sum_r e_k z_k/(m-1)$.

4.4 Technique du Jackknife

La technique du jackknife a pour principe de recalculer l'estimateur après avoir enlevé une ou plusieurs unités de l'échantillon. La variance entre les estimations obtenues est utilisée pour obtenir une estimation de la variance de l'estimateur calculé sur l'ensemble de l'échantillon. Lorsque l'unité j est enlevée, l'estimateur du total Y_U est donné par $\hat{Y}_{\bullet s}^{(j)} = N\sum_{k\neq j\in s} y_{\bullet k}/(n-1)$. L'estimateur par le jackknife est $\hat{V} = \frac{n-1}{n}\sum_{j\in s}(\hat{Y}_{\bullet s}^{(j)} - \hat{Y}_{\bullet s})^2$.

En présence d'imputation, le jackknife, tel que défini ci-haut, sous-estime V_{TOT}. Pour cette situation, Rao and Shao (1992) proposent un estimateur de variance qui

corrige le jackknife en ajustant les valeurs imputées lorsque l'unité enlevée fait partie de l'ensemble des répondants. L'ensemble de données ajustées est donné par

$$y_{\bullet k}^{(aj)} = \begin{cases} y_k & \text{si } k \in r \\ \hat{y}_k + a_k^{(j)} & \text{si } k \in o \text{ et } j \in r \\ \hat{y}_k & \text{si } k \in o \text{ et } j \in o \end{cases}$$

où $y_{\bullet k}^{(aj)}$ est la valeur imputée ajustée et $a_k^{(j)}$ est l'ajustement. L'estimateur de variance par le jackknife est alors

$$\hat{V}_{\text{JK}} = \frac{n-1}{n} \sum_{j \in s} (\hat{Y}_{\bullet s}^{(aj)} - \hat{Y}_{\bullet s}^{(a)})^2$$

où $\hat{Y}_{\bullet s}^{(aj)} = \frac{N}{n-1} \sum_{k \neq j \in s} y_{\bullet k}^{(aj)}$ et $\hat{Y}_{\bullet s}^{(a)} = \frac{1}{n} \sum_{j \in s} \hat{Y}_{\bullet s}^{(aj)}$. Cet estimateur fonctionne bien lorsque la correction pour population finie (cpf), $1-f$ avec $f = n/N$, n'est pas nécessaire. Pour les situations où la cpf est requise, Lee, Rancourt et Särndal (juillet 1995) proposent la correction suivante à l'estimateur de variance par le jackknife $\hat{V}_{\text{JK}}^{*} = \hat{V}_{\text{JK}} - N\hat{S}_{yU}^2$, où $\hat{S}_{yU}^2$ est un estimateur sans biais de S_{yU}^2.

Les ajustements $a_k^{(j)}$ dépendent de la méthode d'imputation. On peut trouver les ajustements pour l'imputation par la moyenne et hot-deck dans Rao (1991), et Rao et Shao (1992), pour l'imputation par quotient dans Rao (1991) et Rao et Sitter (1995), et pour l'imputation par plus proche voisin dans Kovar and Chen (1994). Il est également intéressant de noter que Rao et Sitter (1995) présentent une linéarisation de l'estimateur de variance par la méthode du jackknife.

4.5 Bootstrap

La technique du bootstrap pour l'estimation de la variance en présence de données imputées a été développée par Shao et Sitter (1996). De même que la technique habituelle du bootstrap, elle consiste à tirer plusieurs échantillons à partir de l'échantillon d'origine en reproduisant la méthode d'échantillonnage. Cependant, dans le cas de données imputées, chaque donnée imputée se retrouvant dans l'échantillon bootstrap doit être ré-imputée par la même procédure d'imputation ayant servi à l'origine. Ainsi, l'estimateur de la variance par bootstrap est

$$\hat{V}_{\text{BOOT}} = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{Y}_{*k}^{(b)} - \overline{Y}_{*s}\right)^2$$

où B est le nombre d'échantillons bootstrap, $\hat{Y}_{*k}^{(b)}$ le total pour l'échantillon b estimé sur les données après ré-imputation, et $\overline{Y}_{*k}$ est la moyenne des totaux sur tous les échantillons bootstrap. Cette formule présuppose un échantillonnage avec remise. Dans le cas de l'échantillonnage sans remise, Shao et Sitter (1996) décrivent trois méthodes qui peuvent être employées.

