N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2003 : Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

2003

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2003 de Statistique Canada
Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

IMPUTATION CALÉE DANS LES ENQUÊTES SOUS L'APPROCHE QUASI ASSISTÉE D'UN MODÈLE

Jean-François Beaumont[1]

RÉSUMÉ

Dans le présent document, nous proposons de recourir à l'imputation calée dans le contexte d'une approche quasi assitée d'un modèle. Cette technique consiste à trouver des valeurs imputées finales qui soient les plus proches possible des valeurs imputées provisoires et qui soient calées de manière à respecter des contraintes. En nous reportant à des contraintes appropriées, nous démontrons que l'estimateur imputé ainsi obtenu est approximativement sans biais pour l'estimation de paramètres de population linéaires comme les totaux de domaines. Nous utilisons la technique de linéarisation de Taylor mise au point par Binder (1983) pour obtenir un estimateur de variance sous un modèle général de non-réponse. Nous montrons que la variance de non-réponse peut être séparée en deux composantes : la première composante est obtenue en supposant que les paramètres du modèle de non-réponse sont connus tandis que la seconde composante correspond à l'effet d'estimer les paramètres du modèle de non-réponse. Il est en outre question du recours à l'imputation calée dans des questions comme celles de la vérification et des valeurs manquantes dans les variables auxiliaires servant à l'obtention des valeurs imputées provisoires.

MOTS CLÉS : Approche quasi assitée d'un modèle, échantillonnage à deux phases, fonction estimante, linéarisation de Taylor, modèle de non-réponse, modèle d'imputation.

1. INTRODUCTION

On recourt souvent à des fonctions estimantes pour justifier la forme des estimateurs imputés dans des enquêtes. Par exemple, l'imputation par la régression linéaire, incluant ses cas spéciaux comme l'imputation par le ratio, utilise les fonctions estimantes afin d'obtenir des valeurs imputées. L'imputation par la régression linéaire est répandue et se retrouve en fait dans bien des systèmes d'imputation comme le Système généralisé de vérification et d'imputation (SGVI) mis au point par Statistique Canada (Bissonnette et Girard, 1998). Il est tout à fait naturel de justifier une imputation par fonction estimante au moyen d'un modèle applicable à la variable d'intérêt, ce qu'on appelle souvent un modèle d'imputation. Il va donc de soi qu'on se serve du modèle d'imputation choisi pour évaluer les propriétés (biais, variance, etc.) de l'estimateur imputé et pour procéder à des inférences. C'est la méthode qu'a proposée Särndal (1992) dans le contexte de l'imputation par le ratio et dont l'application a été étendue par Deville et Särndal (1994) à des plans de sondage et à des méthodes d'imputation par la régression d'un caractère plus général. Quant à l'estimateur par la régression généralisée (REGG), il a été étudié par Gagnon, Lee, Rancourt et Särndal (1996). C'est aussi la technique qui, dans un contexte bayésien, a été avancée par Rubin (1978) et Rubin (1987, chapitre 3).

Même là où un modèle d'imputation sert à l'évaluation et/ou à l'inférence, il est impossible d'éviter d'y aller d'hypothèses au sujet du mécanisme inconnu de non-réponse. En d'autres termes, il est nécessaire de postuler un modèle de non-réponse. Comme dans nombre d'études (par exemple, Särndal, 1992, Deville et Särndal, 1994, et Shao et Steel, 1999), il est courant de supposer seulement que ce mécanisme est indépendant du terme aléatoire du modèle d'imputation (ou « non confondu » avec lui). C'est là une condition suffisante pour rendre la non-réponse ignorable par rapport au modèle d'imputation. Parfois, des modèles de non-réponse plus explicites sont utilisés et les probabilités de réponse sont estimées. Ceci est habituellement requis lorsque l'on suppose que le mécanisme de non-réponse dépend directement de la variable d'intérêt.

Dans le présent document, nous envisageons une démarche différente. Cette démarche peut être définie comme étant *quasi assistée d'un modèle*, en ce sens que les propriétés de l'estimateur imputé sont évaluées par rapport au plan de

[1] Jean-François Beaumont, Statistique Canada, Division des méthodes des enquêtes auprès des ménages, Ottawa (Ontario) K1A 0T6 (courrier électronique : Jean-Francois.Beaumont@statcan.ca).

sondage et à un modèle de non-réponse et que les inférences ne dépendent pas de la validité du modèle d'imputation, le rôle de ce dernier étant simplement de justifier la forme de l'estimateur imputé et d'aider à réduire sa variance due à la non-réponse. Comme dans Oh et Scheuren (1983), nous entendons par « quasi » que le mécanisme de non-réponse est inconnu et qu'un modèle doit être appliqué en conséquence. Cette approche quasi assistée d'un modèle a entre autres été utilisée par Rao et Shao (1992), Rao et Sitter (1995), Fay (1996), Rao (1996), Shao et Sitter (1996) et Shao et Steel (1999), qui ont tous supposé un modèle de non-réponse uniforme (à l'intérieur de classes d'imputation ou non). Nous ne nous limiterons pas ici à une non-réponse uniforme et considérerons des modèles de non-réponse plus généraux. Dans la pratique, cette approche quasi assistée d'un modèle est utile lorsqu'il est difficile de trouver un modèle d'imputation satisfaisant mais qu'un bon modèle de non-réponse peut être obtenu et validé. Dans un tel cas, il est plus raisonnable d'évaluer les propriétés de l'estimateur imputé uniquement par rapport au plan de sondage et au modèle de non-réponse et de ne pas se fier au modèle d'imputation pour procéder à des inférences valides. Bien sûr, ce n'est pas dire qu'on ne devrait pas s'efforcer de trouver le meilleur modèle d'imputation possible, puisque le choix d'un tel modèle influe sur la variance due à la non-réponse de l'estimateur imputé.

Dans le présent document, nous proposons d'utiliser l'imputation calée pour compenser les valeurs manquantes. Cette technique consiste à trouver des valeurs imputées finales qui soient les plus proches possible des valeurs imputées provisoires selon une certaine fonction de distance et qui soient calées de manière à respecter des contraintes. En prenant des contraintes appropriées, nous démontrons que l'estimateur imputé ainsi obtenu est approximativement sans biais pour l'estimation de paramètres de population linéaires et est aussi valide sous une approche quasi assistée d'un modèle Des fonctions estimantes, potentiellement justifiées par un certain modèle d'imputation, sont utilsées afin d'obtenir les valeurs imputées provisoires. L'idée de modifier les valeurs imputées afin de satisfaire certaines contraintes a été proposée par Mantel, Singh et Yu (1995) dans le contexte de l'échantillonnage à deux phases, bien que ces auteurs n'aient pas considéré de fonction de distance spécifique. Ils emploient des techniques d'estimation pour des petites régions pour établir des contraintes appropriées. Beaumont (2000) a aussi envisagé l'imputation calée, mais avec des contraintes différentes qui limitaient l'application au modèle d'imputation linéaire. Dans le cas de variables catégoriques d'intérêt, Favre, Matei et Tillé (2003) et Liu et Rancourt (2001) ont récemment étudié l'imputation calée. Celle-ci est étroitement liée au calage inversé mis de l'avant dans le contexte des données aberrantes et des estimations robustes par Ren (2002) et Beaumont et Alavi (2003). Le calage inversé consiste à trouver des valeurs modifiées qui soient les plus proches possible des valeurs initiales des données aberrantes pour des contraintes imposées par des techniques d'estimation robuste. Ce calage diffère toutefois un peu de l'imputation calée, puisque les valeurs initiales sont inconnues dans le contexte de la non-réponse. On peut aussi relever l'idée de l'établissement de valeurs imputées qui satisfont certaines contraintes dans Deville (2002), qui adopte cependant une orientation différente.

