Assemblée annuelle de la SSC, Juin 2011
Recueil de la Section des méthodes d'enquête

IMPUTATION POUR DES POPULATIONS CONTENANT BEAUCOUP DE ZÉROS

Christian O. Nambeu[1], David Haziza[2] et Guillaume Chauvet[3]

RÉSUMÉ

L'imputation simple est très souvent utilisée dans les enquêtes pour compenser la non-réponse partielle. Dans certaines situations, la variable devant être imputée prend des valeurs égales à zéro un très grand nombre de fois. Ceci est très fréquent dans les enquêtes entreprises qui collectent les variables économiques. Dans cet article, nous étudions les propriétés de deux méthodes d'imputation souvent utilisées en pratique et nous montrons qu'elles produisent des estimateurs biaisés en général. Motivé par un modèle de mélange, nous proposons trois méthodes d'imputation et étudions leurs propriétés en termes de biais. Finalement, nous effectuons une étude par simulations pour étudier la performance des estimateurs ponctuels en termes de biais relatif.

MOTS CLÉS : Non-réponse partielle; Imputation aléatoire équilibrée; Imputation par la régression.

ABSTRACT

Single imputation is often used in surveys to compensate for item nonresponse. In some cases, the variable requiring imputation contains a large number of zeroes. This is especially frequent in business surveys that collect economic variables. In this paper, we study the properties of two imputation procedures frequently used in practice and show that they lead to biased estimators, in general. Motivated by a mixture regression model, we then propose three imputation procedures and study their properties in terms of bias. Finally, we perform a simulation study to evaluate the performance of point estimators in terms of relative bias.

KEY WORDS: Balanced random imputation; Item nonresponse; Regression imputation

1. INTRODUCTION

Une des manières de minimiser l'impact de la non-réponse sur les estimations d'enquêtes est l'utilisation de l'imputation. Cette dernière consiste à remplacer les valeurs manquantes par des valeurs artificielles obtenues au moyen d'un modèle. On peut choisir de modéliser le mécanisme de non-réponse ou la variable d'intérêt en se servant d'information auxiliaire de bonne qualité si disponible. Le choix de l'approche utilisée dépend fortement de la nature du problème et des données collectées. Dans certaines situations, la variable d'intérêt c'est-à-dire celle qui est imputée, peut prendre les valeurs zéros un grand nombre de fois. En pratique, cette situation n'est pas anodine. Par exemple, l'enquête sur les dépenses en immobilisation menée à Statistique Canada, collecte les données sur les investissements faits au Canada et pour tous les types d'industries canadiennes et comporte deux variables principales : les capitaux immobilisés pour la nouvelle construction ainsi que ceux pour la nouvelle machinerie. Pour ces deux variables, la proportion de zéros dépassait 50% en 2009. Dans cet article, nous montrons que les méthodes d'imputation usuelles peuvent aboutir à des estimateurs sévèrement biaisés. Nous proposons, basé sur un modèle de mélange trois méthodes d'imputation que nous étudions en termes de biais.

Cet article est structuré de la manière suivante : En section 2, il est définit des termes ainsi que le modèle d'imputation qui seront utilisés par la suite. En section 3, il est présenté deux méthodes d'imputation habituellement utilisées en pratique. Nous montrons que ces deux méthodes d'imputation peuvent aboutir à des estimateurs biaisés. En section 4,

[1] Christian O. Nambeu, 100 Tunney's Pasture Ottawa, Canada, K1A0T6, christianolivier.nambeu@statcan.gc.ca.
[2] David Haziza, Département de Mathématiques et de Statistique, Université de Montréal, Montréal, Canada
[3] Guillaume Chauvet, Laboratoire de Statistique d'enquête, CREST/ENSAI, Campus de Ker Lann, 35170 Bruz, France

sous les hypothèses d'un modèle de mélange, nous proposons trois méthodes d'imputation et nous montrons qu'elles produisent des estimateurs approximativement sans biais. Finalement, en section 5, il est présenté une étude par simulations dans laquelle, nous illustrons les propriétés de l'estimateur de la moyenne en termes de biais.