4.6 Méthode pour Hot-deck

Dans le cas de l'imputation hot-deck, il existe une approche basée sur le plan de sondage qui permette d'estimer correctement la variance totale pour un mécanisme de réponse uniforme. Cette approche à été développée par Provost (1995). Elle est basée sur le fait que l'imputation hot-deck correspond à un tirage aléatoire simple d'une unité parmi les m répondants, conditionnellement à l'échantillon s et à l'ensemble de répondants r. Pour estimer la variance totale, on doit donc estimer la variance due à l'échantillonnage et la variance due à l'imputation. On a alors

$$\hat{V}_{\text{HD}} = \hat{V}_{\text{ÉCH}} + \hat{V}_{\text{IMP}}.$$

Pour l'échantillonnage aléatoire simple sans remise et un mécanisme de réponse uniforme, les deux termes sont

$$\hat{V}_{\text{ÉCH}} = \frac{mn(n-1)}{(n^2-n+m)(m-1)} N^2 \frac{(1-f)}{n}\sum_{s}\frac{(y_{\bullet k} - \overline{y}_{\bullet s})^2}{n-1}$$

$$\hat{V}_{\text{IMP}} = \frac{\left(n^2+m-n-m^2\right)(n-1)}{(n^2-n+m)(m-1)} N^2 \frac{1}{n}\sum_{s}\frac{(y_{\bullet k} - \overline{y}_{\bullet s})^2}{n-1}.$$

4.7 Méthode d'imputation de tous les cas

La technique d'imputation de tous les cas consiste à imputer une valeur à toutes les unités de l'échantillon y compris les répondants. L'estimation ponctuelle s'effectue alors en utilisant les données imputées de tout l'échantillon. Pour le calcul de la variance, on dispose donc de résidus qui vont permettre d'évaluer l'erreur d'imputation, et donc d'estimer la variance due à l'imputation. Cette méthode,

décrite dans Montaquila et Jernigan (1997) a été développée récemment pour l'imputation par donneur dans le cas de populations infinies.

4.8 La Méthode des échantillons balancés répétés (ou BRR)

L'estimation de variance en présence d'imputation a aussi été développée pour la méthode des échantillons balancés répétés. Elle consiste en un ajustement des données imputés similaire à celui pour le jackknife, pour chacun des échantillons balancés. Une fois ces ajustements effectués, on peut utiliser la formule

$$\hat{V}_{\mathrm{BRR}} = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{Y}_{\bullet k}^{(r)} - \bar{Y}_{\bullet}\right)^2,$$

où R est le nombre d'échantillons balancés, et $\hat{Y}_{\bullet}^{(r)}$ est l'estimateur pour l'échantillon r, calculé avec les ajustements.

La méthode est décrite dans Shao, Chen et Chen (1998), qui paraîtra bientôt.

5. Comparaisons des méthodes

Plusieurs caractéristiques différencient les méthodes d'estimation de variance présentées à la section précédente. La discussion qui suit aborde sept thèmes qui sont ensuite résumés dans un tableau général. Les termes entre parenthèses sont utilisés dans le tableau 1 à la section 5.8 pour référer aux différentes caractéristiques des méthodes.

