Techniques d'enquête, juin 1994
Vol. 20, n° 1, pp. 47–55
Statistique Canada

Méthode du jackknife pour l'estimation de la variance en présence de données imputées

J.G. KOVAR et E.J. CHEN[1]

RÉSUMÉ

L'imputation est une méthode dont se servent couramment les organismes d'enquête afin de corriger le problème posé par la non-réponse à des questions particulières. Bien que dans la plupart des cas, les ensembles de données ainsi complétés offrent de bonnes estimations des moyennes et des totaux, les variances correspondantes, souvent, sont largement sous-estimées. Plusieurs méthodes permettent de remédier à ce problème, mais la plupart dépendent du plan d'échantillonnage et de la méthode d'imputation. Récemment, Rao (1992) et Rao et Shao (1992) ont proposé une méthode jackknife unifiée pour l'estimation de la variance d'ensembles de données ayant fait l'objet d'une imputation. Le présent article évalue cette technique de manière empirique, au moyen d'une population réelle d'entreprises, et selon un plan d'échantillonnage aléatoire simple et un mécanisme de non-réponse uniforme. La possibilité d'étendre cette méthode à des plans d'échantillonnage stratifié à plusieurs degrés est examinée, et l'on se penche brièvement sur la performance de l'estimateur de la variance proposé dans le cas de mécanismes de réponse qui ne sont pas uniformes.

MOTS CLÉS: Non-réponse à des questions particulières; imputation hot-deck; imputation du plus proche voisin; non-réponse non aléatoire; plan d'échantillonnage complexe.

1. INTRODUCTION

Le problème de la non-réponse touche toutes les enquêtes par sondage, à des degrés divers. Dans le cas de la non-réponse totale d'une unité de l'échantillon, cette lacune est souvent corrigée par un rajustement approprié de la pondération de l'enquête, mais si la non-réponse ne touche que des questions particulières, la plupart des organismes d'enquête recourent à l'imputation. Ainsi, des valeurs plausibles sont insérées à la place de valeurs manquantes ou incohérentes, ce qui simplifie l'estimation des moyennes et des totaux à tous les niveaux d'agrégation. Dès les années 1950, toutefois, Hansen, Hurwitz et Madow (1953) ont reconnu que le fait de traiter les valeurs imputées comme des valeurs observées peut entraîner une sous-estimation des variances des estimateurs si les formules habituelles sont utilisées, sous-estimation qui s'amplifie à mesure que la proportion des réponses imputées s'accroît.

Plusieurs solutions ont été suggérées pour résoudre ce problème. En particulier, Rubin (1987) a proposé de recourir à l'imputation multiple pour estimer la variance attribuable à l'imputation, en répétant le processus plusieurs fois et en estimant la variation entre les divers ensembles de données ainsi obtenus. Plus récemment, Särndal (1990) a présenté un certain nombre d'estimateurs de la variance fondés sur un modèle, tandis que Rao et Shao (1992) ont proposé une technique de rajustement des valeurs imputées qui permet de corriger l'estimateur de la variance jackknife habituel (ou simpliste). Les méthodes de Särndal et de Rao et Shao sont attrayantes du fait que seul le fichier soumis à l'imputation (dans lequel les valeurs imputées sont marquées) est nécessaire à l'estimation de la variance. Aucun fichier auxiliaire n'est requis. La méthode de Särndal fondée sur un modèle donne des estimateurs sans biais de la variance pourvu que le modèle se vérifie (Lee, Rancourt et Särndal 1991). La méthode jackknife rajustée de Rao et Shao donne une estimation convergente selon le plan et non biaisée selon le modèle (Rao 1992). Toutefois, la méthode fondée sur un modèle exige des estimateurs de la variance différents pour chaque méthode d'imputation, tandis que la méthode jackknife rajustée offre une approche unifiée exigeant la mise en oeuvre d'un seul estimateur, l'estimateur jackknife, pourvu que les valeurs imputées soient convenablement rajustées à l'étape de l'estimation de la variance.

Dans le présent article, nous décrivons une étude de simulation qui permet d'évaluer l'estimateur de la variance jackknife rajusté de Rao et Shao (1992). Dans la section 2, nous faisons valoir le bien-fondé de la présente étude empirique en montrant les caractéristiques de l'estimateur de la variance simpliste selon quatre méthodes d'imputation, dans le cas d'un échantillonnage aléatoire simple. Dans la section 3, nous décrivons brièvement la méthode de rajustement de Rao et Shao et nous présentons les résultats empiriques. Des extensions à des plans plus complexes et à des expériences comportant des mécanismes de non-réponse non aléatoires sont examinées à la section 4. Enfin, la section 5 contient des conclusions et des recommandations, et propose notamment des voies futures de recherche.

[1] J.G. Kovar, Division des méthodes d'enquêtes-entreprises; E.J. Chen, Division des méthodes d'enquêtes sociales, Statistique Canada, Immeuble R.H. Coats, Parc Tunney, Ottawa, Ontario, K1A 0T6.