À la section 2, nous commençons par présenter une certaine notation. Nous limitons notre propos à l'estimation de paramètres de population linéaires comme les totaux de domaines. Nous supposons aussi que l'estimateur REGG (ou tout autre estimateur par calage) serait utilisé en cas d'absence de non-réponse. À la section 3, nous décrivons l'imputation calée dans le contexte d'une approche quasi assistée d'un modèle. À la section 4, nous employons la technique de linéarisation de Taylor mise au point par Binder (1983) afin d'obtenir un estimateur de variance pour un modèle général de non-réponse. Nous montrons que la variance de non-réponse peut être séparée en deux composantes : la première composante est obtenue en supposant que les paramètres du modèle de non-réponse sont connus tandis que la seconde composante correspond à l'effet d'estimer les paramètres du modèle de non-réponse. À la section 5, nous discutons du recours à l'imputation calée dans des questions comme celles de vérification et des valeurs manquantes dans les variables auxiliaires servant à l'obtention des valeurs imputées provisoires. Nous concluons par un bref résumé et une discussion à la dernière section.

2. CONTEXTE ET NOTATION

Supposons que notre but est d'estimer un paramètre de population linéaire $t_{dy} = \sum_{k \in P} d_k y_k$ pour une population P, où y_k est la valeur d'une variable d'intérêt y pour l'unité de la population k et d_k, la valeur d'une certaine variable d pour cette même unité. Dans la pratique, d est souvent une variable indicatrice d'un domaine. En d'autres termes, $d_k = 1$ si l'unité k appartient au domaine d'intérêt et $d_k = 0$ dans les autres cas. Comme il est

habituellement impossible d'observer l'ensemble des unités de la population, nous tirons un échantillon aléatoire *s* selon un certain plan de sondage $p(s)$. La probabilité de sélection de l'unité *k* est désignée par π_k, la probabilité de sélection conjointe des unités *k* et *l*, par π_{kl}, et les espérances par rapport au plan de sondage, par $\mathrm{E}_p(\cdot)$. Nous supposons aussi qu'un vecteur de variables auxiliaires $\mathbf{x_1}$ est disponible pour toutes les unités de l'échantillon et que les totaux de population, $\mathbf{t}_{\mathbf{x}_1} = \sum_{k\in P} \mathbf{x_{1k}}$, sont connus pour ces variables. En absence de non-réponse, il serait possible d'utiliser un estimateur par calage $\hat{t}_{dy} = \sum_{k\in s} \tilde{w}_k d_k y_k$ de t_{dy}, où $\tilde{w}_k$, pour $k \in s$, sont les poids d'estimation (voir Deville et Särndal, 1992). L'estimateur par la régression généralisée (REGG) est un important cas spécial d'estimateurs par calage avec des poids d'estimations donnés par $\tilde{w}_k = w_k g_{1k}$, où $w_k = 1/\pi_k$ et

$$g_{1k} = 1 + \frac{\mathbf{x'_{1k}}}{v_{1k}} \left(\sum_{k\in s} \frac{w_k}{v_{1k}} \mathbf{x_{1k}} \mathbf{x'_{1k}} \right)^{-1} \left(\mathbf{t}_{\mathbf{x}_1} - \sum_{k\in s} w_k \mathbf{x_{1k}} \right). \tag{2.1}$$

La constante v_{1k} correspond à la structure de variance du modèle d'estimation sous-jacent à l'estimateur REGG. Par une approximation de Taylor au premier degré (voir, par exemple, Särndal, Swensson et Wretman, 1992, chapitre 6), on peut démontrer que l'estimateur REGG est approximativement sans biais par rapport à *p*. En d'autres termes, $\mathrm{E}_p(\hat{t}_{dy}) \approx t_{dy}$ et sa variance peut s'estimer par

$$\hat{\mathrm{V}}_p(\hat{t}_{dy}) = \sum_{k\in s} \sum_{l\in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl} \pi_k \pi_l} e_{1k} e_{1l}\ , \tag{2.2}$$

où $e_{1k} = d_k y_k - \mathbf{x'_{1k}} \hat{\mathbf{B}}_\mathbf{1}$ et

$$\hat{\mathbf{B}}_\mathbf{1} = \left(\sum_{k\in s} \frac{w_k}{v_{1k}} \mathbf{x_{1k}} \mathbf{x'_{1k}} \right)^{-1} \left(\sum_{k\in s} \frac{w_k}{v_{1k}} \mathbf{x_{1k}} d_k y_k \right).$$

Dans presque toute enquête, il y a des valeurs manquantes et la variable *y* est observée seulement pour une partie de l'échantillon *s*. On observe un ensemble aléatoire de répondants s_r selon un mécanisme de non-réponse $q(s_r \mid s)$. La non-réponse peut être considérée comme une deuxième phase de sélection, la différence étant que l'échantillonneur n'est pas maître du mécanisme de non-réponse et que celui-ci est inconnu, d'où la nécessité de postuler un modèle de non-réponse. Nous supposons que le mécanisme inconnu de non-réponse dépend d'un vecteur de variables auxiliaires $\mathbf{z}$ observées pour chaque unité de l'échantillon, ainsi que d'un vecteur de paramètres inconnus $\boldsymbol{\alpha}$. En d'autres termes, le mécanisme de non-réponse peut s'exprimer comme $q\left(s_r \mid s, \{\mathbf{z_k}; k \in s\}; \boldsymbol{\alpha}\right)$. La probabilité de réponse de l'unité échantillonnale *k* et la probabilité de réponse conjointe de deux unités *k* et *l* sont respectivement désignées par $p_k(\boldsymbol{\alpha})$ et $p_{kl}(\boldsymbol{\alpha})$. Les espérances par rapport au modèle de non-réponse sont désignées par $\mathrm{E}_q(\cdot \mid s)$ et les espérances rar rapport au plan de sondage et du modèle de non-réponse, par $\mathrm{E}_{pq}(\cdot)$. Nous supposons également qu'un estimateur $\boldsymbol{\alpha}^*$ de $\boldsymbol{\alpha}$ s'obtient en utilisant fonction estimante sans biais par rapport *q*, $\mathbf{U_1}(.)$, c'est-à-dire que $\mathbf{E_q}\left(\mathbf{U_1}(\boldsymbol{\alpha}) \mid s\right) = \mathbf{0}$. L'estimateur $\boldsymbol{\alpha}^*$ est donc implicitement défini par l'équation $\mathbf{U_1}(\boldsymbol{\alpha}^*) = \mathbf{0}$. La probabilité de réponse $p_k(\boldsymbol{\alpha})$ et la probabilité de réponse conjointe $p_{kl}(\boldsymbol{\alpha})$ sont respectivement estimées par $p_k(\boldsymbol{\alpha}^*)$ et $p_{kl}(\boldsymbol{\alpha}^*)$.