5.1 Possibilité d'obtenir une estimation de $V_{ÉCH}$ et de V_{IMP} (V_{IMP})

Dans toutes les méthodes sauf le jackknife, le bootstrap et le BRR, on obtient différents termes qui représentent plus ou moins bien $\hat{V}_{\mathrm{ÉCH}}$ et $\hat{V}_{\mathrm{IMP}}$. Il peut être avantageux de disposer de cette décomposition de l'estimation de la variance totale pour mieux connaître l'impact de l'imputation. En connaissant l'importance relative de $\hat{V}_{\mathrm{ÉCH}}$ et $\hat{V}_{\mathrm{IMP}}$ par rapport à $\hat{V}_{\mathrm{TOT}}$, on dispose ainsi d'une mesure qui pourrait permettre de mieux répartir, dans les enquêtes répétées, les ressources entre deux

objectifs importants, soient l'amélioration du plan de sondage ou du système d'imputation.

5.2 Nombre d'imputations requises pour chaque unité (# Imp)

Une caractéristique qui est propre aux méthodes d'estimation de variance en présence d'imputation est le nombre d'ensembles de données créés. C'est-à-dire que pour l'imputation simple, il n'y a qu'un seul ensemble de données alors qui pour l'imputation multiple, il y en a plusieurs. Ce surplus d'ensembles à entreposer et à maintenir peut engendrer des coûts supplémentaires. Par contre, l'imputation multiple permet l'utilisation des méthodes habituelles d'estimation de variance. Il est ainsi possible de les appliquer sur chaque ensemble de données pour ensuite combiner les résultats afin d'obtenir une estimation de la variance totale qui inclut la variance due à l'imputation. Il est aussi à noter que le bootstrap est en fait une technique d'imputation multiple, puisque le processus d'imputation doit être répété à chaque itération du bootstrap.

5.3 Nécessité d'identifier les répondants et les non-répondants (Flag)

Pour les méthodes d'estimation de variance qui s'appliquent à l'imputation simple, il est nécessaire de savoir si les données de l'échantillon font partie de l'ensemble des répondants ou de l'ensemble des non-répondants. En d'autres termes, on doit disposer d'identificateurs d'imputation pour chaque unité de l'échantillon. De plus, on doit également connaître la méthode d'imputation utilisée pour obtenir la bonne formule d'estimation de variance et pour utiliser la bonne correction pour la technique du jackknife.

5.4 Restriction sur les méthodes d'imputation (Méthode)

Toutes les méthodes d'estimation de variance ont évidemment des limites, mais certaines ont été développées pour des méthodes d'imputation spécifiques. C'est le cas de la méthode hot-deck qui, comme son nom l'indique, est conçue pour l'imputation hot-deck et de la méthode d'imputation de tous les cas qui s'applique présentement au cas de l'imputation par donneur. En ce qui concerne les autres méthodes, elles semblent pouvoir s'appliquer à de plus grandes familles de méthodes d'imputation. Également, la méthode de l'imputation multiple requiert que la méthode d'imputation soit « propre » au sens décrit dans Rubin (1987).

5.5 Hypothèses sur le mécanisme de non-réponse (Non-rép)

Les mécanismes de non-réponse peuvent être classés en trois groupes, comme décrit dans Rubin (1976) et dans Rancourt, Lee et Särndal (1994) :

1) Mécanismes où les données sont manquantes de façon aléatoire (MCAR[2]. ou uniforme) où la non-réponse ne dépend d'aucune variable de l'échantillon et est distribuée uniformément.

2) Mécanismes où les données sont manquantes de façon aléatoire (MAR ou non-confondu) conditionnellement à une ou plusieurs variables auxiliaires. Dans ce cas, la non-réponse peut dépendre d'une variable auxiliaire mais ne dépend pas de la variable d'intérêt.

3) Mécanismes où les données sont manquantes de façon non aléatoire (NMAR ou confondu) où la non-réponse dépend de la variable d'intérêt.

Toutes les méthodes décrites dans cet article fonctionnent dans le cas 1 et sont vulnérables au mécanisme de réponse dans le cas 3. Dans le cas 2, les méthodes utilisant un modèle (et donc des variables auxiliaires) et les méthodes de ré-échantillonnage se comportent assez bien selon les conditions de l'enquête. Par contre, une méthode comme l'approche en deux phases suppose que l'ensemble des répondants soit un échantillon aléatoire simple de l'échantillon. La méthode est donc très bien adaptée à la situation 1 mais n'est pas tellement robuste aux situations 2 et 3.