2. CONTEXTE

Suivant la notation de Rao (1992), nous supposons que dans un échantillon s de taille n, m unités répondent à la question y, tandis que $n - m$ unités ne le font pas. Désignons par y_i^* la valeur imputée pour l'unité i, $i \in s\text{-}s_r$, où s_r est l'ensemble des unités qui ont répondu. L'estimateur habituel de la moyenne $\bar{Y}$ basé sur le fichier soumis à l'imputation, dans le cas d'un échantillonnage aléatoire simple, est donné par

$$\bar{y}_I = \frac{1}{n}\left(\sum_{i \in s_r} y_i + \sum_{i \in s\text{-}s_r} y_i^* \right). \tag{1}$$

2.1 Méthodes d'imputation

Dans la présente étude de simulation, nous examinons quatre méthodes d'imputation simple: moyenne des répondants, quotient, plus proche voisin et hot-deck. Le lecteur est invité à lire l'article de Kalton et Kasprzyk (1986) pour un examen approfondi de la question de l'imputation. La méthode d'imputation la plus simple et la plus intuitive quand on veut estimer la moyenne des réponses à une question y consiste à imputer la moyenne des unités répondantes observées aux unités non répondantes. La valeur imputée y_i^* de l'unité i est donc la suivante, dans le cas de la méthode d'imputation de la moyenne:

$$y_i^* = \bar{y}_m = \sum_{j \in s_r} y_j / m. \tag{2}$$

L'estimateur de la moyenne de la population $\bar{Y}$ donné en (1) se réduit alors à l'estimateur $\bar{y}_I = \bar{y}_m$. Comme cette méthode a l'inconvénient de causer une distorsion des distributions, elle n'est généralement utilisée qu'en dernier recours. Nous la présentons ici à des fins d'illustration.

En second lieu, nous examinons une méthode d'imputation par quotient reposant sur l'hypothèse qu'une variable auxiliaire corrélée x est disponible et que le quotient $\bar{y}/\bar{x}$ est le même dans les ensembles s_r et $s - s_r$, comme ce serait le cas si la non-réponse survenait au hasard, par exemple. Selon cette méthode du quotient, nous imputons la valeur suivante à la place de la valeur manquante y_i:

$$y_i^* = \frac{\bar{y}_m}{\bar{x}_m} x_i, \tag{3}$$

où $\bar{x}_m$ est la moyenne des valeurs x de l'ensemble des répondants s_r. L'estimateur de la moyenne de la population $\bar{Y}$ donné en (1) se réduit à l'estimateur d'échantillonnage double $\bar{y}_I = (\bar{y}_m/\bar{x}_m)\bar{x}$, si on considère les répondants comme l'échantillon de deuxième phase.

La troisième méthode d'imputation examinée est celle du plus proche voisin (PPV). Cette méthode consiste à imputer à une donnée manquante la valeur observée d'une autre unité de l'ensemble s_r, dont la distance par rapport à l'unité non répondante est minimale. En pratique, les fonctions de distance utilisées sont habituellement les normes ζ_1, ζ_2, ou ζ_∞ de Minkowski fondées sur les variables auxiliaires x, qu'on suppose observées pour toutes les unités de s. Ainsi

$$y_i^* = y_j, j \in s_r, \quad \text{tel que} \quad \| x_i - x_j \| \quad \text{est minimisé,} \tag{4}$$

où $\| \cdot \|$ désigne l'une des normes ci-dessus.

Les trois méthodes qui viennent d'être décrites sont souvent qualifiées de déterministes, car pour un échantillon de répondants donné, les valeurs imputées sont déterminées de façon unique. La quatrième méthode d'imputation examinée dans cette étude, la méthode du hot-deck (HD), est non déterministe, puisque les valeurs imputées sont choisies au hasard dans l'échantillon de répondants. Bien qu'en pratique des classes d'imputation soient souvent créées et qu'un processus séquentiel quelconque soit généralement mis en oeuvre, nous examinons ici le hot-deck pur, dans lequel le donneur (j) est choisi au hasard, avec remplacement, dans l'ensemble s_r complet, c.-à-d.

$$y_i^* = y_j, \; j \in s_r. \tag{5}$$

2.2 Variance due à l'imputation

Si l'on traite les valeurs imputées comme des valeurs observées, on a l'estimateur incorrect de la variance

$$v_{naive} = (1 - f)s_I^2/n, \tag{6}$$

où s_I^2 est la variance de l'échantillon complet (valeurs des répondants et valeurs imputées) et $(1 - f)$ est le facteur de correction pour population finie $(f = n/N)$. Il est facile de montrer que la variance vraie de l'estimateur $\bar{y}_I$ dans (1), $V(\bar{y}_I)$, peut s'écrire ainsi (Särndal 1990):

$$V(\bar{y}_I) = V_{sam} + V_{imp} + V_{mix}, \tag{7}$$

où V_{sam} est la variance d'échantillonnage, V_{imp} est la variance introduite par la méthode d'imputation en cause et V_{mix} est un terme de covariance entre V_{sam} et V_{imp} qui, dans la plupart des cas, est négligeable ou nul. On pourrait obtenir une estimation de V_{sam} en ajoutant à v_{naive} un terme de correction tenant compte du fait que la formule habituelle sous-estime la variance d'échantillonnage quand l'ensemble de données contient des valeurs imputées. Pour estimer $V(\bar{y}_I)$, toutefois, il est nécessaire d'estimer une composante de variance additionnelle, V_{imp}, attribuable au mécanisme d'imputation. Cela peut se faire explicitement, comme dans l'imputation multiple de Rubin (1987), ou encore en modifiant des formules de variance commune comme dans Särndal (1990) et dans Rao et Shao (1992). Notons que l'intérêt réside dans l'estimation de la variance de l'estimateur en cause, soit $V(\bar{y}_I)$, et non de la variance d'un estimateur qui aurait été obtenu en l'absence de non-réponse.