En présence de non-réponse, il est impossible de calculer $\hat{t}_{dy}$, car la variable *y* n'est pas observée pour l'ensemble aléatoire de non-répondants s_m. Cependant, nous supposons que la variable *d* n'est pas sujette à la non-réponse et est observée pour chaque unité de l'échantillon. Une solution courante au problème de non-réponse consiste à imputer les valeurs manquantes de *y*, ce qui mène à l'estimateur imputé $\hat{t}_{dy}^{I} = \sum_{k\in s} \tilde{w}_k d_k y_{\bullet k}$, où

$$y_{\bullet k} = \begin{cases} y_k & \text{si } k \in s_r, \\ y_k^* & \text{dans les autres cas}, \end{cases} \tag{2.3}$$

et où y_k^* est la valeur imputée finale de l'unité *k*.

3. IMPUTATION CALÉE

Avant de pouvoir établir des valeurs imputées finales par imputation calée, il faut d'abord déterminer les valeurs imputées provisoires $\hat{\mu}_k^*$. Dans le présent document, nous utilisons l'imputation simple déterministe pour trouver les valeurs imputées provisoires. Nous nous intéressons plus particulièrement à l'imputation par fonction estimante. Dans cette méthode d'imputation, les valeurs imputées provisoires $\hat{\mu}_k^*$ dépendent d'un vecteur de variables auxiliaires **x**, observées pour chaque unité de l'échantillon, ainsi que d'un vecteur de paramètres estimés $\hat{\mathbf{B}}^*$. En d'autres termes, $\hat{\mu}_k^* = h(\mathbf{x_k}; \hat{\mathbf{B}}^*)$ pour une certaine fonction connue $h(.;.)$. Le vecteur **x** peut contenir un certain nombre de variables auxiliaires incluses dans les modèles d'estimation et de non-réponse et/ou d'autres variables. L'exigence que **x** soit observé pour toute unité de l'échantillon est assouplie à la section 5.1. Le vecteur $\hat{\mathbf{B}}^*$ est implicitement défini par l'équation $\mathbf{U}_2(\boldsymbol{\alpha}^*, \hat{\mathbf{B}}^*) = \mathbf{0}$, où $\mathbf{U}_2(.,.)$ est une fonction estimante sans biais par rapport à q pour $\hat{\mathbf{B}}$, en ce sens qu'il existe un vecteur $\hat{\mathbf{B}}$ tel que $\mathbf{E_q}\left(\mathbf{U}_2(\boldsymbol{\alpha}, \hat{\mathbf{B}}) \mid s\right) = \mathbf{0}$. Il est ordinairement acceptable de justifier le choix des variables auxiliaires **x**, de la fonction estimante $\mathbf{U}_2(.,.)$ et de la fonction $h(.;.)$ par un modèle d'imputation *m* tel que $\mathrm{E}_m(y_k \mid \mathbf{x_k}) = h(\mathbf{x_k}; \boldsymbol{\beta})$, où $\boldsymbol{\beta}$ est un vecteur de paramètres inconnus du modèle. Toutefois, ce n'est pas une exigence de notre méthode d'obtenir un estimateur imputé qui soit approximativement sans biais par rapport à *q* et **x**, $\mathbf{U}_2(.,.)$ et $h(.;.)$ peuvent être choisis d'une manière tout à fait arbitraire. Il ne faut néanmoins pas faire ces choix au hasard, car ils influent sur la variance due à la non-réponse de l'estimateur imputé, ainsi que nous le démontrons à la section 4.

Dans la pratique, nous désirons souvent estimer plus d'un paramètre de population, en l'occurrence un vecteur de paramètres $\mathbf{t}_{\ddot{\mathbf{d}}y} = \sum_{k\in P} \ddot{\mathbf{d}}_\mathbf{k} y_k$, où $\ddot{\mathbf{d}}_\mathbf{k}$ est un vecteur de valeurs pour l'unité de population *k*. Le vecteur $\ddot{\mathbf{d}}_\mathbf{k}$ n'est pas sujet aux valeurs manquantes et d_k et t_{dy} sont respectivement des éléments de $\ddot{\mathbf{d}}_\mathbf{k}$ et $\mathbf{t}_{\ddot{\mathbf{d}}y}$. Une fois que les vecteurs $\boldsymbol{\alpha}^*$ et $\hat{\mathbf{B}}^*$ sont obtenus, il est possible d'imputer les valeurs manquantes et d'estimer $\mathbf{t}_{\ddot{\mathbf{d}}y}$. Il est courant et naturel de définir les valeurs imputées finales $y_k^* = \hat{\mu}_k^*$, mais cela ne conduit pas nécessairement à un estimateur imputé qui soit approximativement sans biais par rapport à *p et à q* pour le vecteur de paramètres de population $\mathbf{t}_{\ddot{\mathbf{d}}y}$. Pour rendre négligeable le biais dû à la non-réponse, nous proposons une imputation calée. Dans l'approche quasi assistée d'une modèle, cette technique consiste à trouver des valeurs imputées finales y_k^* pour $k \in s_m$ qui soient les plus proches possible des valeurs imputées provisoires $\hat{\mu}_k^*$ et qui soient calées de manière à respecter la contrainte $\hat{\mathbf{t}}_{\ddot{\mathbf{d}}y}^{I} = \hat{\mathbf{t}}_{\ddot{\mathbf{d}}y}^{*}$, où $\hat{\mathbf{t}}_{\ddot{\mathbf{d}}y}^{I} = \sum_{k\in s} \tilde{w}_k \ddot{\mathbf{d}}_\mathbf{k} y_{\bullet k}$ et

$$\hat{\mathbf{t}}_{\ddot{\mathbf{d}}y}^{*} = \sum\nolimits_{k\in s} \tilde{w}_k \ddot{\mathbf{d}}_\mathbf{k} \hat{\mu}_k^* + \sum\nolimits_{k\in s_r} \frac{\tilde{w}_k}{p_k(\boldsymbol{\alpha}^*)} \ddot{\mathbf{d}}_\mathbf{k} \left(y_k - \hat{\mu}_k^*\right). \tag{3.1}$$

Le but de cette contrainte est d'obtenir un estimateur imputé qui soit approximativement sans biais par rapport à q, ce qu'on peut voir en notant que l'estimateur (3.1) a la forme d'un estimateur REGG (non linéaire). On peut aussi aisément constater par (3.1) que l'estimateur imputé $\hat{t}_{dy}^{I}$ est implicitement défini par l'équation

$$\mathrm{U}_3(\boldsymbol{\alpha}^*, \hat{\mathbf{B}}^*, \hat{t}_{dy}^{I}) = \hat{t}_{dy}^{I} - \sum_{k\in s} \tilde{w}_k d_k \hat{\mu}_k^* - \sum_{k\in s_r} \frac{\tilde{w}_k}{p_k(\boldsymbol{\alpha}^*)} d_k \left(y_k - \hat{\mu}_k^*\right) = 0 \ , \tag{3.2}$$

où $\mathrm{U}_3(.,.,.)$ est une fonction estimante sans biais par rapport à q pour $\hat{t}_{dy}$, c'est-à-dire que $\mathrm{E}_q\left(\mathrm{U}_3(\boldsymbol{\alpha}, \hat{\mathbf{B}}, \hat{t}_{dy}) \mid s\right) = 0$. Cette dernière égalité se vérifie à condition que l'hypothèse suivante vaille :

A1) Le mécanisme de non-réponse est indépendant de $\tilde{w}_k$, $\hat{\mu}_k = h(\mathbf{x_k}; \hat{\mathbf{B}})$, y_k et d_k pour $k \in s$ après conditionnement par s et $\mathbf{z_k}$ pour $k \in s$. Toutes les quantités dans $\mathrm{U}_3(\boldsymbol{\alpha}, \hat{\mathbf{B}}, \hat{t}_{dy})$ peuvent être considérées comme fixes sauf l'ensemble aléatoire s_r.

