Recueil du Symposium 2002 de Statistique Canada
Modélisation des données d'enquête pour la recherche économique et sociale

SYSTÈME POUR L'ESTIMATION DE LA VARIANCE DUE À LA NON-RÉPONSE ET À L'IMPUTATION (SEVANI)

Jean-François Beaumont et Charles Mitchell[1]

RÉSUMÉ

Le Système pour l'estimation de la variance due à la non-réponse et à l'imputation (SEVANI) est un système prototype programmé en SAS et doté d'une interface utilisateur graphique. L'estimation de la variance se fonde sur le cadre de travail « quasi-multi-phases ». Ce cadre nécessite l'utilisation d'un modèle de non-réponse et permet aussi celle d'un modèle d'imputation. Le SEVANI peut faire face à deux catégories de méthodes de traitement de la non-réponse, à savoir la repondération et l'imputation. Si l'on choisit l'imputation, le SEVANI suppose l'utilisation de l'une des quatre méthodes suivantes, à savoir l'imputation déterministe par la régression linéaire, l'imputation aléatoire par la régression linéaire, l'imputation par une valeur auxiliaire ou l'imputation par le plus proche voisin. Dans le présent article, nous décrivons la méthodologie sur laquelle s'appuie le SEVANI et nous présentons un exemple basé sur des données réelles pour illustrer l'application de la théorie.

MOTS CLÉS : Quasi-randomisation; échantillonnage à deux phases; modèle de non-réponse; modèle d'imputation.

1. INTRODUCTION

La plupart des enquêtes, sinon toutes, doivent faire face au problème de la non-réponse. Celle-ci peut avoir plusieurs causes, dont le refus par la personne interviewée de répondre à au moins une question ou l'impossibilité de prendre contact avec une unité donnée. La non-réponse peut aussi être créée à l'étape de la vérification des données en vue de résoudre des problèmes d'incohérence ou de réponses suspectes. Par conséquent, la non-réponse a ici la signification générale de données manquantes dans l'échantillon.

La non-réponse produit inévitablement un échantillon observé de plus petite taille que l'échantillon sélectionné au départ. Cette réduction de la taille de l'échantillon s'accompagne généralement d'une augmentation de la variance des estimations, quelle que soit la méthode choisie pour traiter la non-réponse. Cette augmentation de la variance est appelée la variance due à la non-réponse. Nous définissons ici la variance due à l'imputation comme étant une composante de la variance due à la non-réponse qui résulte de l'application d'une méthode d'imputation aléatoire.

Au début des années 80, la plupart des travaux de recherche sur la non-réponse se sont concentrés sur l'évaluation et la réduction du biais dû à la non-réponse. Depuis, de nombreuses études ont été consacrées à l'estimation de la variance due à la non-réponse, surtout lorsqu'on recourt à l'imputation pour remplacer les valeurs manquantes (voir Lee, Rancourt et Särndal, 2001 pour une revue). Les résultats de certains de ces travaux ont mené au développement du Système d'estimation de la variance due à l'imputation (SIMPVAR), qui a été présenté au Symposium 1997 de Statistique Canada (Rancourt, Gagnon, Lee, Provost et Särndal 1997). Les travaux plus récents ont visé à donner la possibilité de considérer un plus grand nombre de méthodes de traitement de la non-réponse dans un cadre unifié d'estimation de la variance appelé cadre de travail « quasi-multi-phases » (ou quasi-randomisation selon la terminologie de Oh et Scheuren, 1983). Särndal et Swensson (1987), Särndal (1992), Rao et Sitter (1995) et Beaumont (2000) sont des exemples d'articles fondés sur le cadre de travail « quasi-deux-phases ». Étant donné les modifications importantes qui lui ont été apportées, SIMPVAR a changé de nom pour devenir le Système pour l'estimation de la variance due à la non-réponse et à l'imputation (SEVANI). Le SEVANI est un système prototype développé à l'aide de SAS v8. Par conséquent, il

[1] Statistique Canada, Immeuble R.-H.-Coats, 16e étage, Pré Tunney, Canada, K1A 0T6
(jean-francois.beaumont@statcan.ca et charles.mitchell@statcan.ca)

n'accepte que les fichiers et les bibliothèques SAS v8. Il s'agit d'une série de macro-instructions qui peuvent être exécutées individuellement ou en se servant de l'interface utilisateur graphique.

Pour utiliser la version 1.0 du SEVANI, il faut que les paramètres de population que l'on doit estimer soient le total ou la moyenne d'un domaine et que l'estimateur en présence de réponse complète soit l'estimateur d'Horvitz-Thompson ou fasse partie de la famille des estimateurs par calage. Le SEVANI peut faire face à deux catégories de méthodes de traitement de la non-réponse, à savoir la repondération et l'imputation. Si l'on choisit l'imputation, le SEVANI suppose l'utilisation de l'une des quatre méthodes suivantes, à savoir l'imputation déterministe par la régression linéaire, l'imputation aléatoire par la régression linéaire, l'imputation par une valeur auxiliaire ou l'imputation par le plus proche voisin.

2. POURQUOI FAUT-IL ESTIMER LA VARIANCE DUE À LA NON-RÉPONSE?

Comme le font remarquer Rancourt et coll. (1997), plusieurs raisons justifient l'estimation de la variance d'un estimateur, qu'il y ait ou non non-réponse. En cas de non-réponse, on peut mettre l'accent sur les quatre raisons principales d'estimer la variance due à la non-réponse, soit :

i) obtenir des inférences valides en présence d'une non-réponse;
ii) mesurer convenablement la qualité des estimations d'enquête et informer les utilisateurs de la qualité des données;
iii) mieux répartir les ressources d'enquête;
iv) comparer diverses méthodes de traitement de la non-réponse et prendre de meilleures décisions.

La troisième raison signifie que, si la variance due à la non-réponse est importante comparativement à la variance d'échantillonnage dans une strate particulière, il pourrait être indiqué de consacrer davantage de ressources à la prévention de la non-réponse (p. ex., un plus grand nombre de suivis des non-répondants) pour cette strate particulière. Pour réaliser cet objectif en respectant le budget de l'enquête, il se pourrait qu'on doive réduire la taille souhaitée de l'échantillon. Une telle mesure risque d'augmenter la variance d'échantillonnage, mais on peut s'attendre à une réduction importante de la variance due à la non-réponse, donc, à une diminution éventuelle de la variance totale.