2.3 Sous-estimation de la variance

Pour montrer l'ampleur de la sous-estimation de $V(\bar{y}_I)$ par v_{naive}, et le lien entre le degré de sous-estimation et la méthode d'imputation, nous décrivons d'abord l'étude de simulation que nous avons réalisée à cette fin. Nous examinons un ensemble de données de 5,620 unités auxquelles sont associées deux variables: une variable auxiliaire x, le revenu d'exploitation brut, connu pour toutes les unités et pouvant servir de mesure de la taille, et une variable connexe relative aux achats, y. La corrélation entre x et y dans cet ensemble de données particulier est d'environ 0.92. Des échantillons aléatoires simples de taille 200 étaient prélevés sans remplacement. Une proportion fixe d'unités étaient désignées au hasard comme des non-répondants; leur valeur y était supprimée, puis imputée selon l'une des quatre méthodes décrites ci-dessus. Divers taux de non-réponse ont été produits, mais pour l'essentiel, nous ne présentons que les résultats obtenus avec les taux de non-réponse de 5% et de 30%.

Pour évaluer la performance des estimateurs de la variance proposés, nous calculons le biais relatif en pourcentage de l'estimateur de la variance $v.$, donné par

$$\text{Biais rel.}(v.) = \sum_{k=1}^{K} \frac{(v_k - V(\bar{y}_I))/K}{V(\bar{y}_I)} \times 100, \qquad (8)$$

où $V(\bar{y}_I)$ est obtenue par simulation, et v_k est la $k$ième réalisation des K estimations de la variance simulées en question. De même, la stabilité relative en pourcentage des estimateurs de la variance est donnée par

$$\text{Stab. rel.}(v.) = \sum_{k=1}^{K} \frac{\sqrt{(v_k - V(\bar{y}_I))^2/K}}{V(\bar{y}_I)} \times 100. \qquad (9)$$

Toutes les simulations ont été faites sur un PC d'IBM avec le logiciel FORTRAN 77, version 5.0, de Microsoft. Dans le cas de l'échantillonnage aléatoire simple, les résultats se fondent sur des moyennes de 100,000 répétitions (K = 100,000). Avec un tel nombre de répétitions, nous avons observé que les biais relatifs présentés ne variaient pas de plus d'un point de pourcentage. Les résultats sont résumés ci-dessous au tableau 1 pour les taux de non-réponse de 5% et de 30%.

L'examen du tableau 1 nous indique d'abord que l'estimateur simpliste sous-estime la variance vraie de $\bar{y}_I$ de 10.7% dans le cas de l'imputation de la moyenne, à un niveau de non-réponse de 5%. Cette sous-estimation est attribuable environ pour moitié au fait que v_{naive} sous-estime V_{sam}, et pour le reste au fait que v_{naive} ne tient pas compte de la composante V_{imp}. Särndal (1990) obtient des résultats très semblables pour ce qui est de la ventilation de la sous-estimation dans le cas de l'imputation de la moyenne. Nous observons ensuite, dans la première ligne du tableau 1, que la variance vraie de $\bar{y}_I$ est plus élevée dans le cas de l'imputation hot-deck que dans celui de l'imputation de la moyenne, en raison de la variabilité

Tableau 1

Sous-estimation de la variance de $\bar{y}_I$ par l'estimateur simpliste (v_{naive}) selon quatre méthodes d'imputation, à des taux de non-réponse de 5% et de 30%

Taux de non-réponse	Estimateur de la variance	Méthode d'imputation			
		Moyenne	HD	Quotient	PPV
5%	$V(\bar{y}_I)$	9.9	10.3	9.5	9.5
	v_{naive}	8.9	9.4	9.2	9.3
	Biais rel. (v_{naive})	−10.7%	−9.4%	−2.5%	−2.2%
30%	$V(\bar{y}_I)$	13.5	16.5	10.1	10.3
	v_{naive}	6.5	9.4	8.5	9.0
	Biais rel. (v_{naive})	−51.4%	−43.4%	−15.3%	−12.8%

inhérente au hot-deck (c.-à-d. que la composante V_{imp} est plus grande). Par contre, $V(\bar{y}_I)$ est légèrement plus faible dans le cas des méthodes du quotient et du plus proche voisin, puisque V_{imp} diminue avec la capacité de la méthode d'imputation d'estimer les valeurs manquantes vraies (Särndal 1990), comme c'est le cas dans la présente étude en raison de la corrélation relativement élevée entre les variables x et y. Un autre fait que révèle le tableau 1, c'est que $V(\bar{y}_I)$ augmente, tandis que v_{naive} diminue, avec l'élévation du taux de non-réponse. Ainsi, la sous-estimation de $V(\bar{y}_I)$, quand les valeurs imputées sont traitées comme des valeurs observées, s'amplifie à mesure que s'accroît la proportion de valeurs manquantes. Le problème est plus prononcé dans le cas des méthodes d'imputation de la moyenne et du hot-deck, qui n'utilisent pas d'information auxiliaire. Signalons qu'une sous-estimation de la variance de l'ordre de 50% comme celle observée ici peut donner des intervalles de confiance trop étroits d'environ 30% et amener à déclarer comme significatifs des résultats qui ne le sont pas. Il est intéressant, par ailleurs, de noter le comportement semblable des méthodes du quotient et du plus proche voisin, caractéristique dont nous nous servirons plus loin.

3. ESTIMATEUR DE LA VARIANCE JACKKNIFE

Soit $\bar{y}_I(j)$ l'estimateur imputé de $\bar{Y}$ obtenu lorsque la $j$ième unité est retirée de l'échantillon. Dans le cas de l'échantillonnage aléatoire simple, un estimateur de la variance jackknife simpliste de $\bar{y}_I$ est alors donné par

$$\bar{v}_J = \frac{n-1}{n} \sum_{j=1}^{n} [\bar{y}_I(j) - \bar{y}_I]^2, \qquad (10)$$

qui se réduit à v_{naive}, comme cela a été démontré (Rao 1992).