Ainsi, il est primordial de bien choisir les variables auxiliaires **z** à inclure dans le modèle de non-réponse pour que cette hypothèse se vérifie raisonnablement.

Plus explicitement, nous désirons trouver des valeurs imputées finales y_k^* pour $k \in s_m$ qui minimisent la fonction de distance

$$\sum_{k\in s} u_k (y_{\bullet k} - \hat{\mu}_k^*)^2 \tag{3.3}$$

en fonction de la contrainte $\hat{\mathbf{t}}_{\ddot{\mathbf{d}}y}^{I} = \hat{\mathbf{t}}_{\ddot{\mathbf{d}}y}^{*}$. La quantité u_k est un poids qui peut être calculé pour chaque unité de l'échantillon (ou du moins pour les non-répondants). Si un modèle d'imputation est utilisé pour justifier la forme des valeurs imputées provisoires, un choix qui va de soi serait $u_k = \tilde{w}_k / \hat{\sigma}_k^2$, où $\hat{\sigma}_k^2$ est une estimation de la variance du modèle $\mathrm{V}_m(y_k \mid \mathbf{x_k})$. Des fonctions de distance autres que (3.3) pourraient être envisagées pour juger de la proximité entre les valeurs imputées finales et les valeurs imputées provisoires. Nous optons pour la fonction de distance (3.3), puisque c'est un choix qui va de soi dans le contexte de l'estimation par les moindres carrés généralisés et que le calcul des valeurs imputées s'en trouve facilité. Par la méthode des multiplicateurs de Lagrange et par une application algébrique simple, on peut montrer que les valeurs imputées qui minimisent (3.3) en fonction de $\hat{\mathbf{t}}_{\ddot{\mathbf{d}}y}^{I} = \hat{\mathbf{t}}_{\ddot{\mathbf{d}}y}^{*}$ sont données par

$$y_k^* \;=\; \hat{\mu}_k^* \;+\; \frac{\tilde{w}_k}{u_k}\left(\ddot{\mathbf{d}}_\mathbf{k}\right)'\left(\sum_{k\in s_m} \frac{\tilde{w}_k^2}{u_k}\left(\ddot{\mathbf{d}}_\mathbf{k}\right)\left(\ddot{\mathbf{d}}_\mathbf{k}\right)'\right)^{-1} \sum_{k\in s_r} \tilde{w}_k \frac{\left(1 - p_k(\boldsymbol{\alpha}^*)\right)}{p_k(\boldsymbol{\alpha}^*)}\left(y_k - \hat{\mu}_k^*\right)\ddot{\mathbf{d}}_\mathbf{k} \ . \tag{3.4}$$

Ainsi, nous obtenons les valeurs imputées finales en ajoutant un terme d'ajustement pour le biais dû à la non-réponse aux valeurs imputées provisoires $\hat{\mu}_k^*$. Ce terme d'ajustement disparaît et $y_k^* = \hat{\mu}_k^*$ lorsque le second terme de (3.4) est nul. Tel est le cas si, par exemple, les deux conditions suivantes sont réunies :

(i) $\mathbf{U_2}(\boldsymbol{\alpha}^*, \hat{\mathbf{B}}^*) = \sum_{k\in s_r} a_k (y_k - \hat{\mu}_k^*) \dfrac{\mathbf{x_k}}{v_k} = \mathbf{0}$, où $a_k = \tilde{w}_k \left(1 - p_k(\boldsymbol{\alpha}^*)\right) / p_k(\boldsymbol{\alpha}^*)$ est un poids de régression et où v_k est une constante connue qui peut être justifiée par une hypothèse quant à la variance du modèle $\mathrm{V}_m(y_k \mid \mathbf{x_k})$;

(ii) $\ddot{\mathbf{d}}_\mathbf{k} v_k = \boldsymbol{\Lambda}\mathbf{x_k}$, où $\boldsymbol{\Lambda}$ est une matrice de constantes connues.

La condition (ii) est respectée en incluant $\ddot{\mathbf{d}}_{\mathbf{k}} v_k$ dans le vecteur $\mathbf{x}_{\mathbf{k}}$, mais cela peut ne pas toujours convenir là où le nombre de paramètres de population à estimer est très grand ou, pour l'exprimer autrement, là où $\ddot{\mathbf{d}}_{\mathbf{k}}$ est d'une très grande dimension. D'ordinaire, pour éviter que l'estimateur $\hat{\mathbf{B}}^*$ soit instable, la dimension de $\mathbf{x}_{\mathbf{k}}$ n'est pas trop grande et seules les variables qui sont suffisamment corrélées avec *y* font partie du vecteur **x**. Ainsi, la dimension de $\ddot{\mathbf{d}}_{\mathbf{k}}$ sera probablement supérieure à celle de $\mathbf{x}_{\mathbf{k}}$ en pratique. Il n'est pas toujours bon, par conséquent, d'écarter le terme d'ajustement pour le biais dû à la non-réponse.

L'estimateur $\hat{\boldsymbol{\theta}}^* = \left((\boldsymbol{\alpha}^*)', (\hat{\mathbf{B}}^*)', \hat{t}_{dy}^I\right)'$ est implicitement défini par l'équation

$$\mathbf{U}(\hat{\boldsymbol{\theta}}^*) = \begin{pmatrix} \mathbf{U}_1(\boldsymbol{\alpha}^*) \\ \mathbf{U}_2(\boldsymbol{\alpha}^*, \hat{\mathbf{B}}^*) \\ \mathrm{U}_3(\boldsymbol{\alpha}^*, \hat{\mathbf{B}}^*, \hat{t}_{dy}^I) \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ 0 \end{pmatrix}, \tag{3.5}$$

où la fonction estimante $\mathbf{U}(.)$ est sans biais par rapport à *q* pour $\hat{\boldsymbol{\theta}} = \left(\boldsymbol{\alpha}', \hat{\mathbf{B}}', \hat{t}_{dy}\right)'$, c'est-à-dire que $\mathbf{E}_{\mathbf{q}}\left(\mathbf{U}(\hat{\boldsymbol{\theta}}) \mid s\right) = \mathbf{0}$. En utilisant une approximation de Taylor au premier degré (voir Binder, 1983), on peut facilement démontrer que l'estimateur $\hat{\boldsymbol{\theta}}^*$ est approximativement sans biais par rapport à *q* pour $\hat{\boldsymbol{\theta}}$ et que, en particulier, l'estimateur imputé $\hat{t}_{dy}^I$ est approximativement sans biais par rapport à *q* pour l'estimateur par calage de réponse complète $\hat{t}_{dy}$, c'est-à-dire que $\mathrm{E}_q(\hat{t}_{dy}^I \mid s) \approx \hat{t}_{dy}$. La grande différence d'avec ce que propose Binder (1983) est que les espérances s'évaluent dans le cas présent par rapport au modèle de non-réponse, alors que, dans le cas de Binder, elles s'évaluent par rapport au plan de sondage. Comme $\mathrm{E}_q(\hat{t}_{dy}^I \mid s) \approx \hat{t}_{dy}$, $\hat{t}_{dy}^I$ est aussi approximativement sans biais par rapport à *p* et à *q* pour le paramètre de population t_{dy}, c'est-à-dire que $\mathrm{E}_{pq}(\hat{t}_{dy}^I) \approx t_{dy}$. Cette propriété de l'estimateur $\hat{t}_{dy}^I$ d'être approximativement sans biais par rapport à *q* se vérifie quelle que soit la validité du modèle d'imputation (si on en a spécifié un). Aussi le modèle d'imputation a-t-il simplement pour rôle de donner un moyen d'accroître l'efficience de l'estimateur imputé sans toutefois compter sur sa compétence à être approximativement sans biais par rapport à *q*. Cependant, pour être valide, le modèle de non-réponse doit être valide pour cette approche quasi assistée d'un modèle.