2. TERMINOLOGIES

On considère une population finie U de taille N. Dans cet article, on s'intéresse à l'estimation de la moyenne de la population $\overline{Y} = N^{-1}\sum_{k=1}^{N} y_k$ d'une variable d'intérêt y. À cette fin, on sélectionne un échantillon s de taille n selon un plan de sondage $p(s)$. Désignons par I_k, la variable indicatrice de sélection dans l'échantillon associée à l'unité k, tel que I_k=1, si l'unité est dans l'échantillon et zéro sinon. En présence de non-réponse, on utilise habituellement l'estimateur suivant :

$$\hat{\overline{Y}}_I = \frac{1}{N}\left(\sum_{k\in s} w_k r_k y_k + \sum_{k\in s} w_k \left(1-r_k\right) y_k^{\,*}\right), \tag{1}$$

où r_k désigne l'indicatrice de réponse associée à l'unité k tel que r_k = 1 si l'unité k a répondu à l'item y_k et 0 sinon. On désigne par w_k le poids de sondage associé à l'unité k et y_k^* la valeur imputée utilisée pour imputer y_k. Dans la suite de cet article, l'échantillon de répondants et de non répondants seront désignés respectivement par s_r et s_m.
Pour étudier les propriétés de l'estimateur en (1), on utilise la décomposition habituelle de l'erreur totale donnée par :

$$\hat{\overline{Y}}_I - \overline{Y} = \left(\hat{\overline{Y}}_\pi - \overline{Y}\right) + \left(\hat{\overline{Y}}_I - \hat{\overline{Y}}_\pi\right) \tag{2}$$

Le premier terme à droite de l'égalité (2) est appelé l'erreur due à l'échantillonnage tandis que le second terme est appelé l'erreur due à la non-réponse. Dans cet article, on se concentre sur l'erreur due à la non-réponse qui peut être exprimée comme suit

$$\hat{\overline{Y}}_I - \hat{\overline{Y}}_\pi = -N^{-1}\left[\sum_{k\in s} w_k (1-r_k)(y_k - y_k^*)\right]. \tag{3}$$

Dans le cas d'une méthode d'imputation aléatoire, l'erreur totale en (3) peut être décomposée comme suit :

$$\hat{\overline{Y}}_I - \hat{\overline{Y}}_\pi = \left(\tilde{\overline{Y}}_I - \hat{\overline{Y}}_\pi\right) - \left(\hat{\overline{Y}}_I - \tilde{\overline{Y}}_I\right), \tag{4}$$

où $\tilde{\overline{Y}}_I = E_I\left(\hat{\overline{Y}}_I \mid \mathbf{y},\mathbf{r},\mathbf{I}\right)$ dénote l'estimateur imputé qui aurait été utilisé dans le cas d'une méthode d'imputation déterministe avec $\mathbf{y} = (y_1, \ldots, y_N)'$, $\mathbf{I} = (I_1,\ldots,I_N)'$, $\mathbf{r} = (r_1,\ldots,r_N)'$ et l'indice I dénote le mécanisme d'imputation aléatoire.
Pour étudier les propriétés de l'estimateur (1), on considère deux approches d'inférence : (i) l'approche basée sur le modèle de non-réponse (NM), étudiée entre autres par Rao (1990), Rao et Sitter (1995); (ii) et l'approche basée sur le modèle d'imputation (IM), étudiée entre autres par Särndal (1992), Deville et Särndal (1994).
Sous l'approche NM, des hypothèses explicites sont postulées sur le mécanisme de non-réponse. L'inférence est basée sur la distribution conjointe induite par le plan de sondage et le mécanisme de non-réponse. Le modèle de non-réponse correspond donc à un ensemble d'hypothèses faites sur la distribution inconnue des variables indicatrices de réponse $\mathbf{r}$ qu'on appelle encore mécanisme de non-réponse. Désignons par $p_k = P(r_k = 1 \mid I_k = 1)$ la probabilité de réponse de l'unité k. On suppose de plus que les unités répondent indépendamment les unes des autres. Dans le cas où toutes les probabilités de réponse sont égales, on parle alors de non-réponse uniforme (UNM). Tandis que sous l'approche IM, des hypothèses explicites sont postulées sur le modèle d'imputation et l'inférence est basée sur la distribution conjointe induite par le plan de sondage, le mécanisme de non-réponse et le modèle d'imputation. Le modèle d'imputation est un ensemble d'hypothèses sur la distribution inconnue de $\mathbf{y}$. Dans cette approche, on suppose que la distribution des erreurs du modèle n'est pas liée à celle de $\mathbf{I}$ ni à celle de $\mathbf{r}$ une fois qu'on a conditionné sur les variables auxiliaires appropriées. En présence de zéros sur la variable y, la population U peut être vue comme un mélange de deux sous-populations : $U_0 \subset U$ de taille N_0, qui représente l'ensemble des unités de U dont les valeurs de y sont nulles et $U_1 \subset U$ de taille N_1, qui représente l'ensemble des unités de U dont les valeurs de y sont positives. Dans le cas de l'imputation par la régression, le modèle d'imputation sous-jacent est donné par :