3.1 Rajustement des valeurs imputées

Afin de produire la version "appropriée" (Rao 1990) de l'estimateur de la variance jackknife, Rao (1992) a proposé de rajuster les valeurs imputées de la façon décrite ci-dessous. Intuitivement, un tel rajustement est nécessaire dès qu'une unité répondante est retirée d'une répétition jackknife, car pour la plupart des méthodes d'imputation, toutes les valeurs imputées dépendent directement ou indirectement de la valeur observée qui a été supprimée. Cela est clair dans le cas de l'imputation de la moyenne et de l'imputation par quotient, car tous les répondants contribuent directement à la moyenne $\bar{y}_m$, mais est moins évident pour les méthodes du plus proche voisin et du hot-deck, dans lesquelles l'unité supprimée contribue au processus d'imputation uniquement en ce sens qu'elle n'est pas disponible pour être choisie comme donneur. Ainsi, dès qu'une unité répondante est supprimée, *toutes* les valeurs imputées de l'échantillon doivent être rajustées avant que l'estimateur imputé "à une suppression" de la moyenne soit calculé. Le rajustement, de toute évidence, doit être fonction de la méthode d'imputation utilisée. Dans le cas des méthodes d'imputation de la moyenne et du hot-deck, on peut montrer que le rajustement suivant est approprié (Rao 1992; Rao et Shao 1992). Soit $z_i^*(j)$ la valeur rajustée de la $i$ième unité imputée y_i^*, quand la $j$ième unité a été supprimée. Alors, $z_i^*(j)$ est donnée par

$$z_i^*(j) = \begin{cases} y_i^* + [\bar{y}_m(j) - \bar{y}_m] & \text{si } j \in s_r \\ y_i^* & \text{si } j \in s\text{-}s_r. \end{cases} \quad (11)$$

Autrement dit, aucun rajustement n'est nécessaire si l'unité supprimée (j) a elle-même été imputée, c'est-à-dire si l'unité j est un non-répondant. Dans le cas de l'imputation de la moyenne, par exemple, quand $j \in s_r$, la valeur rajustée se réduit à $\bar{y}_m(j)$, la moyenne des $m - 1$ répondants restants, comme souhaité.

On évalue l'estimateur de la variance jackknife d'abord en calculant l'estimateur imputé rajusté $\bar{y}_I^a(j)$ suivant

$$\bar{y}_I^a(j) = \sum_{\substack{i \in s \\ i \neq j}} z_i^*(j)/(n-1), \quad (12)$$

puis en posant

$$v_J(\bar{y}_I) = \frac{n-1}{n} \sum_{j=1}^{n} [\bar{y}_I^a(j) - \bar{y}_I]^2. \quad (13)$$

Il peut être démontré que l'estimateur de la variance jackknife rajusté se réduit à l'estimateur de la variance approprié dans le cas de l'imputation de la moyenne (Rao 1990) et offre une estimation convergente dans le cas de l'imputation hot-deck (Rao et Shao 1992).

Pour ce qui est de l'imputation par quotient, les valeurs rajustées sont données par

$$z_i^*(j) = \begin{cases} y_i^* + \left[\dfrac{\bar{y}_m(j)}{\bar{x}_m(j)} x_i - \dfrac{\bar{y}_m}{\bar{x}_m} x_i\right] & \text{si } j \in s_r \\ y_i^* & \text{si } j \in s\text{-}s_r, \end{cases} \quad (14)$$

où $\bar{x}_m(j)$ est la moyenne des $m - 1$ valeurs de x pour les unités répondantes lorsque l'unité j est supprimée. L'estimateur de la variance jackknife $v_J(\bar{y}_I)$ est alors calculé comme en (13) ci-dessus, ce qui donne l'estimateur de la variance approprié. En outre, Rao (1992) montre non seulement que l'estimateur de la variance jackknife rajusté est convergent selon le plan (p-convergent) dans le cas d'une non-réponse uniforme et indépendamment du modèle, mais aussi qu'il est non biaisé selon le plan et le modèle (pm-non biaisé), en vertu du modèle (15) et de tout mécanisme de non-réponse qui ne dépend pas des valeurs y.

$$E_m(y_i) = \beta x_i, \quad V_m(y_i) = \sigma^2 x_i, \quad \operatorname{cov}_m(y_i, y_j) = 0 \quad i \neq j \in s. \quad (15)$$

Puisque l'estimateur de la variance simpliste en vertu de l'imputation du plus proche voisin a affiché un comportement très semblable à celui de l'estimateur de la variance simpliste en vertu de l'imputation par quotient, le rajustement relatif à l'imputation par quotient donné en (14) a été appliqué à l'imputation du plus proche voisin. Nous avons aussi examiné un autre rajustement, qui consistait à faire une nouvelle imputation de l'unité i au moyen de la méthode du plus proche voisin, lorsque l'unité supprimée (j) avait servi à l'imputation de l'unité i. Autrement dit, le rajustement a lieu seulement si l'unité supprimée est un répondant (comme ci-dessus), et seulement les non-répondants de la $j$ième répétition jackknife à qui l'unité j a été imputée font l'objet d'une nouvelle imputation pour recevoir la valeur de l'un des $m - 1$ donneurs restants. (Cela équivaut à l'imputation du deuxième voisin le plus proche pour ces unités.) Il convient de signaler qu'il n'existe pas de justification théorique pour l'un ou l'autre de ces rajustements. Puisque le dernier rajustement a donné un rendement inférieur à celui du rajustement de la méthode du quotient dans nos exemples, et qu'il serait lourd à appliquer en pratique, nous ne l'avons pas analysé davantage, bien qu'il ait toujours produit des estimations prudentes.