4. ESTIMATION DE LA VARIANCE

On peut procéder par analogie avec l'échantillonnage à deux phases (voir, par exemple, Hidiroglou et Särndal, 1998) pour estimer la variance de l'estimateur imputé $\hat{t}_{dy}^I$. Étant donné que $\mathrm{E}_q(\hat{t}_{dy}^I \mid s) \approx \hat{t}_{dy}$, la variance de $\hat{t}_{dy}^I$ par rapport à *p* et à *q* peut être approximativement donnée par :

$$\mathrm{V}_{pq}(\hat{t}_{dy}^I) \approx \mathrm{V}_p(\hat{t}_{dy}) + \mathrm{E}_p \mathrm{V}_q(\hat{t}_{dy}^I \mid s) \ . \tag{4.1}$$

Le premier terme du côté droit de (4.1), $\mathrm{V}_p(\hat{t}_{dy})$, est souvent appelé variance due à l'échantillonnage et le second, $\mathrm{E}_p \mathrm{V}_q(\hat{t}_{dy}^I \mid s)$, variance due à la non-réponse.

En cas de réponse complète, la variance due à l'échantillonnage peut s'estimer par (2.2), mais (2.2) ne peut se calculer en présence de non-réponse, puisque y_k n'est pas observé pour les unités non répondantes $k \in s_m$. Par analogie avec l'échantillonnage à deux phases, la variance due à l'échantillonnage peut néanmoins s'estimer par

$$\hat{\mathrm{V}}_p^*(\hat{t}_{dy}) = \sum_{k\in s_r}\sum_{l\in s_r}\frac{1}{p_{kl}(\boldsymbol{\alpha}^*)}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}\pi_k\pi_l}\, e_{1k}^* e_{1l}^* \,, \tag{4.2}$$

où $e_{1k}^* = d_k y_k - \mathbf{x}_{\mathbf{1k}}'\hat{\mathbf{B}}_{\mathbf{1}}^*$ et

$$\hat{\mathbf{B}}_{\mathbf{1}}^* = \left(\sum_{k\in s_r}\frac{w_k}{p_k(\boldsymbol{\alpha}^*)}\frac{1}{v_{1k}}\mathbf{x}_{\mathbf{1k}}\mathbf{x}_{\mathbf{1k}}'\right)^{-1}\left(\sum_{k\in s_r}\frac{w_k}{p_k(\boldsymbol{\alpha}^*)}\frac{1}{v_{1k}}\mathbf{x}_{\mathbf{1k}}d_k y_k\right).$$

Un estimateur approximativement sans biais par rapport à *p* et à *q* de la variance due à la non-réponse $\mathrm{E}_p\mathrm{V}_q(\hat{t}_{dy}^I \mid s)$ peut s'obtenir en trouvant un estimateur approximativement sans biais par rapport à *q* de la variance conditionnelle due à la non-réponse $\mathrm{V}_q(\hat{t}_{dy}^I \mid s)$. Si on emploie la technique de linéarisation de Taylor proposée par Binder (1983) et qu'on remplace les espérances par rapport au plan de sondage par les espérances par rapport au modèle de non-réponse, nous constatons que $\mathbf{V_q}(\hat{\boldsymbol{\theta}}^* \mid s)$ peut être approximativement donnée par $\dot{\mathbf{V}}_{\mathbf{q}}(\hat{\boldsymbol{\theta}}^* \mid s)$, où

$$\dot{\mathbf{V}}_{\mathbf{q}}(\hat{\boldsymbol{\theta}}^* \mid s) = \left[\mathbf{H}(\hat{\boldsymbol{\theta}})\right]^{-1}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\left[\mathbf{H}'(\hat{\boldsymbol{\theta}})\right]^{-1} \tag{4.3}$$

et où

$$\mathbf{H}(\tilde{\boldsymbol{\theta}}) = \mathbf{E_q}\left(\frac{\partial\mathbf{U}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\theta}}'} \mid s\right) \text{ pour un certain vecteur } \tilde{\boldsymbol{\theta}} = \left(\tilde{\boldsymbol{\alpha}}', \tilde{\mathbf{B}}', \tilde{t}_{dy}\right)' \text{ et } \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}) = \mathbf{E_q}\left(\mathbf{U}(\hat{\boldsymbol{\theta}})\mathbf{U}'(\hat{\boldsymbol{\theta}}) \mid s\right).$$

Ainsi, la variance conditionnelle due à la non-réponse $\mathrm{V}_q(\hat{t}_{dy}^I \mid s)$ est approximativement donnée par la valeur de la dernière ligne et de la dernière colonne de la matrice définie du côté droit de (4.3).

Pour obtenir une expression plus explicite de $\mathrm{V}_q(\hat{t}_{dy}^I \mid s)$, désignons d'abord l'estimateur imputé $\hat{t}_{dy}^I$ par $\hat{t}_{dy}^I(\boldsymbol{\alpha}^*)$ pour bien souligner que cet estimateur imputé dépend d'un vecteur de paramètres estimés du modèle de non-réponse $\boldsymbol{\alpha}^*$. L'estimateur imputé que l'on obtiendrait si les paramètres du modèle de non-réponse étaient connus est donc désigné par $\hat{t}_{dy}^I(\boldsymbol{\alpha})$. Il est alors possible de démontrer (voir Beaumont, 2004), après un traitement algébrique simple mais quelque peu fastidieux, que la variance approximative $\dot{\mathrm{V}}_q(\hat{t}_{dy}^I(\boldsymbol{\alpha}^*) \mid s)$ pour $\mathrm{V}_q(\hat{t}_{dy}^I(\boldsymbol{\alpha}^*) \mid s)$ est donnée par

$$\dot{\mathrm{V}}_q\left(\hat{t}_{dy}^I(\boldsymbol{\alpha}^*) \mid s\right) \;=\; \dot{\mathrm{V}}_q\left(\hat{t}_{dy}^I(\boldsymbol{\alpha}) \mid s\right) \;+\; \left\{-\mathbf{H_{31}}(\hat{\boldsymbol{\theta}})\dot{\mathbf{V}}_{\mathbf{q}}\left(\boldsymbol{\alpha}^* \mid s\right)\mathbf{H}_{\mathbf{31}}'(\hat{\boldsymbol{\theta}}) - 2\,\mathbf{H_{31}}(\hat{\boldsymbol{\theta}})\dot{\mathbf{C}}_{\mathbf{q}}\left(\boldsymbol{\alpha}^*, \hat{t}_{dy}^I \mid s\right)\right\}, \tag{4.4}$$