5.6 Utilisation d'un modèle (Modèle)

Les méthodes d'imputation peuvent toutes être exprimées de façon explicite ou implicite par un modèle. Si le modèle supposé pour l'imputation n'est pas bon, l'estimation de variance sera moins exacte. Par contre, si le modèle est bon, l'estimation de la variance sera sans biais. Un modèle peut être utilisé de façon implicite ou explicite dans le développement d'une approche d'estimation de variance. Seule la méthode assistée d'un modèle en fait une utilisation explicite.

2. Les acronymes MCAR, MAR et NMAR proviennent des termes anglais : « Missing Completely At Random », « Missing At Random », et « Not Missing At Random ».

5.7 Nature des utilisateurs des données après imputation (À qui)

Après la diffusion des résultats d'une enquête, ce sont les clients se procurant les données qui seront les utilisateurs. Par contre, en ce qui a trait au calcul de la variance, on peut distinguer deux types d'analystes de données :

1)Les « imputeurs », qui font eux-mêmes l'analyse et fournissent ensuite les estimations et une mesure de leur précision aux clients.

2)Les clients, qui font le calcul des mesures de précision lors de leurs analyses, après avoir reçu des « imputeurs » l'ensemble de données après imputation.

Dans le premier cas, il n'est pas nécessaire de construire plusieurs ensembles de données pouvant être analysées à l'aide de logiciels simples, puisque l'imputeur dispose des connaissances et de toute l'information requise pour évaluer la précision des estimations. Dans le deuxième cas, il est par contre important de fournir les outils dont l'analyste a besoin pour effectuer ses calculs. C'est dans ce dernier cas, que l'imputation multiple apparaît comme une excellente solution. En effet, une fois les multiples ensembles de données créés, l'analyste n'a qu'à répéter son travail sur chacun des ensembles de données, pour ensuite combiner les résultats selon la formule donnée à la section 4.1. Par contre, l'analyste doit être disposé à prendre le temps de répéter son analyse et le stockage de J ensembles de données n'est pas intéressant.

5.8 Différences entre les méthodes

Le tableau 1 présente un résumé de chacune des caractéristiques mentionnées ci-haut. Pour chaque caractéristique, les entrées sont les suivantes :

Vimp :	Peut-on séparer $\hat{V}_{\text{ÉCH}}$ et $\hat{V}_{\text{IMP}}$?	(Oui, Non)
#Imp :	Nombre d'imputations requises :	(1, >1)
Flag :	Besoin d'identifier répondants et non-répondants?	(Oui, Non)
Méthodes :	Restriction sur la méthode d'imputation :	(Méthode)
Non-rép :	Hypothèse sur la non-réponse :	(Uniforme, confondu)
Modèle :	La méthode utilise-t-elle explicitement un modèle?	(Oui, Non)
À qui :	Utilisateurs de l'ensemble de données :	(Interne, Externe)

Tableau 1
Caractéristiques des méthodes d'estimation de variance

	Vimp	#imp	Flag	Méthodes	Non-rép.	Modèle	À qui
Imputation Multiple	Oui	>1	Non	Propre	Pas confondu	Non	Externe
Assistée Modèle	Oui	1	Oui		Pas confondu	Oui	Interne
Jackknife	Non	1	Oui		Pas confondu	Non	Interne
2 phases	Oui	1	Oui		Uniforme	Non	Interne
Hot-deck	Oui	1	Oui	Hot-deck	Pas confondu	Non	Interne
Bootstrap	Non	>1	Oui		Pas confondu	Non	Interne
Tous les cas	Oui	1	Oui	Donneur	Pas confondu	Non	Interne
BRR	Non	1 ou >1	Oui		Pas confondu	Non	Interne

Pour plus de renseignement sur les différences, voir Lee, Rancourt et Särndal (1994) et Kovar et Chen (1994) qui ont effectué des expériences de Monte-Carlo et comparé certaines de ces méthodes sous différents modèles de population et types de non-réponse.