Il importe de souligner que pour toutes les méthodes d'imputation, les rajustements ne servent qu'à l'estimation de la variance et peuvent par conséquent n'être appliqués que temporairement, au moment de l'exécution du programme d'estimation de la variance. Aucun rajustement permanent du fichier imputé n'est nécessaire pour l'estimation des moyennes et des totaux, bien qu'il faille marquer de façon appropriée les champs ayant reçu une valeur imputée.

3.2 Résultats empiriques

L'estimateur de la variance jackknife, avec rajustements correspondant aux quatre méthodes d'estimation décrites ci-dessus, a été calculé en sus de v_{naive} dans l'étude de simulation décrite à la section 2. Des taux de non-réponse

Tableau 2

Biais relatifs de l'estimateur de la variance simpliste (v_{naive}) et de l'estimateur de la variance jackknife rajusté, à des taux de non-réponse de 5% et de 30%

Taux de non-réponse	Estimateur de la variance	Méthode d'imputation			
		Moyenne	HD	Quotient	PPV
		en pourcentage			
5%	v_{naive}	−10.7	−9.4	−2.5	−2.2
	v_J	2.7	3.6	3.4	3.7
30%	v_{naive}	−51.4	−43.4	−15.3	−12.8
	v_J	3.3	1.9	3.0	5.3

de 5% et de 30% ont été utilisés et les biais relatifs ont été calculés. Les résultats sont résumés dans le tableau 2.

Puisque l'estimateur de la variance jackknife rajusté est convergent selon le plan (p-convergent) (Rao 1992), il donne de bons résultats, comme prévu, dans le cas de l'imputation de la moyenne, de l'imputation hot-deck et de l'imputation par quotient pour une structure de réponse uniforme. (Nous avons observé d'aussi bons résultats avec d'autres ensembles de données qui ne suivent pas le modèle (15), mais il faudrait approfondir davantage l'analyse dans leur cas.) Il convient de signaler la performance relativement bonne obtenue dans le cas de l'imputation du plus proche voisin. L'estimateur proposé semble être quelque peu prudent, ce qui s'explique, pour une faible part, par le fait qu'il n'intègre pas la correction pour population finie.

4. EXTENSIONS

Bien que l'estimateur de la variance jackknife rajusté ait donné de bons résultats dans le cas de l'échantillonnage aléatoire simple, pour un mécanisme de non-réponse uniforme et une seule classe d'imputation, nous examinons ici des extensions possibles à un plan plus complexe, à plus d'une classe d'imputation et à des mécanismes de réponse non aléatoires.

4.1 Plans complexes

Dans la présente section, nous décrivons une étude de simulation qui évalue l'estimateur de la variance jackknife rajusté de Rao et Shao (1992) par rapport à l'estimateur de la variance simpliste, dans le cas d'un échantillonnage stratifié à plusieurs degrés et d'une imputation hot-deck. Plus précisément, nous utiliserons des données de l'enquête canadienne sur les finances des consommateurs (EFC), dont le plan de sondage est identique à celui de l'enquête sur la population active. La variable à l'étude, y, est le revenu total du ménage. L'EFC se fonde sur un plan complexe à plusieurs degrés avec stratification, et les unités primaires d'échantillonnage (upé) des strates utilisées dans la présente étude sont sélectionnées avec probabilité proportionnelle au nombre de logements. De façon générale, les upé sont des ensembles de logements, plus précisément des îlots urbains dans les villes et des groupes de secteurs de dénombrement (SD) du recensement dans les régions rurales. Nous avons pris comme population un échantillon de 3,870 ménages appartenant à 30 strates, et nous avons prélevé deux upé dans chaque strate. Comme dans notre étude sur l'échantillonnage aléatoire simple, des taux de non-réponse uniformes de 5% et de 30% ont été simulés, au niveau des ménages. Les valeurs manquantes ont ensuite fait l'objet d'une imputation selon la méthode du hot-deck décrite dans Rao et Shao (1992). En bref, la méthode d'imputation consiste à sélectionner les donneurs dans l'ensemble des répondants, avec remplacement, et avec probabilité proportionnelle au poids de l'enquête des donneurs.

Nous examinons d'abord le cas où il n'y a qu'une seule classe d'imputation. Soit y_{hik} la valeur observée pour la $k$ième unité de la $i$ième upé et de la $h$ième strate ($k = 1, \ldots, n_{hi}, i = 1, \ldots, n_h, h = 1, \ldots, L, n = \sum\sum n_{hi}$) et soit y^*_{hik} la valeur imputée correspondante quand l'unité (hik) est un non-répondant, c'est-à-dire quand $(hik) \in s\text{-}s_r$. L'estimateur imputé de Y est alors donné par

$$\hat{Y}_I = \sum_{(hik)\in s_r} w_{hik} y_{hik} + \sum_{(hik)\in s\text{-}s_r} w_{hik} y^*_{hik}, \qquad (16)$$

où w_{hik} est le poids de l'enquête correspondant à l'unité (hik). En vertu de la méthode d'imputation hot-deck ci-dessus, $\hat{Y}_I$ est asymptotiquement sans biais (Rao et Shao 1992).