Un bon effort de modélisation est toujours requis en vue de réduire autant que possible le bais dû à la non-réponse et de choisir une méthode de traitement de la non-réponse. Si un modèle donne de meilleurs résultats que les autres, il n'est pas nécessaire d'estimer la variance due à la non-réponse afin de choisir la méthode. Cependant, si plusieurs modèles se font concurrence, l'estimation de la variance due à la non-réponse peut servir de critère pour décider de la méthode de traitement de la non-réponse que l'on adoptera, comme le souligne la quatrième raison susmentionnée.

3. LE CAS DE LA RÉPONSE COMPLÈTE

Supposons que l'on veuille estimer le total ou la moyenne d'un domaine pour une population donnée *U*, et représentons ce paramètre inconnu de la population par θ. Dans le cas d'un total de domaine, $\theta = \sum_{k \in U} d_k y_k$, où y_k est la valeur de la variable d'intérêt *y* pour l'unité de population *k* et d_k est une variable indicatrice égale à 1 si l'unité *k* appartient au domaine d'intérêt et à 0, autrement. Dans le cas d'une moyenne de domaine, $\theta = \sum_{k \in U} d_k y_k / \sum_{k \in U} d_k$.

Comme il est d'habitude impossible d'étudier toutes les unités de la population, on sélectionne un échantillon aléatoire *S* conformément à un plan d'échantillonnage donné. Représentons la probabilité de sélection de l'unité *k* par π_k et la probabilité qu'un échantillon donné *s* soit sélectionné par $p(s)$. Représentons par $\mathrm{E}_p(\cdot)$ les espérances par rapport au plan d'échantillonnage considéré. On obtient habituellement un estimateur $\hat{\theta}$ de θ en calculant les poids d'estimation $\tilde{w}_k$, pour $k \in S$. Dans le cas d'un total de domaine, $\hat{\theta} = \sum_{k \in S} \tilde{w}_k d_k y_k$, et, dans le cas d'une moyenne de domaine,

$\hat{\theta} = \sum_{k \in S} \tilde{w}_k d_k y_k / \sum_{k \in S} \tilde{w}_k d_k$. Si l'on ne dispose d'aucun renseignement auxiliaire au niveau de la population, alors on obtient l'estimateur d'Horvitz-Thompson (HT) habituel de θ en prenant $\tilde{w}_k = 1/\pi_k$.

Il n'est pas inhabituel qu'on dispose d'un vecteur de variables auxiliaires $\mathbf{x_1}$ pour toutes les unités échantillonnées et que l'on connaisse les totaux de population, $\mathbf{t}_{\mathbf{x}_1} = \sum_{k \in U} \mathbf{x_{1k}}$, pour ces variables. Dans ce cas, nous pouvons utiliser ces renseignements auxiliaires sous la forme d'un estimateur par calage pour améliorer l'estimateur HT habituel de θ. Par exemple, nous pouvons utiliser l'estimateur par la régression généralisée (GREG) ainsi que les poids d'estimation $\tilde{w}_k = w_k g_k$, où

$$g_k = 1 + \frac{\mathbf{x}'_{\mathbf{1k}}}{c_{1k}} \left(\sum_{k \in S} \frac{w_k}{c_{1k}} \mathbf{x_{1k}} \mathbf{x}'_{\mathbf{1k}} \right)^{-1} \left(\mathbf{t}_{\mathbf{x}_1} - \sum_{k \in S} w_k \mathbf{x_{1k}} \right),$$

$w_k = 1/\pi_k$ et c_{1k} est une fonction positive connue de $\mathbf{x_{1k}}$ qui correspond à la structure de variance du modèle d'estimation qui sous-tend l'estimateur GREG. L'estimateur GREG est asymptotiquement sans biais par rapport au plan *p*, autrement dit, $\mathrm{E}_p(\hat{\theta}) \approx \theta$.

La variance d'échantillonnage de $\hat{\theta}$ est représentée par $\mathrm{V}_{\mathrm{sam}} = \mathrm{V}_p(\hat{\theta})$. Nous supposons qu'il existe un estimateur de la variance pour $\mathrm{V}_{\mathrm{sam}}$, représenté par $\hat{\mathrm{V}}_{\mathrm{sam}} = \hat{\mathrm{V}}_p(\hat{\theta})$. Par exemple, nous pourrions obtenir $\hat{\mathrm{V}}_{\mathrm{sam}}$ pas des méthodes telles que la linéarisation de Taylor, le bootstrap ou le jackknife.

4. LE CADRE DE TRAVAIL « QUASI-DEUX-PHASES »

Dans le cadre de travail « quasi-deux-phases », la non-réponse est considérée comme une deuxième phase de la sélection et la variable *y* est observée uniquement pour une partie de l'échantillon *s*. Soit R_1 l'échantillon aléatoire de répondants (l'indice 1 indique qu'il s'agit de la première phase de non-réponse). La probabilité que l'unité échantillonnée *k* réponde est représentée par $p_{1k} = P(k \in R_1 \mid s)$. La probabilité que deux unités échantillonnées distinctes *k* et *l* répondent est représentée par $p_{1kl} = P(k \in R_1, l \in R_1 \mid s)$. La probabilité qu'un échantillon donné de répondants r_1 soit observé est représentée par $q_1(r_1 \mid s)$ et les espérances sous le mécanisme de non-réponse sont représentées par $\mathrm{E}_{q_1}(\cdot \mid s)$. Les espérances sous le plan d'échantillonnage et sous le mécanisme de non-réponse sont représentées par $\mathrm{E}_{pq_1}(\cdot) = \mathrm{E}_p \mathrm{E}_{q_1}(\cdot \mid s)$.