$$m : y_k = \delta_k\left(\mathbf{z}_k'\boldsymbol{\beta} + \varepsilon_k\right), \tag{5}$$

où $\mathbf{z}$ est une matrice de variables auxiliaires de dimension $n\times q$ disponible pour tous les unités de l'échantillon, $\boldsymbol{\beta}$ un vecteur de q paramètres inconnus et δ_k une variable indiquant 1 si l'unité $k \in U_1$ et 0 sinon. Soit $\phi_k = P(\delta_k = 1)$ la

probabilité que l'unité k appartienne à U_1. De plus, on suppose que : $E(\varepsilon_k|\delta_k=1)=0$, $E(\varepsilon_k\varepsilon_l|\delta_k=1,\delta_l=1)=0$ et $V(\varepsilon_k|\delta_k=1)=\sigma^2 c_k$, où $c_k = \boldsymbol{\lambda}'\mathbf{z}_k$ avec $\boldsymbol{\lambda}$ un vecteur de constantes connues de dimension q. Dans cet article, le but ultime est de proposer une méthode d'imputation qui satisfait simultanément les trois critères suivants :

(a) l'estimateur imputé est asymptotiquement sans biais sous les approches UNM et IM;
(b) les valeurs imputées sont réalistes;
(c) l'estimateur imputé est complètement efficace.

Les méthodes d'imputation satisfaisant le critère (a) sont habituellement appelées les méthodes d'imputation doublement robustes, voir par exemple, Haziza et Rao (2006). Le critère (b) suppose qu'étant donné que la population comporte un grand nombre de zéros, les valeurs imputées devraient refléter cette situation c'est-à-dire être un mélange de zéros et de valeurs positives. Finalement, le critère (c) est satisfait lorsque l'estimateur imputé proposé n'est pas affecté par une variabilité additionnelle due à la sélection aléatoire des valeurs imputées dans le cas d'une méthode d'imputation aléatoire, voir par exemple Kim et Fuller (2004). Ce dernier critère implique que toutes les méthodes d'imputation déterministe sont efficaces.