où

$$\dot{\mathrm{V}}_q\left(\hat{t}_{dy}^I(\boldsymbol{\alpha}) \mid s\right) = V_q\left(\mathrm{U}_3(\boldsymbol{\alpha}, \hat{\mathbf{B}}, \hat{t}_{dy}) \mid s\right) = \sum_{k\in s}\sum_{l\in s}\frac{p_{kl}(\boldsymbol{\alpha}) - p_k(\boldsymbol{\alpha})p_l(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})p_l(\boldsymbol{\alpha})}\,[\tilde{w}_k(y_k - \hat{\mu}_k)d_k]\,[\tilde{w}_l(y_l - \hat{\mu}_l)d_l],$$

$$\dot{\mathbf{V}}_{\mathbf{q}}\left(\boldsymbol{\alpha}^* \mid s\right) = \left[\mathbf{H_{11}}(\hat{\boldsymbol{\theta}})\right]^{-1}\,\mathbf{E_q}\left(\mathbf{U_1}(\boldsymbol{\alpha})\,\mathbf{U_1}'(\boldsymbol{\alpha}) \mid s\right)\left[\mathbf{H}_{\mathbf{11}}'(\hat{\boldsymbol{\theta}})\right]^{-1},$$

$$\dot{\mathbf{C}}_{\mathbf{q}}\left(\boldsymbol{\alpha}^*, \hat{t}_{dy}^I \mid s\right) = -\dot{\mathbf{V}}_{\mathbf{q}}\left(\boldsymbol{\alpha}^* \mid s\right)\mathbf{H}_{\mathbf{31}}'(\hat{\boldsymbol{\theta}}) + \mathbf{H}_{\mathbf{11}}^{-1}(\hat{\boldsymbol{\theta}})\,\mathbf{E_q}\left(\mathbf{U_1}(\boldsymbol{\alpha})\,\mathrm{U}_3(\boldsymbol{\alpha}, \hat{\mathbf{B}}, \hat{t}_{dy}) \mid s\right),$$

$$\mathbf{H_{31}}(\hat{\boldsymbol{\theta}}) = \mathbf{E_q}\left(\frac{\partial\,\mathrm{U}_3(\tilde{\boldsymbol{\alpha}}, \tilde{\mathbf{B}}, \tilde{t}_{dy})}{\partial\tilde{\boldsymbol{\alpha}}'} \mid s\right)\Bigg|_{\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}} = \sum_{k\in s}\frac{\tilde{w}_k}{p_k(\boldsymbol{\alpha})}(y_k - \hat{\mu}_k)d_k\left(\frac{\partial p_k(\tilde{\boldsymbol{\alpha}})}{\partial\tilde{\boldsymbol{\alpha}}'}\Bigg|_{\tilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha}}\right),$$

$$\mathbf{H_{11}}(\tilde{\boldsymbol{\theta}}) = \mathbf{E_q}\left(\frac{\partial\mathbf{U_1}(\tilde{\boldsymbol{\alpha}})}{\partial\tilde{\boldsymbol{\alpha}}'} \mid s\right).$$

La variance conditionnelle approximative due à la non-réponse $\dot{\mathrm{V}}_q(\hat{t}_{dy}^I(\boldsymbol{\alpha}^*)\,|\,s)$ peut ainsi se diviser en deux composantes. La première composante, $\dot{\mathrm{V}}_q\left(\hat{t}_{dy}^I(\boldsymbol{\alpha})\,|\,s\right)$, correspond à la variance conditionnelle approximative due à la non-réponse que l'on obtiendrait si les paramètres du modèle de non-réponse étaient connus et la seconde, qui est l'élément en accolade dans (4.4), correspond à l'effet d'estimer les paramètres du modèle de non-réponse. Cet effet peut être positif ou négatif. On peut démontrer qu'il est négatif pour le modèle de non-réponse logistique (voir les détails dans Beaumont, 2004).

Dans la pratique, on peut être tenté de supposer que $\boldsymbol{\alpha}^*$ est suffisamment proche de $\boldsymbol{\alpha}$ pour approximer la variance conditionnelle due à la non-réponse $\mathrm{V}_q(\hat{t}_{dy}^I(\boldsymbol{\alpha}^*)\,|\,s)$ par la première composante de (4.4), $\dot{\mathrm{V}}_q\left(\hat{t}_{dy}^I(\boldsymbol{\alpha})\,|\,s\right)$. Cette approximation simplifie grandement l'estimation de variance. Elle a en fait été employée par Beaumont et Mitchell (2002) dans l'élaboration du Système pour l'estimation de la variance due à la non-réponse et à l'imputation (System for Estimation of Variance due to Nonresponse and Imputation ou SEVANI). Dans le contexte de l'estimation *bootstrap* de la variance, Mantel, Nadon et Yeo (2000) ont empiriquement constaté que, si on considère $\boldsymbol{\alpha}^*$ comme fixe, on obtient souvent des estimations de variance un peu supérieures à celles qui seraient obtenues sans cette approximation. De (4.4), on peut voir que cette approximation est appropriée lorsque $\mathbf{H}_{31}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$. Cette dernière équation se vérifie si toutes les unités de l'échantillon répondent avec la même probabilité et si les conditions (i) et (ii) énoncées à la section 3 sont satisfaites. Cependant, la seconde composante de (4.4) peut ne pas toujours être négligeable, surtout si le modèle d'imputation *m* n'est pas convenablement spécifié. Dans un tel cas, non seulement $\mathbf{H}_{31}(\hat{\boldsymbol{\theta}})$ ne sera-t-il probablement pas négligeable, mais $\mathbf{E}_{\mathbf{m}}\left\{\mathbf{H}_{31}(\hat{\boldsymbol{\theta}})\,|\,s\right\}$ aussi. Dans une étude par simulation, Beaumont (2004) démontre que, en négligeant la seconde composante de (4.5), on est amené à des inférences raisonnablement conservatrices.

Il convient de noter que l'équation (4.4) ne dépend pas de la fonction estimante $\mathbf{U}_2(.,.)$ et que, par conséquent, cette fonction peut être relativement complexe sans ajouter à la complexité de l'expression de $\dot{\mathrm{V}}_q(\hat{t}_{dy}^I(\boldsymbol{\alpha}^*)\,|\,s)$. On peut aussi constater que la variance approximative $\dot{\mathrm{V}}_q(\hat{t}_{dy}^I(\boldsymbol{\alpha}^*)\,|\,s)$ dépend des résidus $e_k = y_k - \hat{\mu}_k$. Il importe donc de bien choisir le vecteur de variables auxiliaires **x**, la fonction estimante $\mathbf{U}_2(.,.)$ et la fonction $h(.;.)$, de sorte que les résidus e_k restent les plus petits possible, ce qu'on peut faire en trouvant le modèle d'imputation qui convient le mieux en l'occurrence. À noter cependant que la validité de l'expression de la variance (4.4) ne dépend pas de la validité du modèle d'imputation.