5.9 Disponibilité des méthodes

Il n'y a pas encore de logiciel d'estimation qui offre vraiment la possibilité de tenir compte de l'imputation dans l'estimation de la variance. Cependant, pour un système comme le Système généralisé d'estimation (SGE) de Statistique Canada décrit dans Estevao, Hidiroglou et Särndal (1995), on présente dans Lee, Rancourt et Särndal (août 1995) les éléments de base de l'estimation en présence d'imputation. De plus, des développements sont en cours à Statistique Canada sur un prototype appelé SIMPVAR afin de tenir compte de l'imputation. Ce système est brièvement décrit à la section 7.

6. Points à considérer lors de la mise en oeuvre

L'estimation de variance, et plus particulièrement dans le cas où il y a eu de l'imputation, est un problème souvent considéré seulement vers la fin du traitement des données. On devrait plutôt considérer ce problème dans son ensemble. C'est-à-dire que, dès le développement de l'enquête, il faut planifier l'estimation de

variance de façon à ne pas se retrouver face à un manque d'information pouvant être nécessaire au calcul d'une estimation de variance qui tient compte de l'imputation. La liste (non exhaustive) suivante contient des points qui peuvent contribuer à simplifier ou même à rendre possible l'estimation de variance. Plus de détails sont fournis dans Rancourt (1996).

1) L'imputation et l'estimation ne doivent pas être considérées comme deux approches séparées, mais comme deux sous-étapes d'un seul et même processus.

2) La création des groupes d'imputation devrait tenir compte (et se rapprocher) autant que possible des domaines d'estimation.

3) Les méthodes d'imputation utilisant de l'information auxiliaire contribuent à une robustesse accrue face aux différents mécanismes de non-réponse possibles.

4) Les groupes d'imputation ne doivent pas être basés uniquement sur les combinaisons possibles des variables catégoriques, mais également sur l'adéquation du modèle d'imputation.

5) L'utilisation des méthodes d'imputation ayant une composante stochastique permet de préserver les distributions et facilite le calcul de la variance. Voir par exemple Rao et Sitter (1997).

6) Des identificateurs (flags) de réponse et de méthode d'imputation doivent être assignés et conservés.

7) Il est préférable de restreindre le nombre de méthodes d'imputation utilisées à l'intérieur du même groupe d'imputation pour préserver la cohérence des données. Cette situation est abordée dans Rancourt, Lee et Särndal (1993).

8) Non seulement les méthodologistes / statisticiens doivent participer à l'élaboration de l'imputation et de l'estimation, mais il est essentiel d'y faire participer les spécialistes du sujet de l'enquête.

9) La stratégie globale doit être simple.

7. Estimation de variance en présence d'imputation à Statistique Canada

Cette section présente un aperçu des caractéristiques du Système généralisé d'estimation (SGE) de Statistique Canada. Ce système est conçu pour le cas d'ensembles de données complets, mais les plans de développement prévoient l'incorporation de méthodes d'estimation de variance tenant compte de l'imputation.

À cette fin, un prototype, SIMPVAR (Système pour tenir compte de l'imputation dans l'estimation de la variance), est en développement et est décrit ci-après.

7.1 Le Système généralisé d'estimation (SGE) de Statistique Canada

Le Système généralisé d'estimation (SGE), Estevao, Hidiroglou et Särndal (1995) est une application du progiciel SAS développée à Statistique Canada afin de produire des estimations par domaines pour un large éventail de situations. Plusieurs enquêtes font utilisation du SGE pour les calculs de totaux, de moyennes ou de ratios. Le système a été construit afin de satisfaire plusieurs plans de sondage et il fournit un vaste choix d'estimateurs à travers l'utilisation de l'estimateur de régression généralisé ou GREG, Särndal, Swensson et Wretman (1992). De plus, le calcul de la variance peut s'effectuer à l'aide de la formule issue de la linéarisation de Taylor ou de la technique du jackknife.