En présence de non-réponse, on ne peut calculer l'estimateur GREG ou HT représenté par $\hat{\theta}$, puisque y_k n'est pas observé pour $k \in s - r_1$. On recourt habituellement à la repondération ou à l'imputation afin d'obtenir un estimateur $\hat{\theta}^*$ corrigé pour la non-réponse. On suppose que l'estimateur corrigé est asymptomatiquement sans biais par rapport au mécanisme de la première phase de non-réponse q_1, étant donné l'échantillon réalisé *s*, autrement dit, que $\mathrm{E}_{q_1}(\hat{\theta}^* \mid s) \approx \hat{\theta}$. Par conséquent, $\hat{\theta}^*$ est approximativement inconditionnellement sans biais ($\mathrm{E}_{pq_1}(\hat{\theta}^*) = \mathrm{E}_p \mathrm{E}_{q_1}(\hat{\theta}^* \mid s) \approx \theta$) et sa variance peut être approximée par

$$\begin{aligned} \mathrm{V}_{pq_1}(\hat{\theta}^*) &= \mathrm{V}_p \mathrm{E}_{q_1}(\hat{\theta}^* \mid s) + \mathrm{E}_p \mathrm{V}_{q_1}(\hat{\theta}^* \mid s) \\ &\approx \mathrm{V}_p(\hat{\theta}) + \mathrm{E}_p \mathrm{V}_{q_1}(\hat{\theta}^* \mid s) \\ &= \mathrm{V}_{\mathrm{sam}} + \mathrm{V}_{\mathrm{nr1}}, \end{aligned} \tag{4.1}$$

où $\mathrm{V}_{\mathrm{nr1}} = \mathrm{E}_p \mathrm{V}_{q_1}(\hat{\theta}^* \mid s)$ est la variance due à la première phase de non-réponse. Donc, nous pouvons obtenir un estimateur approximativement sans biais de la variance pour $\mathrm{V}_{\mathrm{nr1}}$ en trouvant simplement un estimateur approximativement sans biais pour $\mathrm{V}_{q_1}(\hat{\theta}^* \mid s)$. Puisqu'en pratique, on ne connaît pas le mécanisme de non-réponse, il faut spécifier un modèle de non-réponse pour l'approximer. Par conséquent, l'expression (4.1) de la variance n'est valide que si le modèle de non-réponse postulé est un bon substitut du mécanisme réel, mais inconnu, de non-réponse. C'est la raison pour laquelle le cadre de travail est qualifié de quasi-deux-phases.

Lorsqu'on recourt à l'imputation pour traiter la non-réponse, on consacre habituellement plus d'efforts à la recherche d'un bon modèle d'imputation (modèle pour la variable *y*, représenté par *m*) qu'à celle d'un bon modèle de non-réponse. Dans ces conditions, il serait sans doute préférable de ne pas s'appuyer trop fortement sur le modèle de non-réponse et d'utiliser le modèle d'imputation pour obtenir un estimateur de la variance. Une approche de ce genre est décrite dans Särndal (1992). Au lieu d'estimer $\mathrm{V}_{pq_1}(\hat{\theta}^*)$, il propose d'estimer l'espérance par rapport au modèle de $\mathrm{V}_{pq_1}(\hat{\theta}^*)$, qui est donnée par

$$\begin{aligned} \mathrm{E}_m \mathrm{V}_{pq_1}(\hat{\theta}^*) &\approx \mathrm{E}_m \mathrm{V}_p(\hat{\theta}) + \mathrm{E}_m \mathrm{E}_p \mathrm{V}_{q_1}(\hat{\theta}^* \mid s) \\ &= \mathrm{V}^{\mathrm{m}}_{\mathrm{sam}} + \mathrm{V}^{\mathrm{m}}_{\mathrm{nr1}}, \end{aligned} \tag{4.2}$$

où, ici, la variance d'échantillonnage est $\mathrm{V}^{\mathrm{m}}_{\mathrm{sam}} = \mathrm{E}_m \mathrm{V}_{\mathrm{sam}}$ et la variance due à la première phase de non-réponse est $\mathrm{V}^{\mathrm{m}}_{\mathrm{nr1}} = \mathrm{E}_m \mathrm{V}_{\mathrm{nr1}}$. Sous cette approche, on suppose généralement que le plan d'échantillonnage et le mécanisme de non-réponse sont ignorables par rapport au modèle d'imputation *m*, tel que défini dans Rubin (1976). Pratiquement, cela signifie que l'information pertinente sur le plan de sondage et l'information liée au mécanisme de non-réponse sont incluses dans le modèle d'imputation. Cette hypothèse rend l'estimation de la variance beaucoup plus facile. En particulier, à condition que $\mathrm{E}_{q_1}(\hat{\theta}^* \mid s) \approx \hat{\theta}$, la variance due à la première phase de non-réponse peut s'écrire sous la forme

$$\mathrm{V}^{\mathrm{m}}_{\mathrm{nr1}} = \mathrm{E}_p \mathrm{E}_{q_1} \mathrm{E}_m \left\{ (\hat{\theta}^* - \hat{\theta})^2 \right\}.$$

Par conséquent, nous pouvons obtenir un estimateur approximativement sans biais de la variance pour $\mathrm{V}^{\mathrm{m}}_{\mathrm{nr1}}$ en recherchant simplement un estimateur approximativement sans biais pour $\mathrm{E}_m \left\{ (\hat{\theta}^* - \hat{\theta})^2 \right\}$ qui ne nécessite pas de modélisation du mécanisme inconnu de non-réponse, à part le fait de supposer qu'il est ignorable. Ce modèle de non-réponse (hypothèse selon laquelle le mécanisme de non-réponse est ignorable) est beaucoup plus faible que la spécification du modèle de non-réponse habituellement nécessaire pour estimer (4.1). Cependant, cette approche exige que le modèle d'imputation *m* soit valide. Le choix entre l'estimation de (4.1) ou (4.2) devrait donc dépendre de la confiance qu'a le méthodologiste d'enquête dans le modèle de non-réponse ou dans le modèle d'imputation. S'il considère que le modèle de non-réponse est de meilleure qualité que le modèle d'imputation, alors l'estimation de (4.1) devrait être préférée à celle de (4.2) et vice-versa.

5. MODÈLES UTILISÉS DANS LE SEVANI

Comme on ne connaît pas le mécanisme de non-réponse, on ne connaît pas non plus les probabilités de réponse p_{1k}, pour $k \in s$ ni les probabilités de réponse conjointes p_{1kl}, pour $k, l \in s$. Afin d'estimer la variance selon l'expression (4.1), on estime généralement ces probabilités au moyen d'un modèle de non-réponse. Il s'agit de l'approche fondée sur un modèle de non-réponse. Un de ces modèles utilisé couramment est le modèle de non-réponse uniforme dans la classe, où on suppose que toutes les unités appartenant à une classe donnée répondent indépendamment les unes des autres avec la même probabilité de réponse.