3. MÉTHODES D'IMPUTATION USUELLES

Les deux méthodes d'imputation présentées dans cette section sont les méthodes les plus utilisées en pratique dans le contexte des populations contenant beaucoup de zéros. On étudie leur propriété en termes de biais sous les conditions du modèle d'imputation donnée en (5)

3.1 Imputation utilisant uniquement les unités avec valeur strictement positives

Dans cette section, on étudie les propriétés de l'estimateur imputé (1) lorsque les valeurs imputées sont construites uniquement sur la base des répondants positifs. Autrement dit, les unités dont la variable d'intérêt est égale à zéro, sont supprimées et une régression est ajustée sur le reste des observations, c'est-à-dire sur $s_1 = s \cap U_1$. Dans ce cas, les valeurs imputées sont données par :

$$y_k^* = \mathbf{z}_k'\hat{\mathbf{B}}_{r_1}, \tag{6}$$

où

$$\hat{\mathbf{B}}_{r_1} = \left(\sum_{i\in s_1} w_k r_k c_k^{-1}\mathbf{z}_k\mathbf{z}_k'\right)^{-1}\left(\sum_{i\in s_1} w_k r_k c_k^{-1}\mathbf{z}_k y_k\right). \tag{7}$$

On appelle cette méthode l'imputation par la Régression Déterministe Positive (**RDP**). On peut montrer que le biais conditionnel de non-réponse de $\hat{\bar{Y}}_I$ sous l'approche NM est donné par :

$$B_q(\hat{\bar{Y}}_I) \doteq -N^{-1}\sum_{i\in s} w_k(1-p_k)(y_k - \mathbf{z}_k'\hat{\mathbf{B}}_{p_1}), \tag{8}$$

où $\hat{\mathbf{B}}_{p_1}=\left(\sum_{i\in s_1} w_k p_k c_k^{-1}\mathbf{z}_k\mathbf{z}_k'\right)^{-1}\left(\sum_{i\in s_1} w_k p_k c_k^{-1}\mathbf{z}_k y_k\right)$. Le biais asymptotique en (8) n'est pas nul en général, sauf si p_k=1 pour tout k. Sous les hypothèses du modèle donné par (5), on peut montrer que le biais conditionnel de non-réponse sous l'approche IM est donné par :

$$B_{qm}(\hat{\bar{Y}}_I) = N^{-1}\sum_{i\in s} w_k(1-p_k)(1-\phi_k)\mathbf{z}_k'\boldsymbol{\beta}. \tag{9}$$

Une fois de plus, le biais (9) ne s'annule pas sauf dans les cas extrêmes où (i) p_k=1, $\forall k$, (ii) ϕ_k=1, $\forall k$. C'est-à-dire en présence de réponse complète et/ou en absence de zéros dans la population.

3.2 Imputation utilisant toutes les unités répondantes

On étudie les propriétés de l'estimateur $\hat{\bar{Y}}_I$ sous la méthode d'imputation par la Régression Déterministe (**RD**). Cette méthode d'imputation est l'une des méthodes d'imputation les plus utilisées dans les enquêtes entreprises dont la population d'intérêt contient beaucoup de zéros. Les valeurs imputées sont données par :

$$y_k^* = \mathbf{z}_k' \hat{\mathbf{B}}_r, \tag{10}$$

où $\hat{\mathbf{B}}_r = \left(\sum_{i \in s} w_k r_k c_k^{-1} \mathbf{z}_k \mathbf{z}_k' \right)^{-1} \left(\sum_{i \in s} w_k r_k c_k^{-1} \mathbf{z}_k y_k \right)$.

Sous l'approche UNM, on peut montrer que le biais de non-réponse conditionnel est nul sans égard de la relation qui existe entre les variables *y* et **z**.

Sous l'approche IM, le biais de non-réponse conditionnel de l'estimateur $\hat{\bar{Y}}_I$ est donné par :

$$B_{qm}(\hat{\bar{Y}}_I) \doteq N^{-1} \sum_{k \in s} w_k z_k' (\hat{\mathbf{T}}_p^{-1} \hat{\mathbf{T}}_{p\phi} - \phi_k \mathbf{I}_q) \boldsymbol{\beta}, \tag{11}$$

où $\hat{\mathbf{T}}_p = \sum_{k \in s} w_k p_k c_k^{-1} \mathbf{z}_k \mathbf{z}_k'$, $\hat{\mathbf{T}}_{p\phi} = \sum_{k \in s} w_k p_k \phi_k c_k^{-1} \mathbf{z}_k \mathbf{z}_k'$ et $\mathbf{I}_q$ est la matrice identité de rang *q*.