L'espérance de $\hat{Y}_I$ en vertu de la méthode du hot-deck peut s'écrire ainsi (Rao et Shao 1992):

$$E_*(\hat{Y}_I) = \left[\sum_{(hik)\in s_r} w_{hik} y_{hik} \Big/ \sum_{(hik)\in s_r} w_{hik}\right] \times \sum_{(hik)\in s} w_{hik}$$

$$= [\hat{S}/\hat{T}] \times \hat{U}, \qquad (17)$$

notation introduisant les termes $\hat{S}$, $\hat{T}$ et $\hat{U}$. Les valeurs jackknife "à une suppression" sont alors données par

$$\hat{S}(gj) = \sum_{\substack{(hik)\in s_r \\ h\neq g}} w_{hik} y_{hik} + \frac{n_g}{n_g - 1} \sum_{\substack{(gik)\in s_r \\ i\neq j}} w_{gik} y_{gik},$$

$$\hat{T}(gj) = \sum_{\substack{(hik)\in s_r \\ h\neq g}} w_{hik} + \frac{n_g}{n_g - 1} \sum_{\substack{(gik)\in s_r \\ i\neq j}} w_{gik}, \qquad (18)$$

lorsque la $j$ième upé de la $g$ième strate est supprimée. On effectue le rajustement des valeurs imputées quand la (gj)ième upé est supprimée, $(hi) \neq (gj)$, et $(hik) \in s\text{-}s_r$, en posant

$$z^{(gj)}_{hik} = y^*_{hik} + \left[\frac{\hat{S}(gj)}{\hat{T}(gj)} - \frac{\hat{S}}{\hat{T}}\right]. \qquad (19)$$

Ensuite, comme dans les équations (12) et (13), on évalue l'estimateur de la variance jackknife en calculant d'abord, de la manière suivante, l'estimateur imputé rajusté $\hat{Y}^a_I$ quand la (gj)ième upé est supprimée:

$$\hat{Y}_I^a(gj) = \hat{S}(gj) + \sum_{(hik)\in s-s_r} w_{hik}\, z_{hik}^{(gj)} + \frac{n_g}{n_g - 1} \sum_{\substack{(hik)\in s-s_r \\ i\neq j}} w_{gik}\, z_{gik}^{(gj)}, \quad (20)$$

puis en posant

$$v_J(\hat{Y}_I) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_I^a(gj) - \hat{Y}_I)^2. \quad (21)$$

Il peut être démontré que l'estimateur v_J défini en (21) est un estimateur convergent de la variance de $\hat{Y}_I$ (Rao et Shao 1992).

Nous avons produit 10,000 échantillons de 60 upé choisies avec probabilité proportionnelle à la taille, et nous avons soumis les ménages sélectionnés à des taux de non-réponse uniformes de 5% et de 30%. Nous avons ensuite calculé l'estimateur de la variance simpliste, ainsi que l'estimateur de la variance jackknife rajusté, v_J, donné en (21). Le biais relatif (8) et la stabilité relative (9) ont été calculés pour les deux estimateurs de la variance, et les résultats sont résumés au tableau 3.

Tableau 3

Biais relatif et stabilité relative (entre parenthèses) de l'estimateur de la variance simpliste (v_{naive}) et de l'estimateur de la variance jackknife rajusté, à des taux de non-réponse de 5% et de 30%, dans le cas d'un échantillonnage stratifié à plusieurs degrés

Estimateur de la variance	Taux de non-réponse	
	5%	30%
	en pourcentage	
v_{naive}	−10.3 (88)	−43.7 (84)
v_J	−0.9 (97)	1.2 (124)

Comme en fait foi le tableau 3, l'estimateur de la variance simpliste sous-estime la variance vraie de Y en des proportions comparables à celles observées dans le cas de l'échantillonnage aléatoire simple (tableau 2), et la sous-estimation s'aggrave avec l'augmentation du taux de non-réponse. L'estimateur de la variance jackknife rajusté, par contre, donne de bons résultats aux deux niveaux de non-réponse, au prix relativement modique d'une légère diminution de sa stabilité, en comparaison de v_{naive}.

4.2 Classes d'imputation

Selon le même plan d'échantillonnage que celui décrit à la section 4.1, nous avons aussi examiné le cas où il existe plus d'une classe d'imputation, qui reflète la situation réelle. La taille du ménage, connue pour tous les ménages de l'échantillon, a été utilisée pour former deux classes d'imputation, soit les ménages à un seul membre et les ménages à plus d'un membre. L'hypothèse sous-jacente était que la propension à répondre est différente entre ces deux classes; on a supposé, en revanche, une probabilité de réponse uniforme à l'intérieur des classes d'imputation. Deux structures de non-réponse ont été évaluées. La première suppose une non-réponse uniforme de 5% dans la classe des ménages à un seul membre et une non-réponse uniforme de 10% dans la classe des ménages à plusieurs membres, tandis que la deuxième suppose des taux de non-réponse de 25% et de 30% respectivement pour ces deux classes. L'imputation hot-deck, les rajustements des valeurs imputées et les calculs des totaux rajustés donnés en (20), $\hat{Y}_{Iv}^a(gj)$, ont été effectués indépendamment dans chaque classe d'imputation, désignée par v. Les termes $\hat{Y}_{Iv}^a(gj)$ ont ensuite été additionnés pour les deux classes d'imputation, ce qui a donné $\hat{Y}_I^a(gj)$, dont on s'est servi pour obtenir l'estimation v_J, selon l'équation (21). Les résultats sont résumés au tableau 4.

Tableau 4

Biais relatif et stabilité relative (entre parenthèses) de l'estimateur de la variance simpliste (v_{naive}) et de l'estimateur de la variance jackknife rajusté, selon deux structures de non-réponse, dans le cas d'un échantillonnage stratifié à plusieurs degrés et de deux classes d'imputation

Estimateur de la variance	Taux de non-réponse	
	5% et 10%	25% et 30%
	en pourcentage	
v_{naive}	−16.7 (87)	−40.2 (84)
v_J	−1.0 (103)	1.1 (127)

Comme le montre le tableau 4, l'estimateur de la variance jackknife rajusté v_J donne de bons résultats en vertu des deux structures de non-réponse. Ces résultats, jumelés à ceux du tableau 3, démontrent la convergence et la stabilité relativement bonne de l'estimateur de la variance jackknife rajusté, même dans le cas de taux de non-réponse élevés.