En pratique, on n'estime jamais directement les probabilités conjointes de réponse p_{1kl}. Par conséquent, il faut émettre l'hypothèse d'une certaine forme d'indépendance pour se débarrasser de l'estimation de la probabilité de réponse conjointe et simplifier l'estimation de la variance (4.1). Dans le SEVANI, il est supposé que les grappes d'unités *c*, pour $c \in s'$, répondent indépendamment les unes des autres, où s' est l'échantillon sélectionné de grappes. Autrement dit, toutes les unités échantillonnées à l'intérieur d'une grappe *c* sont observées simultanément ou elles manquent toutes simultanément. La probabilité de réponse pour toutes les unités $k \in s_c$ est représentée par p'_{1c}, où s_c est l'échantillon d'unités dans la grappe *c*. La probabilité de réponse conjointe d'une unité *k* dans la grappe *c* et d'une unité *l* dans une grappe différente c' est égale à $p'_{1c}\, p'_{1c'}$, tandis que la probabilité de réponse conjointe de toute paire d'unités dans la même grappe *c* est égale à p'_{1c}. Un exemple typique où l'on peut supposer que les grappes d'unités répondent indépendamment est celui où les logements (grappes) sont sélectionnés, mais que l'information souhaitée est recueillie pour toutes les personnes vivant dans les logements sélectionnés. Notons qu'une grappe pourrait aussi contenir une seule unité échantillonnée et, donc, correspondre exactement à une unité d'analyse *k*.

À condition que le modèle de non-réponse choisi soit satisfaisant, les probabilités de réponse estimées devraient être proches des probabilités de réponse réelles inconnues. Dans le SEVANI, on suppose que les probabilités de réponse estimées sont exactement égales aux probabilités de réponse réelles p_{1k} (ou p'_{1c}). Cette hypothèse devrait produire une sous-estimation de la variance due à la non-réponse $\mathrm{V}_{\mathrm{nr1}}$ quand $\hat{\theta}^*$ dépend de ces probabilités de réponse, comme cela est le cas lorsqu'on fait une repondération ou, parfois, lorsqu'on fait une imputation. Cependant, selon certaines études par simulation (par exemple, Mantel, Nadon et Yeo, 2000), la sous-estimation est souvent négligeable.

Pour trouver un estimateur de la variance (4.2), le SEVANI suppose que le plan d'échantillonnage et le mécanisme de non-réponse sont ignorables par rapport au modèle d'imputation. En outre, le système utilise le modèle d'imputation *m* où $\mathrm{E}_m(y_k \mid \mathbf{x_k}) = \boldsymbol{\beta}'\mathbf{x_k}$, $\mathrm{V}_m(y_k \mid \mathbf{x_k})$ est proportionnelle à c_k et où les observations sont indépendantes les unes des autres. Le vecteur $\mathbf{x_k}$ contient les variables auxiliaires, disponibles pour tout $k \in s$, utilisées pour obtenir l'estimateur imputé $\hat{\theta}^*$, $\boldsymbol{\beta}$ est un vecteur de paramètres inconnus du modèle et c_k est une fonction positive connue de $\mathbf{x_k}$ qui correspond à la structure de variance du modèle d'imputation *m*. Il s'agit de l'approche fondée sur un modèle d'imputation. Parfois, on exécute l'imputation indépendamment à l'intérieur des classes d'imputation. Le SEVANI traite cette situation en ajustant un modèle distinct pour chaque classe. Il est également possible de traiter les classes d'imputation en incluant dans le vecteur $\mathbf{x_k}$ des variables indicatrices qui précisent à quelle classe appartient l'unité *k*.

Un bon effort de modélisation est toujours nécessaire pour réduire autant que possible le biais dû à la non-réponse et la variance due à la non-réponse, quelle que soit la méthode choisie de traitement de la non-réponse. Durant la modélisation, la sélection convenable des variables auxiliaires, qui sont utilisées pour obtenir l'estimateur corrigé $\hat{\theta}^*$, est l'élément clé de la réduction de l'effet de la non-réponse.

6. MÉTHODES DE TRAITEMENT DE LA NON-RÉPONSE

6.1 Repondération

Un moyen courant de remédier à la non-réponse totale consiste à faire une repondération. Si l'on dispose des variables auxiliaires $\mathbf{x_1}$, on obtient l'estimateur corrigé résultant $\hat{\theta}^*$ de θ en considérant la non-réponse totale comme une deuxième phase de sélection et en utilisant l'estimateur GREG sous les conditions de l'échantillonnage à deux phases. Dans le cas du total d'un domaine, $\hat{\theta}^* = \sum_{k \in R_1} \tilde{w}_k^* d_k y_k$, et dans celui de la moyenne d'un domaine, $\hat{\theta}^* = \sum_{k \in R_1} \tilde{w}_k^* d_k y_k \Big/ \sum_{k \in R_1} \tilde{w}_k^* d_k$, où $\tilde{w}_k^* = w_k^* g_k^*$ est le poids d'estimation de l'unité *k*,

$$g_k^* = 1 + \frac{\mathbf{x'_{1k}}}{c_{1k}} \left(\sum_{k \in R_1} \frac{w_k^*}{c_{1k}} \mathbf{x_{1k}} \mathbf{x'_{1k}} \right)^{-1} \left(\mathbf{t_{x_1}} - \sum_{k \in R_1} w_k^* \mathbf{x_{1k}} \right) \tag{6.1}$$

et $w_k^* = w_k / p_{1k}$ est le poids ajusté pour la non-réponse de l'unité *k*. Si l'on ne dispose d'aucune variable auxiliaire, l'estimateur corrigé $\hat{\theta}^*$ a la même forme, avec remplacement de g_k^* dans (6.1) par $g_k^* = 1$.

6.2 Imputation

On utilise généralement l'imputation pour compenser la non-réponse partielle et, parfois, pour tenir compte de la non-réponse totale. L'imputation consiste à remplacer la valeur manquante y_k d'une unité non répondante $k \in S - R_1$ par une valeur imputée y_k^*. L'estimateur corrigé $\hat{\theta}^*$ de θ est $\hat{\theta}^* = \sum_{k \in S} \tilde{w}_k d_k y_{\bullet k}$ dans le cas d'un total de domaine et $\hat{\theta}^* = \sum_{k \in S} \tilde{w}_k d_k y_{\bullet k} / \sum_{k \in S} \tilde{w}_k d_k$ dans le cas d'une moyenne de domaine, où $y_{\bullet k} = R_{1k} y_k + (1 - R_{1k}) y_k^*$ et R_{1k} est une variable aléatoire indiquant si l'unité *k* a répondu ($R_{1k} = 1$) ou non ($R_{1k} = 0$). Chaque méthode d'imputation mène à une façon différente d'obtenir les valeurs imputées y_k^* et à un estimateur corrigé $\hat{\theta}^*$ différent.