On peut obtenir un estimateur de variance pour $\mathrm{V}_q(\hat{t}_{dy}^I(\boldsymbol{\alpha}^*)\,|\,s)$ en estimant les quantités inconnues dans (4.4). Il est donné par

$$\hat{\mathrm{V}}_q\left(\hat{t}_{dy}^I(\boldsymbol{\alpha}^*)\,|\,s\right) \;=\; \hat{\mathrm{V}}_q\left(\hat{t}_{dy}^I(\boldsymbol{\alpha})\,|\,s\right) \;+\; \left\{-\hat{\mathbf{H}}_{31}(\hat{\boldsymbol{\theta}}^*)\hat{\mathbf{V}}_{\mathbf{q}}\left(\boldsymbol{\alpha}^*\,|\,s\right)\hat{\mathbf{H}}'_{31}(\hat{\boldsymbol{\theta}}^*) - 2\,\hat{\mathbf{H}}_{31}(\hat{\boldsymbol{\theta}}^*)\hat{\mathbf{C}}_{\mathbf{q}}\left(\boldsymbol{\alpha}^*, \hat{t}_{dy}^I\,|\,s\right)\right\}, \qquad (4.5)$$

où

$$\hat{\mathrm{V}}_q\left(\hat{t}_{dy}^I(\boldsymbol{\alpha})\,|\,s\right) = \sum_{k\in s_r}\sum_{l\in s_r}\frac{p_{kl}(\boldsymbol{\alpha}^*) - p_k(\boldsymbol{\alpha}^*)p_l(\boldsymbol{\alpha}^*)}{p_{kl}(\boldsymbol{\alpha}^*)p_k(\boldsymbol{\alpha}^*)p_l(\boldsymbol{\alpha}^*)}\left[\tilde{w}_k(y_k - \hat{\mu}_k^*)d_k\right]\left[\tilde{w}_l(y_l - \hat{\mu}_l^*)d_l\right],$$

$$\hat{\mathbf{V}}_{\mathbf{q}}\left(\boldsymbol{\alpha}^*\,|\,s\right) = \left[\hat{\mathbf{H}}_{\mathbf{11}}(\hat{\boldsymbol{\theta}}^*)\right]^{-1}\hat{\boldsymbol{\Sigma}}_{\mathbf{11}}(\hat{\boldsymbol{\theta}}^*)\left[\hat{\mathbf{H}}'_{\mathbf{11}}(\hat{\boldsymbol{\theta}}^*)\right]^{-1},$$

$$\hat{\mathbf{C}}_{\mathbf{q}}\left(\boldsymbol{\alpha}^*, \hat{t}_{dy}^I\,|\,s\right) = -\hat{\mathbf{V}}_{\mathbf{q}}\left(\boldsymbol{\alpha}^*\,|\,s\right)\hat{\mathbf{H}}'_{\mathbf{31}}(\hat{\boldsymbol{\theta}}^*) + \hat{\mathbf{H}}_{\mathbf{11}}^{-1}(\hat{\boldsymbol{\theta}}^*)\,\hat{\boldsymbol{\Sigma}}_{\mathbf{13}}(\hat{\boldsymbol{\theta}}^*),$$

$$\hat{\mathbf{H}}_{\mathbf{31}}(\hat{\boldsymbol{\theta}}^*) = \sum_{k\in s_r}\frac{\tilde{w}_k}{p_k^2(\boldsymbol{\alpha}^*)}(y_k - \hat{\mu}_k^*)d_k\left(\left.\frac{\partial p_k(\tilde{\boldsymbol{\alpha}})}{\partial\tilde{\boldsymbol{\alpha}}'}\right|_{\tilde{\boldsymbol{\alpha}}=\boldsymbol{\alpha}^*}\right),$$

$$\hat{\mathbf{H}}_{\mathbf{11}}(\hat{\boldsymbol{\theta}}^*) = \left.\frac{\partial\mathbf{U}_1(\tilde{\boldsymbol{\alpha}})}{\partial\tilde{\boldsymbol{\alpha}}'}\right|_{\tilde{\boldsymbol{\alpha}}=\boldsymbol{\alpha}^*}.$$

et où $\hat{\boldsymbol{\Sigma}}_{11}(\hat{\boldsymbol{\theta}}^*)$ et $\hat{\boldsymbol{\Sigma}}_{13}(\hat{\boldsymbol{\theta}}^*)$ sont respectivement des estimateurs de $\mathbf{E_q}\left(\mathbf{U_1}(\boldsymbol{\alpha})\mathbf{U_1'}(\boldsymbol{\alpha}) \mid s\right)$ et $\mathbf{E_q}\left(\mathbf{U_1}(\boldsymbol{\alpha})\,\mathrm{U}_3(\boldsymbol{\alpha},\hat{\mathbf{B}},\hat{t}_{dy}) \mid s\right)$. On peut alors estimer la variance totale $\mathrm{V}_{pq}\left(\hat{t}_{dy}^{I}(\boldsymbol{\alpha}^*)\right)$ en ajoutant l'estimateur de variance due à l'échantillonnage (4.2) à l'estimateur de variance due à la non-réponse (4.5).

5. QUELQUES QUESTIONS

5.1 Valeurs manquantes dans les variables x

Jusqu'à présent, nous avons supposé que les variables auxiliaires servant à l'obtention des valeurs imputées provisoires n'étaient pas sujettes aux valeurs manquantes, mais il est courant d'utiliser certaines variables d'intérêt pour en imputer d'autres. Nous sommes alors dans une situation où le vecteur de variables auxiliaires $\mathbf{x_k}$ peut ne pas être entièrement observé pour un certain nombre d'unités de l'échantillon et où plusieurs méthodes d'imputation peuvent être utilisées. Par exemple, si nous avons deux variables d'intérêt *y* et *x*, nous pourrions vouloir imputer la valeur manquante de *y* pour une unité *k* à l'aide d'un modèle de régression linéaire entre *y* et *x* si x_k est connu, et à l'aide d'une imputation par la moyenne dans les autres cas. Ainsi, la valeur imputée provisoire d'une unité *k*, $\hat{\mu}_k^*$, dépend du patron de non-réponse $\mathbf{r_k^{(x)}}$, où la j^{e} composante de $\mathbf{r_k^{(x)}}$ sera 0 ou 1 selon que la j^{e} composante de $\mathbf{x_k}$ est manquante ou non. En d'autres termes, $\hat{\mu}_k^* = h(\mathbf{x_k}, \mathbf{r_k^{(x)}}; \hat{\mathbf{B}}^*)$. Bien sûr, $\hat{\mu}_k^*$ ne peut dépendre des valeurs non observées du vecteur $\mathbf{x_k}$. La fonction estimante $\mathbf{U}_2(.,.)$ peut devenir plutôt complexe si on recourt à plusieurs méthodes d'imputation différentes. Heureusement, cela n'ajoute pas de complexité à l'estimation de la variance puisque l'estimateur de variance (4.5) ne dépend pas de cette fonction estimante. Ainsi, les éléments théoriques présentés aux sections 3 et 4 peuvent s'appliquer directement en cas de valeurs manquantes dans les variables **x**. À noter que l'hypothèse (A1) énoncée à la section 3 doit toujours être valide avec $\hat{\mu}_k = h(\mathbf{x_k}; \hat{\mathbf{B}})$ qui est remplacé par $\hat{\mu}_k = h(\mathbf{x_k}, \mathbf{r_k^{(x)}}; \hat{\mathbf{B}})$. Il est donc important de songer à inclure $\mathbf{r^{(x)}}$ dans le vecteur de variables auxiliaires **z** pour que cette hypothèse soit raisonnablement satisfaite, car $\mathbf{r_k^{(x)}}$ pour $k \in s$ pourrait être relié au patron de non-réponse de la variable *y*.