4.3 Non-réponse non aléatoire

Comme nous l'avons vu ci-dessus, l'estimateur de la variance jackknife rajusté donne de bons résultats quand la non-réponse est aléatoire à l'intérieur des classes d'imputation. Pour étudier sa robustesse à l'égard de l'hypothèse d'un mécanisme de réponse uniforme, nous avons utilisé l'ensemble de données décrit à la section 2, et produit une non-réponse de la manière décrite dans Lee, Rancourt et Särndal (1991). Plus précisément, nous avons supposé que la probabilité de non-réponse était reliée à la variable x de deux façons distinctes:

$$P_L = 1 - \exp(-c_L x), \tag{22}$$

$$P_S = \exp(-c_S x), \tag{23}$$

où les constantes c_L et c_S sont choisies de façon à donner un taux de non-réponse probable de 30%. Dans le modèle P_L donné en (22), la non-réponse est en corrélation positive avec la variable x, ce qui signifie qu'une non-réponse est plus probable dans le cas des grandes unités (L). L'inverse est vrai pour le modèle P_S donné en (23), en vertu duquel ce sont les petites unités (S) qui sont le plus susceptibles de ne pas répondre. Les méthodes d'imputation qui ne tiennent pas compte de la variable x (moyenne et hot-deck) devraient normalement produire des estimateurs de $\bar{Y}$ qui sous-estiment la moyenne vraie en vertu du modèle de non-réponse (22) et qui surestiment la moyenne vraie en vertu du modèle (23). Toutefois, les méthodes d'imputation qui utilisent la variable auxiliaire (quotient et plus proche voisin) devraient produire de meilleures estimations de la moyenne. Une simulation, dont les résultats sont présentés au tableau 5 ci-dessous, a permis de confirmer ces suppositions. Comme auparavant, nous avons utilisé 100,000 répétitions.

Tableau 5

Estimations de la moyenne $\hat{Y}$ en pourcentage de la moyenne vraie lorsque la non-réponse n'est pas aléatoire et que le taux de non-réponse probable est de 30%

Modèle de non-réponse	Méthode d'imputation			
	Moyenne	HD	Quotient	PPV
	en pourcentage			
P_L	60.4	60.4	94.7	93.5
P_S	132.7	132.7	102.0	101.4

De toute évidence, l'estimation de la variance n'offre aucun intérêt lorsque les estimateurs ponctuels eux-mêmes sont fortement biaisés, comme c'est le cas pour les méthodes de la moyenne et du hot-deck. Toutefois, dans le cas des méthodes du quotient et du plus proche voisin, pour lesquelles les estimateurs ponctuels sont supérieurs, nous avons examiné la performance de l'estimateur de la variance jackknife rajusté, ainsi que d'un estimateur proposé par Särndal (1990), qui peut s'écrire (Rao 1992):

$$v_S(\bar{y}_I) = \left(\frac{\bar{x}}{\bar{x}_m}\right)^2 \frac{1}{m(m-1)} \sum_{i \in s_r} \left(y_i - \frac{\bar{y}_m}{\bar{x}_m} x_i\right)^2$$

$$+ \left(\frac{\bar{y}_m}{\bar{x}_m}\right) \frac{2m}{n^2(m-1)} \sum_{i \in s_r} \left(y_i - \frac{\bar{y}_m}{\bar{x}_m}\right) x_i \tag{24}$$

$$+ \left(\frac{\bar{y}_m}{\bar{x}_m}\right)^2 \frac{1}{n(n-1)} \sum_{i \in s} (x_i - \bar{x})^2,$$

pourvu que le facteur de correction pour population finie soit omis et que $(n-1)/n \cong 1$ et $(m-1)/m \cong 1$. Les résultats sont résumés au tableau 6.

Tableau 6

Biais relatif de l'estimateur de la variance simpliste (v_{naive}), de l'estimateur de la variance jackknife rajusté et de l'estimateur de la variance de Särndal en vertu d'une non-réponse non aléatoire de 30%

Modèle de non-réponse	Estimateur de la variance	Méthode d'imputation	
		Quotient	PPV
		en pourcentage	
P_L	v_{naive}	−22.7	−54.6
	v_J	3.9	−37.5
	v_S	−2.6	−36.8
P_S	v_{naive}	−4.0	−0.7
	v_J	3.7	7.2
	v_S	2.8	4.5

Dans le cas de l'imputation par quotient, l'estimateur de la variance simpliste donne des résultats très différents pour les deux modèles de non-réponse (−22.7% contre −4.0%). En effet, s'il est vrai que la réduction de la taille effective de l'échantillon tend à réduire la variance dans les deux cas, il y a surreprésentation des unités de grande taille parmi les unités manquantes dans le modèle P_L, ce qui tend à accentuer cet effet, alors que dans le modèle P_S, où il y a surreprésentation des petites unités parmi les unités manquantes, cet effet tend à être partiellement compensé. Deuxième observation, l'estimateur de la variance jackknife rajusté donne de bons résultats dans le cas de l'imputation par quotient, mais est plutôt décevant dans le cas de l'imputation du plus proche voisin. Cela s'explique par le fait que le présent ensemble de données suit assez bien le modèle linéaire habituel (15) et que l'estimateur de la variance jackknife rajusté s'est révélé non biaisé selon le modèle (Rao 1992) pour l'imputation par quotient. Par ailleurs, le rajustement de la méthode du quotient n'apporte rien d'intéressant à l'imputation du plus proche voisin quand la non-réponse n'est pas uniforme. L'autre rajustement applicable à l'imputation du plus proche voisin, que nous avons décrit à la section 3, donne également de piètres résultats en termes absolus (valeurs non présentées ici), bien que les estimations soient toujours prudentes. Troisième observation, la performance de l'estimateur de Särndal, v_S, équivaut grosso modo à celle de l'estimateur jackknife rajusté, en vertu aussi bien de la méthode du quotient que de celle du plus proche voisin, et d'une non-réponse non aléatoire qui dépend seulement de x.