Dans le cas de l'imputation déterministe par la régression linéaire (DRL), la valeur imputée y_k^* est égale à la valeur prédite $\hat{\boldsymbol{\beta}}' \mathbf{x_k}$ obtenue d'après le modèle d'imputation *m*, où

$$\hat{\boldsymbol{\beta}} = \left(\sum_{k \in R_1} \frac{a_k}{c_k} \mathbf{x_k} \mathbf{x'_k} \right)^{-1} \sum_{k \in R_1} \frac{a_k}{c_k} \mathbf{x_k} y_k$$

est l'estimateur de $\boldsymbol{\beta}$ et a_k est un poids de régression. L'imputation DRL comporte de nombreux cas particuliers, comme l'imputation par la moyenne ou l'imputation par le ratio. S'il existe des classes d'imputation, on estime $\hat{\boldsymbol{\beta}}$ séparément pour chaque classe.

Pour chaque méthode d'imputation DRL, il existe une méthode d'imputation aléatoire par la régression linéaire (ARL) correspondante. Dans le cas de l'imputation ARL, la valeur imputée y_k^* est égale à $\hat{\boldsymbol{\beta}}' \mathbf{x_k} + e_k$, où e_k est une composante aléatoire. Autrement dit, $e_k = (y_l - \hat{\boldsymbol{\beta}}' \mathbf{x_l}) \sqrt{c_k / c_l}$ et l'unité *l* est un répondant ($l \in r_1$) appartenant à la même classe d'imputation que l'unité *k*, sélectionné aléatoirement avec remise et avec probabilité proportionnelle à a_k. L'imputation aléatoire hot-deck, qui est une version aléatoire de l'imputation par la moyenne, est un exemple d'imputation ARL. Dans le cas de l'imputation aléatoire hot-deck, on remplace une valeur manquante par la valeur d'un répondant sélectionné aléatoirement et qui se trouve dans la même classe d'imputation que celle du non-répondant.

Dans le cas de l'imputation par une valeur auxiliaire (VA), la valeur manquante y_k pour une unité non répondante $k \in s - r_1$ est remplacée par une valeur auxiliaire z_k. La valeur imputée $y_k^* = z_k$ est obtenue en utilisant uniquement des données provenant de l'unité non répondante *k*, comme des données historiques. L'imputation par la valeur précédente est un exemple d'imputation VA, où la valeur auxiliaire z_k est la valeur de la variable *y* observée à une période antérieure de l'enquête. On peut justifier cette méthode d'imputation au moyen du modèle d'imputation *m* où $\boldsymbol{\beta}' \mathbf{x_k}$ est connu. Par conséquent, il n'y a aucun vecteur de paramètres inconnus à estimer, contrairement aux méthodes d'imputation par régression.

Enfin, dans le cas de l'imputation par le plus proche voisin (PPV), la valeur imputée y_k^* est égale à y_l, où l'unité *l* est le répondant le plus proche de l'unité *k* en ce qui a trait aux variables auxiliaires $\mathbf{x}$ et *l* fait partie de la même classe

d'imputation que *k*. Pour déterminer le répondant le plus proche de l'unité *k*, une mesure de distance est nécessaire. Dans le SEVANI, il est supposé qu'on a utilisé la norme L_p. Il est également supposé que les variables auxiliaires sont numériques et qu'elles ont subi un changement d'échelle. Le changement d'échelle se fait pour chaque variable auxiliaire en soustrayant la moyenne et en divisant par l'écart-type ou en remplaçant les valeurs de chaque variable auxiliaire par leur rang.

7. ESTIMATION DE LA VARIANCE

7.1 Estimation de la variance d'échantillonnage

En présence de non-réponse, on ne peut calculer les estimateurs habituels de la variance d'échantillonnage $\hat{\mathrm{V}}_{\mathrm{sam}}$ de $\mathrm{V}_{\mathrm{sam}} = \mathrm{V}_p(\hat{\theta})$, puisque y_k n'est pas observé pour $k \in s - r_1$. Cependant, on peut estimer $\mathrm{V}_p(\hat{\theta} \mid r_1)$ par $\hat{\mathrm{V}}_p(\hat{\theta} \mid r_1)$ en appliquant les techniques usuelles d'estimation de la variance. On peut pour cela considérer que R_{1k}, pour $k \in U$, est fixe et utiliser les systèmes d'estimation de la variance existants qui sont valides en cas de réponse complète. On peut alors obtenir un estimateur $\hat{\mathrm{V}}^*_{\mathrm{sam}} = \hat{\mathrm{V}}_p(\hat{\theta} \mid r_1) + \hat{\mathrm{V}}_{\mathrm{cor}}$ approximativement sans biais de la variance $\mathrm{V}_{\mathrm{sam}}$, où $\hat{\mathrm{V}}_{\mathrm{cor}}$ est construit de telle façon que $\mathrm{E}_q(\hat{\mathrm{V}}_{\mathrm{cor}} \mid s) \approx \hat{\mathrm{V}}_{\mathrm{sam}} - \mathrm{E}_q\left\{\hat{\mathrm{V}}_p(\hat{\theta} \mid r_1) \mid s\right\}$. Dans sa version 1.0, le SEVANI n'estime pas la variance d'échantillonnage. Cependant, dans le cas de la repondération pour la non-réponse, le SEVANI calcule $\hat{\mathrm{V}}_{\mathrm{cor}}$, s'il l'est demandé, par la technique de linéarisation de Taylor.

Quand on recourt à l'imputation, il pourrait être plus approprié d'estimer $\mathrm{V}^{\mathrm{m}}_{\mathrm{sam}} = \mathrm{E}_m \mathrm{V}_{\mathrm{sam}}$ que $\mathrm{V}_{\mathrm{sam}}$, comme il l'est mentionné à la section 4. En fait, ce qui est vraiment souhaitable est de trouver un prédicteur pour $\hat{\mathrm{V}}_{\mathrm{sam}}$ approximativement sans biais par rapport au modèle, c'est-à-dire un prédicteur $\hat{\mathrm{V}}^*_{\mathrm{sam}}$ tel que $\mathrm{E}_m\left\{\left(\hat{\mathrm{V}}^*_{\mathrm{sam}} - \hat{\mathrm{V}}_{\mathrm{sam}}\right) \mid s, r_1\right\} \approx 0$. Il est facile de montrer qu'un tel prédicteur $\hat{\mathrm{V}}^*_{\mathrm{sam}}$ est sans biais par rapport à *m, p* et *q*.