7.2 Le Système pour tenir compte de l'imputation dans l'estimation de variance (SIMPVAR)

Le SGE a été développé pour le cas d'ensembles de données complets, c'est-à-dire sans imputation. Pour le cas de données après imputation, l'approche assistée d'un modèle présentée à la section 4.2 est présentement en développement et a été incorporée dans un système appelé SIMPVAR. Les bases de ce système ont été jetées dans Gagnon, Lee, Rancourt et Särndal (1996). Éventuellement, ce système sera intégré au SGE. La version courante de SIMPVAR est très conviviale et fonctionne avec un système de menus. La plupart des conditions présentes dans le SGE peuvent être traitées dans SIMPVAR ; il suffit de fournir un identificateur pour les répondants et les non-répondants, d'indiquer la méthode d'imputation, d'identifier s'il y a lieu la variable auxiliaire utilisée pour l'imputation et d'indiquer le donneur pour l'imputation par donneur.

L'essence de SIMPVAR est de calculer la variance due à l'imputation et de l'ajouter à la variance due à l'échantillonnage, qui aurait été calculée au préalable par un autre système. On a donc la formule

$$\hat{V}_{\text{TOT}} = \hat{V}_{\text{ÉCH}} + \hat{V}_{\text{IMP}},$$

où $\hat{V}_{\text{IMP}}$ est obtenu par SIMPVAR et $\hat{V}_{\text{ÉCH}}$ par le SGE ou un autre système.

Ainsi SIMPVAR peut être perçu comme un module qui ajoute simplement une colonne (variance due à l'imputation) au fichier final contenant les estimations et qui en modifie une autre (variance totale).

8. Conclusion

Le problème d'estimation de la variance en présence d'imputation en est un qui mérite l'attention. Plusieurs méthodes ont été développées et peuvent être utilisées dans les enquêtes. À Statistique Canada, la méthode assistée d'un modèle a été retenue et est présentement en développement. Elle existe dans un système développé sous la forme d'un prototype appelé SIMPVAR et sera éventuellement incorporée dans le Système généralisé d'estimation (SGE). Plusieurs enquêtes pourront alors l'utiliser pour estimer la variance due à l'imputation, et donc mieux connaître la précision des estimations. Ceci permettra de rendre plus efficace l'assignation des ressources pour l'imputation en plus de pouvoir informer avec plus de précision les utilisateurs sur la qualité des données.

Bibliographie

DEVILLE, J.-C., SÄRNDAL, C.-E., « Estimation de la variance en présence de données imputées », *Proceedings of Invited Papers for the 48th Session of the International Statistical Institute*, Book 2, Subject 17, 3e17, 1991.

DEVILLE, J.-C., SÄRNDAL, C.-E., « Variance Estimation for the Regression Imputed Horvitz-Thompson Estimator », *Journal of Official Statistics*, 10, 381-394, 1994.

ESTEVAO, V., HIDIROGLOU, M.A., SÄRNDAL, C.-E., « Methodological principles for a generalized estimation system at Statistics Canada », *Journal of Official Statistics,* 11, 181-204, 1995.

GAGNON, F., LEE, H., RANCOURT, E., SÄRNDAL, C.-E., « Estimating the variance of the Generalized regression estimator in the presence of imputation for the Generalized Estimation System », *Recueil de la section de méthodologie d'enquête*, Société Statistique du Canada, 151-156, juin 1996.

GAGNON, F., LEE, H., PROVOST, M., RANCOURT, E., SÄRNDAL, C.-E., « Estimation de la variance en présence d'imputation », *Recueil du Symposium 97 : Nouvelles orientations pour les enquêtes et les recensements,* à paraître, Statistique Canada, Ottawa, novembre 1997.

KOVAR, J.G., CHEN, E., « Méthode du jackknife pour l'estimation de la variance en présence de données imputées », *Technique d'enquête*, 20, 47-55, 1994.