5.2 Vérification

On applique souvent des règles de vérification pour restreindre les valeurs possibles à être imputées. Dans l'imputation calée, on peut en tenir compte en considérant l'ensemble des règles de vérification comme une contrainte supplémentaire de la procédure de minimisation. Il n'y a aucune incidence sur les propriétés du vecteur d'estimateurs imputés $\hat{\mathbf{t}}_{\mathbf{d}y}^{I}$, puisque les valeurs imputées finales sont quand même calées de façon à satisfaire la contrainte $\hat{\mathbf{t}}_{\mathbf{d}y}^{I} = \hat{\mathbf{t}}_{\mathbf{d}y}^{*}$. Cependant, cela a un effet sur les valeurs imputées finales. Qui plus est, il est généralement impossible d'obtenir une solution explicite par la méthode des multiplicateurs de Lagrange, et il faut habituellement un algorithme numérique pour obtenir ces valeurs finales, et ce, parce que les règles de vérification sont souvent plus complexes que de simples équations. De plus, comme ces règles de vérification font intervenir dans bien des cas plusieurs variables d'intérêt à la fois, une fonction de distance multivariée s'impose. C'est un secteur où on doit pousser la recherche, plus particulièrement dans le contexte des variables catégoriques d'intérêt.

6. CONCLUSION

Dans le présent document, nous avons proposé d'utiliser l'imputation calée dans le cadre d'une approche quasi assistée d'un modèle. Cette technique consiste à trouver des valeurs imputées finales qui soient les plus proches possible des valeurs imputées provisoires et qui soient calées de manière à respecter des contraintes, celles-ci étant choisies de sorte que l'estimateur imputé ainsi obtenu soit approximativement sans biais par rapport au modèle de non-réponse sans que ce dernier soit nécessairement un modèle de non-réponse uniforme. S'il n'est pas nécessaire

d'établir un modèle d'imputation sous une forme explicite, il est naturel d'en employer un pour justifier la forme de l'estimateur imputé. Même si un modèle d'imputation est utilisé, il n'est pas nécessaire qu'il soit spécifié correctement afin d'obtenir des inférences valides.

Dans le présent document, nous avons supposé que tous les paramètres de population étaient pris en compte au stade de l'imputation d'une enquête. Bien que plusieurs paramètres de population sont souvent connus d'avance ou qu'on s'attende à ce qu'ils soient d'intérêt, tel n'est pas toujours le cas, surtout si on produit un fichier de microdonnées à grande diffusion. Le problème avec des paramètres de population non planifiés est que nous ne pouvons être sûrs que le biais de non-réponse de l'estimateur imputé est négligeable. Même si on compte davantage sur un modèle d'imputation comme dans Särndal (1992) ou sur une imputation multiple, on sait fort bien (voir, par exemple, Meng, 1994, et Rubin, 1996) qu'il faut tenir compte, au moment de choisir le modèle d'imputation, des variables devant servir à définir les paramètres de population à estimer, sinon rien ne garantit que des inférences valides basées sur un modèle d'imputation seront possibles. Par contre, si le modèle d'imputation a soigneusement été validé et se révèle suffisamment riche, la perspective de tirer des inférences basées sur le modèle d'imputation pourrait toutefois représenter un choix attrayant pour les analystes qui utilisent des fichiers de microdonnées à grande diffusion. Dans ce cas, des méthodes de ré-échantillonnage comme le bootstrap ou l'imputation multiple peuvent grandement simplifier l'estimation de la variance.

REMERCIEMENTS

L'auteur désire remercier J.N.K. Rao, de l'Université Carleton, ainsi que David Binder, Steve Matthews et Eric Rancourt, de Statistique Canada, de leurs utiles suggestions qui ont aidé à améliorer la qualité du présent document.

RÉFÉRENCES

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

Beaumont, J.-F. (2004), "Calibrated Imputation in Surveys under a Quasi-Model-Assisted Approach", manuscrit non publié, Statistique Canada.

Beaumont, J.-F., et Alavi, A. (2003), "Robust Generalized Regression Estimation", manuscrit non publié soumis à *Techniques d'enquête* .

Beaumont, J.-F., et Mitchell, C. (2002), "Système pour l'estimation de la variance due à la non-réponse et à l'imputation (SEVANI)", *Recueil du Symposium 2002 de Statistique Canada, Modélisation des données d'enquête pour la recherche économique et sociale*, Statistique Canada.

Binder, D.A. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys", *International Statistical Review*, 51, pp. 279-292.

Bissonnette, J., et Girard, J. (1998), "Improvements in Imputation Estimators: the ESTIMP Module", manuscrit non publié, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada.

Deville, J.-C. (2002), "Imputation par prédiction ou imputation avec aléa?", *Recueil des Journées de méthodologie statistique*, INSEE.

Deville, J.-C., et Särndal, C.-E. (1992), "Calibration Estimators in Survey Sampling", *Journal of the American Statistical Association*, 87, pp. 376-382.

Deville, J.-C., et Särndal, C.-E. (1994), "Variance Estimation for the Regression Imputed Horvitz-Thompson Estimator", *Journal of Official Statistics*, 10, pp. 381-394.

Favre, A.-C., Matei, A., et Tillé, Y. (2003), "Calibrated Random Imputation for Qualitative Data", *Journal of Statistical Planning and Inference* (à paraître).

Fay, R.E. (1996), "Alternative Paradigms for the Analysis of Imputed Survey Data", *Journal of the American Statistical Association*, 91, pp. 490-498.

Gagnon, F., Lee, H., Rancourt, E., et Särndal (1996), "Estimating the Variance of the Generalized Regression Estimator in the Presence of Imputation for the Generalized Estimation System", *Proceedings of the Survey Methods Section, Statistical Society of Canada*, pp. 151-156.

Hidiroglou, M.A., et Särndal, C.-E. (1998), "Emploi des données auxiliaires dans l'échantillonnage à deux phases", *Techniques d'enquête*, 24, pp. 11-20.

Liu, T.-P., et Rancourt, E. (2001), "Constrained Categorical Imputation for Nonresponse in Surveys", Document de travail de la direction de la méthodologie, no. HSMD-2001-012E, Statistique Canada.

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

Mantel, H.J., Singh, A.C., et Yu, M. (1995), "Mass Imputation for Two Phase Sampling: Use of Small Area Estimation and Calibration Techniques", *Proceedings of the Survey Methods Section, Statistical Society of Canada*, pp. 57- 62.

Meng, X.L. (1994), "Multiple Imputation with Uncongenial Sources of Input (with Discussion)", *Statistical Science*, 9, pp. 538-574.

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

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

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

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

Ren, R. (2002), "Méthodes d'imputation de valeurs aberrantes pour des données d'enquête" *Recueil des Journées de méthodologie statistique*, INSEE.

Rubin, D.B. (1978), "Multiple Imputation in Sample Surveys - a Phenomenological Bayesian Approach to Nonresponse", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 20-34.

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

Rubin, D.B. (1978), "Multiple Imputation after 18+ Years", *Journal of the American Statistical Association, 91, pp. 473-489*.

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

Särndal, C.-E., Swensson, B., et Wretman, J.H. (1992), *Model Assisted Survey Sampling*, New-York, Springer-Verlag.

Shao, J., et Sitter, R. (1996), "Bootstrap for Imputed Survey Data", *Journal of the American Statistical Association*, 91, pp. 1278-1288.

Shao, J., et Steel, P. (1999), "Variance Estimation for Survey Data with Composite Imputation and Nonnegligible Sampling Fractions", *Journal of the American Statistical Association*, 94, pp. 254-265.