Un moyen d'obtenir un prédicteur pour $\hat{\mathrm{V}}_{\mathrm{sam}}$ approximativement sans biais par rapport au modèle consiste à imputer les valeurs manquantes en tirant des valeurs à partir de la distribution estimée de $\{y_k : k \in s - r_1\}$, étant donné $\{y_k : k \in r_1\}$, avant d'utiliser un estimateur usuel de la variance valide dans le cas de la réponse complète. Essentiellement, cela signifie qu'on ajoute une composante aléatoire aux valeurs imputées dans le cas de l'imputation DRL ou VA avant d'utiliser un système courant d'estimation de la variance. Dans le cas de l'imputation ARL ou PPV, il n'est pas nécessaire de modifier les valeurs imputées. Notons que ces valeurs imputées modifiées ne sont utilisées que pour l'estimation de la variance et qu'elles n'ont aucun effet sur l'estimateur corrigé $\hat{\theta}^*$. Notons aussi que cette méthode de prédiction de $\hat{\mathrm{V}}_{\mathrm{sam}}$ n'a pas été implémentée dans la version courante du SEVANI.

Pour obtenir un estimateur plus efficace, on pourrait tirer plus d'un ensemble de valeurs imputées. On obtiendrait l'estimation finale de la variance d'échantillonnage en calculant la moyenne des estimations de la variance d'échantillonnage associées à chaque ensemble de valeurs imputées. Cette méthode ressemble à l'imputation multiple, sauf qu'ici, il n'est pas nécessaire que l'imputation soit propre. En pratique, il peut être plus commode de n'avoir qu'un seul ensemble de valeurs imputées. Cela devrait produire des estimations raisonnables de la variance d'échantillonnage, à moins que le taux de non-réponse soit très élevé.

7.2 Estimation de la variance due à la non-réponse

Comme nous l'avons mentionné à la section 4, si l'on considère que le modèle de non-réponse est de meilleure qualité que le modèle d'imputation, l'utilisation de l'approche fondée sur un modèle de non-réponse (estimation 4.1) est plus naturelle que celle de l'approche fondée sur un modèle d'imputation (estimation 4.2). Il en est ainsi quand on fait une repondération, puisqu'aucun effort n'est consacré à la recherche d'un modèle d'imputation. On peut obtenir un estimateur approximativement sans biais de la variance $\mathrm{V}_{\mathrm{nr1}}$ en recherchant simplement un estimateur approximativement sans biais

pour $V_{q_1}(\hat{\theta}^* \mid s)$. Dans le SEVANI, $V_{q_1}(\hat{\theta}^* \mid s)$ est estimé par la technique de linéarisation de Taylor en émettant l'hypothèse que les grappes d'unités répondent indépendamment les unes des autres.

Lorsqu'on recourt à l'imputation, il pourrait être plus approprié d'estimer V_{nr1}^{m} que V_{nr1}. Dans le SEVANI, pour estimer V_{nr1}^{m}, on estime $E_m\left\{(\hat{\theta}^* - \hat{\theta})^2\right\}$ et l'on suppose que le plan d'échantillonnage et le mécanisme de non-réponse sont ignorables par rapport au modèle d'imputation *m*. Puisque $\hat{\theta}^*$ et $\hat{\theta}$ sont des estimateurs linéaires, on obtient facilement une expression explicite pour $E_m\left\{(\hat{\theta}^* - \hat{\theta})^2\right\}$.

Que l'on choisisse d'estimer V_{nr1} ou V_{nr1}^{m}, il faut calculer une composante supplémentaire de la variance si l'on utilise la méthode d'imputation ARL. Cette composante supplémentaire de la variance tient compte de la variabilité due au processus d'imputation aléatoire. Donc, dans le cas de l'imputation ARL, la variance due à la non-réponse est égale à la variance due à la non-réponse de la méthode d'imputation DRL correspondante à laquelle s'ajoute la variance supplémentaire due à l'imputation aléatoire. Il est facile d'obtenir une expression explicite de la variance due à l'imputation aléatoire et indépendante de l'approche adoptée (approche fondée sur un modèle de non-réponse ou approche fondée sur un modèle d'imputation).

L'imputation PPV est une méthode d'imputation non paramétrique, puisqu'il n'est pas nécessaire de spécifier le modèle d'imputation pour justifier la forme de l'estimateur corrigé $\hat{\theta}^*$ et pour déterminer sa propriété d'être sans biais par rapport au modèle. Donc, l'utilisation d'un modèle d'imputation par la régression linéaire pour obtenir un estimateur de la variance pourrait ne pas toujours convenir, particulièrement si le modèle linéaire ne tient pas de façon satisfaisante. Si l'on choisit l'approche fondée sur un modèle de non-réponse, l'estimateur corrigé $\hat{\theta}^*$ n'est pas lisse, ce qui cause des complications lors de l'utilisation de la technique de linéarisation de Taylor.

Pour surmonter les problèmes que pose l'imputation PPV, on peut la considérer comme étant une imputation aléatoire hot-deck où le nombre de classes d'imputation est égal au nombre de valeurs distinctes de la variable *y*, ce qui correspond à un modèle d'imputation sans aucun degré de liberté. Par conséquent, il est impossible d'estimer la variabilité à l'intérieur d'une classe d'imputation puisque, par définition de l'imputation PPV, il n'y a aucune variabilité. Le SEVANI résout ce problème en utilisant, comme approximation de l'estimateur corrigé par imputation PPV, l'estimateur corrigé par imputation aléatoire hot-deck à l'intérieur des classes d'imputation, où les classes sont petites, mais contiennent au moins deux répondants. Pour former les classes, le SEVANI utilise l'algorithme de classification à « *k* moyennes » implémenté dans la procédure FASTCLUST de SAS. Le SEVANI utilise la même méthode de changement d'échelle et la même mesure de distance que celles utilisées pour effectuer l'imputation PPV.

8. LE CADRE DE TRAVAIL « QUASI-MULTI-PHASES »

Il est facile d'étendre le cadre de travail « quasi-deux-phases » au cadre de travail « quasi-multi-phases ». Dans ce dernier, il est possible d'estimer la variance due à la non-réponse associée à plus d'un mécanisme de non-réponse, c'est-à-dire plus d'une cause de non-réponse. Par exemple, la plupart des enquêtes souffrent de non-réponse totale et de non-réponse partielle, lesquelles sont vraisemblablement le résultat de mécanismes de non-réponse différents. En outre, on ne leur réserve généralement pas le même traitement. D'habitude, on recourt à la repondération pour compenser la non-réponse totale, mais à l'imputation pour compenser la non-réponse partielle.