4. MÉTHODES D'IMPUTATION PROPOSÉES

Dans cette section, on propose trois méthodes d'imputation motivées par le modèle de mélange (5) et nous étudions leurs propriétés en termes de biais sous les approches d'inférence NM et IM. Par souci de simplicité, on suppose que les probabilités ϕ_k sont connues.

4.1 Méthode d'imputation déterministe

Motivé par le modèle de mélange, on propose premièrement d'utiliser les valeurs imputées suivantes :

$$y_k^* = \phi_k \mathbf{z}_k' \hat{\mathbf{B}}_{r_1}, \tag{12}$$

où $\hat{\mathbf{B}}_{r_1}$ est donné par (7).

En utilisant les valeurs imputées (12) dans l'estimateur(1), on obtient :

$$\hat{\bar{Y}}_I = N^{-1} \left[\sum_{k \in s} w_k r_k y_k + \sum_{k \in s} w_k (1 - r_k) \phi_k \mathbf{z}_k' \hat{\mathbf{B}}_{r_1} \right]. \tag{13}$$

On appelle cette méthode d'imputation, l'imputation par la Régression Déterministe-ϕ (**RD-ϕ**). En utilisant une approximation de Taylor de premier ordre, on peut montrer que l'estimateur (13) est approximativement sans biais sous l'approche UNM et sous l'approche IM. L'imputation par la régression déterministe-ϕ satisfait donc les critères (a) et (c) mais pas le critère (b) car toutes les valeurs imputées par (12) sont strictement positives.

4.2 Méthode d'imputation aléatoire

La méthode d'imputation donnée par (12) ne satisfait pas le critère (b). Pour résoudre ce problème, on propose d'utiliser la méthode d'imputation aléatoire dont les valeurs imputées sont données par l'expression suivante :

$$y_k^* = \begin{cases} \mathbf{z}_k' \hat{\mathbf{B}}_{r_1} & \text{avec probabilité } \phi_k \\ 0 & \text{avec probabilité } 1 - \phi_k, \end{cases} \tag{14}$$

où $\hat{\mathbf{B}}_{r_1}$ est donné par(7). On appelle cette méthode d'imputation l'imputation par la Régression Aléatoire-ϕ (**RA-ϕ**). En notant que $E_I(y_k^* \mid \mathbf{y}, \mathbf{I}, \mathbf{r}) = \phi_k \mathbf{z}_k' \hat{\mathbf{B}}_{r_1}$, il s'ensuit que $\hat{\bar{Y}}_I$ se réduit à (13). L'estimateur imputé donné en (1), obtenu en remplaçant les valeurs imputées par (14) est donc approximativement sans biais sous les approches UNM et IM. Le critère (a) est alors satisfait. Du point de vue de l'utilisateur des micro données, l'imputation par la régression aléatoire-ϕ comporte un attrait particulier car les valeurs imputées sont un mélange de zéros et de valeurs strictement positives. Néanmoins, elle souffre d'une variabilité additionnelle occasionnée par la sélection aléatoire des valeurs imputées. Elle ne satisfait donc par le critère (c).

4.3 Méthode d'imputation équilibrée

Suivant Chauvet, Deville et Haziza (2010), nous avons considéré une méthode d'imputation par la Régression Aléatoire Équilibrée (**RAE-ϕ**) qui consiste à sélectionner les valeurs imputées y_k^* de sorte que l'erreur due à l'imputation, $\hat{\bar{Y}}_I - \tilde{\bar{Y}}_I$ soit égale à zéro. Si l'erreur due à l'imputation est éliminée, on s'attend à ce que la variance due à l'imputation le soit aussi. Le but est donc ici de sélectionner les valeurs imputées (14) sous la contrainte :

$$\hat{\bar{Y}}_I - \tilde{\bar{Y}}_I = N^{-1}\left[\sum_{k\in s} w_k(1-r_k)(y_k^* - \phi_k \mathbf{z}_k{}'\hat{\mathbf{B}}_{r_1})\right] = 0. \tag{15}$$

Si la contrainte en (15) est satisfaite, la variance due à l'imputation sera annulée. Donc en plus des critères (a) et (b) qu'elle satisfait, cette méthode d'imputation que nous appelons imputation par la Régression Équilibrée Aléatoire-ϕ (REA-ϕ) satisfait également le critère (c).