KOVAR, J.G., WHITRIDGE, P.J., « Imputation of Business Survey Data », *Business Survey Methods,* Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J. et Kott, P.S. (éditeurs), 403-423, New York: John Wiley and Sons, 1995.

LEE, H., RANCOURT, E., SÄRNDAL, C.-E., « Experiment with Variance Estimation from Survey Data with Imputed Values », *Journal of Official Statistics*, 10, 231-243, 1994.

LEE, H., RANCOURT, E., SÄRNDAL, C.-E., « Jackknife Variance Estimation for Data with Imputed Values », *Recueil de la section de méthodologie d'enquête,* 111-115, Société Statistique du Canada, juillet 1995.

LEE, H., RANCOURT, E., SÄRNDAL, C.-E., « Variance estimation in the presence of imputed data for the Generalized Estimation System », *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 384-389, août 1995.

MONTAQUILA, J. M., JERNIGAN, R. W., «Variance Estimation in the Presence of Imputed data », *Proceedings of the Section on Survey Research Methods,* American Statistical Association, à paraître, août 1997.

PROVOST, M., *Estimation de la variance dans les sondages utilisant l'imputation hot-deck.* Mémoire de maîtrise, Université de Montréal, 1995.

RANCOURT, E., « Issues in the Combined Use of Statistics Canada's Generalized Edit and Imputation System and Generalized Estimation System », *Survey and Statistical Computing : Proceedings of The Second ASC International Conference,* Association for Survey Computing, 185-194, septembre 1996.

RANCOURT, E., LEE, H., SÄRNDAL, C.-E., « Variance Estimation under More Than one Imputation Method », *Proceedings of the International Conference on Establishment Surveys*, American Statistical Association, 374-379, juin 1993.

RANCOURT, E., LEE, H., SÄRNDAL, C.-E., « Bias Corrections for Survey Estimates from Data with Ratio Imputed Values for Confounded Nonresponse », *Survey Methodology*, 20, 137-147, 1994.

RAO, J.N.K., « Variance Estimation under Imputation for Missing Data ». Rapport technique, Statistique Canada, Ottawa, 1990.

RAO, J.N.K., « Jackknife Variance Estimation under Imputation for Missing Data. Rapport technique, Statistique Canada, Ottawa, 1991.

RAO, J.N.K., SHAO, J., « Jackknife variance estimation with survey data under hot-deck imputation ». *Biometrika*, 79, 811-822, 1992.

RAO, J.N.K., SITTER, R.R., « Variance estimation under two-phase sampling with application to imputation for missing data », *Biometrika*, 82, 453-460, 1995.

RAO, J.N.K., SITTER, R.R., « Efficient Random Imputation for Missing Data in Complex Surveys », Rapport technique, Carleton University et Simon Fraser University, 1997.

RUBIN, D.B., « Inference and missing data », *Biometrika*, 63, 581-590, 1976.

RUBIN, D.B., « Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse ». *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 20-34, 1978.

RUBIN, D.B., *Multiple imputation for nonresponse in surveys*, New York: John Wiley and Sons, 1987.

SÄRNDAL, C.-E., « Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation », *Recueil du Symposium '90: Mesure et amélioration de la qualité des données*, 337-347. Statistique Canada, Ottawa, 1990.

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

SÄRNDAL, C.-E., « For a Better Understanding of Imputation », *Proceedings of the 6th Workshop on Household Survey Nonresponse,* Helsinki, Octobre 1995.

SÄRNDAL, C.-E., SWENSSON, B., WRETMAN, J.H., *Model Assisted Survey Sampling*. New York: Springer-Verlag, 1992.

SHAO, J., CHEN, Y, CHEN, Y., « Balanced Repeated Replication for Stratified Multistage Survey Data under Imputation », *Journal of the American Statistical Association*, à paraître, 1998.

SHAO, J., SITTER, R.R., « Bootstrap for imputed survey data », *Journal of the American Statistical Association*, 91, 1278-1288, 1996.