La version 1.0 du SEVANI permet d'estimer jusqu'à trois composantes de la variance due à la non-réponse; chacune de ces composantes étant associée à un mécanisme différent de non-réponse. Chaque mécanisme de non-réponse supplémentaire est considéré comme une phase supplémentaire de sélection (ou phase de non-réponse). Par conséquent, lorsqu'il existe plus d'un mécanisme de non-réponse, l'approche est qualifiée de « quasi-multi-phases » au lieu de « quasi-deux-phases ». Bien que l'on puisse utiliser diverses méthodes à des phases différentes de non-réponse, le SEVANI limite la repondération à la première phase de non-réponse uniquement.

Lorsqu'il existe deux phases de non-réponse, la variable *y* est observée uniquement pour une partie de r_1. Soit r_2 l'échantillon de répondants à la deuxième phase. La probabilité qu'un échantillon donné de répondants r_2 soit observé est représentée par $q_2(r_2 \mid s, r_1)$ et les espérances sous le mécanisme de la deuxième phase de non-réponse sont représentées par $\mathrm{E}_{q_2}(\cdot \mid s, r_1)$. Les espérances sous le plan d'échantillonnage et les deux mécanismes de non-réponse sont représentées par $\mathrm{E}_{pq_1q_2}(\cdot) = \mathrm{E}_{pq_1}\mathrm{E}_{q_2}(\cdot \mid s, r_1)$.

Lorsqu'il existe deux phases de non-réponse, on ne peut calculer l'estimateur corrigé $\hat{\theta}^*$, puisque y_k n'est pas observé pour $k \in r_1 - r_2$. On procède habituellement à la repondération ou à l'imputation pour la non-réponse afin d'obtenir un estimateur doublement corrigé $\hat{\theta}^{**}$. On suppose que $\hat{\theta}^{**}$ est approximativement sans biais par rapport au mécanisme de la deuxième phase de non-réponse, étant donné les échantillons observés *s* et r_1, c'est-à-dire que $\mathrm{E}_{q_2}(\hat{\theta}^{**} \mid s, r_1) \approx \hat{\theta}^*$. Par conséquent, $\hat{\theta}^{**}$ est approximativement inconditionnellement sans biais et sa variance peut être approximée par

$$\mathrm{V}_{pq_1q_2}(\hat{\theta}^{**}) \approx \mathrm{V}_{pq_1}(\hat{\theta}^*) + \mathrm{E}_{pq_1}\mathrm{V}_{q_2}(\hat{\theta}^{**} \mid s, r_1),$$

où $\mathrm{V}_{\mathrm{nr2}} = \mathrm{E}_{pq_1}\mathrm{V}_{q_2}(\hat{\theta}^* \mid s, r_1)$ est la variance due à la deuxième phase de non-réponse. Si l'on adopte une approche fondée sur un modèle d'imputation, alors on estime $\mathrm{E}_m\mathrm{V}_{pq_1q_2}(\hat{\theta}^{**})$. L'estimation de la variance due à la deuxième phase de non-réponse est fort semblable à l'estimation de la variance due à la première phase de non-réponse et l'on peut utiliser les mêmes techniques. Nous avons déjà discuté de l'estimation de $\mathrm{V}_{pq_1}(\hat{\theta}^*)$ aux sections qui précèdent. Notons que le SEVANI ajoute une composante aléatoire aux valeurs imputées si l'on utilise l'imputation VA ou DRL à la deuxième phase. Cette mesure vise à assurer que la variance due à la première phase de non-réponse soit estimée correctement. Elle est également nécessaire pour l'estimation de la variance d'échantillonnage. L'extension à une troisième phase de non-réponse est simple.

9. UN EXEMPLE

Le présent exemple montre, à l'aide de données de l'Enquête sur la population active (EPA), comment on peut utiliser le SEVANI pour comparer des stratégies différentes de traitement de la non-réponse. Dans le cas de l'EPA, nous aimerions déterminer si l'imputation par la valeur précédente pourrait être remplacée par une méthode aléatoire hot-deck longitudinale. Une façon de déterminer laquelle de ces méthodes donne les meilleurs résultats consiste à estimer la variance due à la non-réponse par chacune d'elles. Dans le présent exemple, le paramètre d'intérêt de la population est le nombre total de personnes occupées dans la population.

L'EPA est une enquête mensuelle où les 5/6 de l'échantillon sont communs d'un mois à l'autre. Donc, s'il y a non-réponse et qu'il existe des données historiques, la pratique courante consiste à reporter ces données au mois courant. Cette pratique porte le nom d'imputation par la valeur précédente.

Comme nous l'avons mentionné à la section 5, l'imputation aléatoire hot-deck consiste à remplacer une valeur manquante par la valeur d'un répondant sélectionné aléatoirement dans la même classe d'imputation que celle du non-répondant. Dans l'imputation aléatoire hot-deck longitudinale, les classes d'imputation sont construites en utilisant des données auxiliaires historiques. L'imputation aléatoire hot-deck longitudinale est une version aléatoire de l'imputation longitudinale par la moyenne, où une valeur manquante dans une classe particulière est remplacée par la moyenne des répondants à l'intérieur de cette classe.

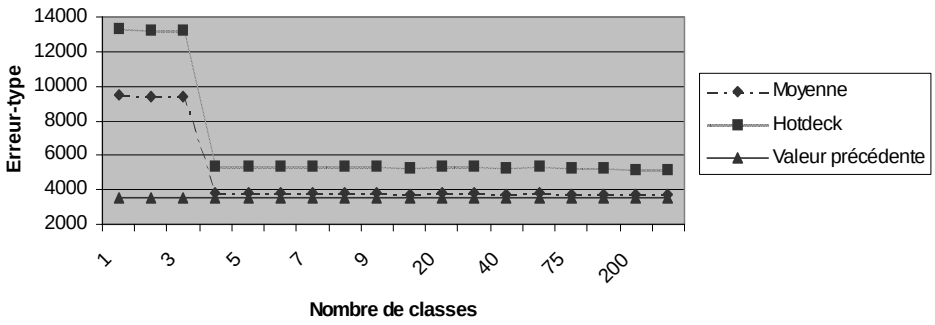
La construction des classes d'imputation se fonde sur la méthodologie des scores (voir, par exemple, Haziza, Charbonnier, Chow et Beaumont, 2001). Pour commencer, on utilise deux modèles de régression logistique, l'un pour la probabilité qu'un individu soit occupé et l'autre pour la probabilité qu'un individu soit chômeur. Puis, on utilise la régression stepwise pour essayer de trouver les meilleures variables auxiliaires historiques pour chaque modèle. Nous

avons utilisé les données de l'EPA portant sur une année complète pour choisir les modèles finaux. Nous avons trouvé que la situation d'activité le mois précédent était la variable la plus importante incluse dans les deux modèles.