5. ETUDE PAR SIMULATIONS

L'objectif visé par cette étude par simulations est d'évaluer numériquement la performance des méthodes d'imputation proposées en termes de biais relatif. Nous comparons le biais des estimateurs obtenus selon les cinq méthodes d'imputation présentées dans cet article. Nous avons commencé par générer une population finie de taille N=1000 consistant en une variable d'intérêt y et en une variable auxiliaire z. Les valeurs de z ont été générées selon une distribution Gamma de paramètre d'échelle α=4 et de paramètre de position β=25. Ensuite, les valeurs de la variable y ont été générées selon le modèle ratio suivant :

$$y_k = \delta_k \times (2z_k + \varepsilon_k), \tag{16}$$

où les ε_k étaient générées selon une distribution Normal de moyenne 0 et de variance σ^2, qui a été choisie de sorte que le coefficient de détermination du modèle était de 0.5. Les valeurs des δ_k ont été générées selon une distribution de Bernoulli de paramètre ϕ_k. Ces probabilités étaient générées selon deux mécanismes distincts : (a) ϕ-mécanisme 1 : le mécanisme aléatoire générant les zéros est uniforme, tous les ϕ_k sont égaux à 0.5 dans la population et (b) ϕ-mécanisme 2 : les probabilités ϕ_k sont générées selon le modèle $\log\left(\frac{\phi_k}{1-\phi_k}\right) = \lambda_0 + \lambda_1 z_k$ et les paramètres λ_0 et λ_1 sont choisis de sorte que la moyenne des ϕ_k est égale à 0.5. À partir de cette population, nous avions sélectionné R = 10 000 échantillons de taille n = 200 selon le plan d'échantillonnage aléatoire simple sans remise. Dans chacun des échantillons sélectionnés, la non-réponse à la variable y a été générée selon les trois mécanismes aléatoires décrits comme suit : (a) mécanisme-p 1 : la probabilité de réponse p_k est constante et égale à 0.7 pour toutes les unités de l'échantillon; (b) mécanisme-p 2 : la probabilité p_k associée à l'unité k est calculée à partir du modèle $\log\left(\frac{p_k}{1-p_k}\right) = \gamma_0 + \gamma_1 z_k$ où γ_0 et γ_1 sont choisis de sorte que le taux de réponse moyen soit de 0.7. Une fois que les probabilités de réponse étaient obtenues, les indicatrices de réponse r_k étaient générées selon une loi de Bernoulli de paramètre p_k. Nous étions intéressés à comparer les méthodes d'imputation suivantes : (a) L'imputation **RDP** dont les valeurs imputées sont données par (6) avec $\mathbf{z_k} = z_k$ et $c_k = z_k$; (b) l'imputation **RD** dont les valeurs imputées sont données par (10) avec $\mathbf{z_k} = z_k$ et $c_k = z_k$; (c) l'imputation **RD-ϕ** dont les valeurs imputées sont données par (12) avec $\mathbf{z_k} = z_k$ et $c_k = z_k$; (d) l'imputation **RA-ϕ** dont les valeurs imputées sont données par (14) avec $\mathbf{z_k} = z_k$ et $c_k = z_k$ (e) et l'imputation **RAE-ϕ** dont les valeurs imputées sont données par (14) avec $\mathbf{z_k} = z_k$ et $c_k = z_k$ sous la contrainte (15). Pour les méthodes d'imputation (c)-(e), les probabilités ϕ_k ont été estimées en utilisant le modèle suivant :

$$\phi_k = \exp\left(\boldsymbol{\mu}_k{}'\boldsymbol{\alpha}\right) / \exp\left(1 + \boldsymbol{\mu}_k{}'\boldsymbol{\alpha}\right). \tag{17}$$