Après avoir obtenu les modèles, nous avons utilisé des données historiques pour obtenir les probabilités estimées d'être occupé et d'être chômeur pour toutes les unités qui avaient répondu le mois précédent. Puis, nous avons utilisé un algorithme de classification (PROC FASTCLUS dans SAS) pour former des classes homogènes par rapport aux deux probabilités. Ensuite, nous avons procédé à l'imputation hot-deck et à l'imputation par la moyenne dans chaque classe.

Comme le montre la figure 1.0, la variance due à la non-réponse est systématiquement la plus faible lorsqu'on utilise l'imputation par la valeur précédente. Cependant, si l'on augmente le nombre de classes, l'imputation hot-deck longitudinale et l'imputation par la moyenne longitudinale donnent presque les mêmes résultats que l'imputation par la valeur précédente.

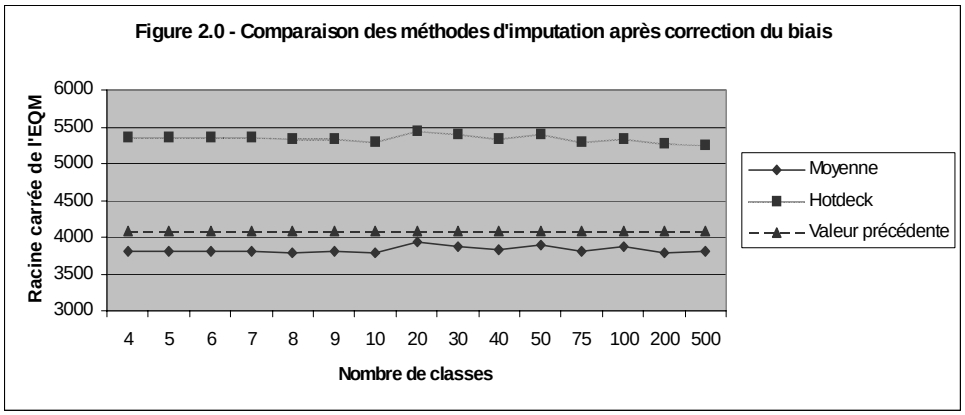
Figure 1.0 - Comparaison des méthodes d'imputation



Notons que le SEVANI n'estime que la variance due à la non-réponse et non le biais dû à la non-réponse. Puisqu'il est probable que l'imputation par la valeur précédente produise un biais dû à la non-réponse plus important que les méthodes longitudinales par la moyenne ou hot-deck, nous avons essayé d'estimer l'erreur quadratique moyenne (EQM) au lieu de la variance. Pour essayer d'estimer le biais, nous avons effectué une imputation par la moyenne à l'intérieur des classes en utilisant un très grand nombre de classes (1 500). Nous avons calculé l'estimation ponctuelle et nous l'avons utilisée comme valeur de référence pour l'estimation du biais et de l'EQM.

Comme le montre la figure 2.0, si l'on s'en tient à l'EQM, l'imputation par la valeur précédente n'est plus la meilleure méthode lorsque le nombre de classes est supérieur à 4. Dans ces conditions, des trois méthodes, l'imputation longitudinale par la moyenne est celle qui donne les meilleurs résultats. L'imputation hot-deck longitudinale produit encore une EQM plus grande que l'imputation par la valeur précédente, mais l'écart n'est pas important. Le fait que le biais associé à l'imputation par la valeur précédente soit plus important pourrait faire de la méthode hot-deck longitudinale une solution plus attrayante.

Comme le montre notre exemple, on peut utiliser le SEVANI comme élément du processus d'évaluation des stratégies de traitement de la non-réponse. Ici, le SEVANI montre que l'imputation par la valeur précédente donne de meilleurs résultats en ce qui concerne la variance due à la non-réponse. Cependant, il faut aussi tenir compte d'autres facteurs, comme le biais dû à la non-réponse.



RÉFÉRENCES

Beaumont, J.-F. (2000), "On Regression Imputation in the Presence of Nonignorable Nonresponse", *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 580-585.

Haziza, D., Charbonnier, C., Chow, O.S.Y., et Beaumont, J.-F. (2001), "Construction of Imputation Cells for the Canadian Labour Force Survey", *Proceedings of the Statistics Canada Symposium* 2001, *Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, Statistics Canada.

Lee, H., Rancourt, E., et Särndal, C.-E. (2001), "Variance Estimation from Survey Data under Single Imputation", in Groves, R.M., Dillman, D.A., Eltinge, J.L., and Little, R.J.A. (eds), *Survey Nonresponse*, New-York: John Wiley & Sons, Inc., pp. 315-328.

Mantel, H.J., Nadon, S., et Yeo, D. (2000), "Effect of Nonresponse Adjustments on Variance Estimates for the National Population Health Survey", *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 221-226.

Oh, H.L., et Scheuren, F.J. (1983), "Weighting Adjustment for Unit Nonresponse", in W.G. Madow, I. Olkin, and D.B. Rubin (eds.), *Incomplete Data in Sample Surveys*, Vol. 2, New-York: Academic Press, pp. 143-184.

Rancourt, E., Gagnon, F., Lee, H., Provost, M., et Särndal, C.-E. (1997), "Estimation of Variance in Presence of Imputation", *Proceedings of the Statistics Canada Symposium* 1997, *New Directions in Surveys and Censuses*, Statistics Canada, pp. 273-277.

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

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

Särndal, C.-E., et Swensson, B. (1987), "A General View of Estimation for Two Phases of Selection with Applications to Two-Phase Sampling and Nonresponse", *International Statistical Review*, 55, pp. 279-294.

Särndal, C.-E. (1992), "Methods for Estimating the Precision of Survey Estimates when Imputation has been Used", *Survey Methodology*, 18, pp. 241-252.