Lorsque les zéros étaient générés selon le ϕ-mécanisme 1, on utilisait le modèle (17) avec $\boldsymbol{\mu}_k$=1 et $\boldsymbol{\alpha} = \alpha_0$ tandis que lorsqu'ils étaient générés selon le ϕ-mécanisme 2, on utilisait une fois de plus le modèle (17) mais cette fois avec $\boldsymbol{\mu}_k = (1, z_k)^{/}$ et $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)^{/}$. Finalement, nous avons calculé l'estimateur imputé (1) dans chaque échantillon. Comme mesure du biais, nous avons utilisé le biais relatif Monte Carlo donné par : $RB(\hat{\bar{Y}}_I) = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{\hat{\bar{Y}}_{I(r)} - \bar{Y}}{\bar{Y}}\right)$.

Les résultats de l'étude par simulations indiquent que l'estimateur imputé obtenu sous la méthode d'imputation RDP est biaisé dans tous les scénarios. Quant à l'imputation RD, elle menait à un estimateur imputé approximativement sans biais uniquement lorsque les probabilités ϕ et p n'étaient pas reliées. Par exemple, on peut voir du tableau 1 que le biais relatif est environ 9.56% lorsque ces probabilités dépendent de z. Dans le cas des méthodes d'imputations proposées, comme on pouvait s'y attendre, le biais est approximativement nul dans tous les scénarios.

Selon d'autres études par simulations que nous avons menées mais dont les résultats ne sont présentés dans cet article, on constatait que l'imputation aléatoire équilibrée menait à un estimateur complètement efficace dans le sens où la variabilité additionnelle créée en sélectionnant les valeurs imputées aléatoirement était complètement éliminée.

Tableau 1: Biais relatif de l'estimateur imputé par méthodes d'imputation

	ϕ–mechanism			
	ϕ(unif)		**ϕ(z)**	
***p*-mechanism**	***p*(unif)**	***p*(z)**	***p*(unif)**	***p*(z)**
RDP	31.18	19.57	16.84	18.44
RD	-0.09	-0.34	0.02	9.56
RD-ϕ	0.3	0.07	0.06	1.58
RA-ϕ	0.3	0.09	0.05	1.62
REA-ϕ	0.31	0.06	0.07	1.62

6. CONCLUSIONS

Les populations contenant un grand nombre de zéros sont fréquentes dans les enquêtes. Dans ce contexte, nous avons proposé plusieurs méthodes d'imputation motivées par un modèle de mélange. L'étude par simulations a montré que ces méthodes d'imputation performaient bien en termes de biais et d'erreur quadratique moyenne. De notre point de vue, l'imputation aléatoire est très attractive dans le sens où elle satisfait aux trois critères visés quant à la double robustesse, aux valeurs imputées réalistes et à l'efficacité des estimateurs.

RÉFÉRENCES

Chauvet, G., Deville, J.C. et Haziza, D. (2010). « On random balanced imputation in surveys ». Soumis pour publication

Deville, J.C. et Särndal, C.E. (1994). « Variance estimation for the regression imputed Horvitz-Thompson estimator ». *Journal of Official Statistics*, **23**, 33-40.

Haziza, D. et Rao, J.N.K. (2006). « A nonresponse model approah to inference under imputation for missing survey data ». *Survey Methodology*, **32**, 53-64.

Kim, J.K. et Fuller, W.A. (2004). « Fractional hot-deck imputation ». *Biometrika*, **91**, 559-578

Rao, J.N.K. (1990). « Variance estimation under imputation for missing data ».Rapport technique, Statistics Canada, Ottawa.

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

Särndal, C.E. (1992). « Methods for estimating the precision of surveys estimates when imputation has been used ». *Survey Methodology*, **18**, 241-252.