Lorsque le mécanisme de réponse n'est pas aléatoire, et que la propension à répondre est reliée à la variable touchée par la non-réponse (y), les estimateurs ponctuels sont eux-mêmes gravement biaisés selon les quatre méthodes d'imputation. L'estimation de la variance a donc peu à

offrir, et le véritable intérêt réside dans l'estimation de l'erreur quadratique moyenne. Autrement dit, des efforts accrus doivent être axés sur l'amélioration des estimations ponctuelles et de leurs biais. Des résultats préliminaires dans ce domaine ont été présentés par Rancourt, Lee et Särndal (1992).

5. CONCLUSION

Il est bien connu que l'estimateur de la variance habituel sous-estime la variance de l'estimation de $\bar{Y}$ en présence de valeurs imputées, si ces valeurs sont traitées comme des valeurs observées. Dans la présente étude, nous avons de nouveau mis en évidence l'importante sous-estimation que produit l'estimateur de la variance simpliste lorsque des données sont imputées. Nous avons examiné plusieurs méthodes d'imputation pour évaluer dans quelle mesure le degré de sous-estimation était lié à la méthode d'imputation. Nous avons évalué un estimateur de la variance jackknife unifié de la forme proposée par Rao et Shao (1992), lequel tient compte de la variance due au processus d'imputation. L'étude a révélé certaines propriétés intéressantes de l'estimateur proposé dans le cas aussi bien de l'échantillonnage aléatoire simple que de plans d'enquête complexes. Nos observations sont résumées dans les paragraphes qui suivent.

(1) L'ampleur de la sous-estimation de la variance est fortement liée à la fois à la capacité de la méthode d'imputation de prédire les valeurs vraies, et à sa capacité de préserver la variation naturelle des données.

(2) L'estimateur de la variance jackknife rajusté qui est proposé offre une approche unifiée pour l'estimation de la variance de données imputées, qui peut être appliquée facilement à un certain nombre de méthodes d'imputation et à des plans de complexité variable.

(3) D'un point de vue pratique, aucune modification du fichier imputé initial n'est nécessaire, de sorte que l'estimation des moyennes et des totaux n'est aucunement perturbée par la nécessité d'estimer les variances.

(4) La méthode proposée peut facilement être étendue à des plans plus complexes, à plusieurs classes d'imputation et, avec prudence, au cas d'une non-réponse non aléatoire qui dépend seulement des variables auxiliaires disponibles.

(5) L'estimateur de la variance jackknife rajusté donne de bons résultats lorsque la non-réponse est uniforme ou que le modèle linéaire habituel est vérifié, ce qui découle du fait que l'estimateur est à la fois convergent selon le plan et non biaisé selon le plan et le modèle.

(6) Dans le cas du modèle P_L, dans lequel les unités ayant des valeurs y élevées sont plus susceptibles d'être non répondantes, les trois estimateurs de la variance ont un rendement extrêmement faible.

(7) Dans le cas de la non-réponse liée à la variable y, de meilleures techniques d'imputation sont nécessaires, et les estimateurs ponctuels doivent être étudiés plus à fond. Ici, il faut plutôt s'intéresser à l'estimation de l'erreur quadratique moyenne qu'à celle de la variance.

Les enquêtes actuelles étant soumises à un degré élevé d'imputation, du moins dans certaines classes d'imputation, il est clair qu'on ne peut passer sous silence l'effet de l'imputation sur l'estimation de la variance. Une surestimation de la précision peut produire des intervalles de confiance trop étroits et amener à présenter comme significatives des données qui ne le sont pas. Si la mise en oeuvre des méthodes suggérées ci-dessus est jugée trop coûteuse dans une situation particulière, il faudrait à tout le moins effectuer des études pour évaluer l'impact de l'imputation dans certains cas représentatifs. Un facteur spécial d'accroissement de la variance pourrait alors être appliqué. Toutefois, avec l'émergence de logiciels généraux d'estimation, il semble de moins en moins justifiable de ne pas recourir à des estimateurs de la variance qui font une juste évaluation de l'effet de l'imputation.

De nombreux problèmes demeurent évidemment non résolus, et sont peut-être impossibles à résoudre. Premièrement, l'imputation du plus proche voisin doit faire l'objet de travaux théoriques beaucoup plus intenses. Les rajustements de la méthode du jackknife que nous avons appliqués à cette méthode d'imputation ont donné un rendement moindre que ceux qui ont été appliqués aux autres méthodes. Il faudra peut-être trouver des fonctions plus lisses pour remplacer la méthode du plus proche voisin. Deuxièmement, la robustesse de l'estimateur proposé doit être évaluée. Il est clair qu'un rendement satisfaisant peut être obtenu si le modèle (15) est vérifié et que la non-réponse est aléatoire. Le non-respect partiel de l'une ou l'autre de ces conditions n'a pas semblé entacher le bon rendement de l'estimateur jackknife dans notre expérience limitée, mais d'autres travaux devront être entrepris dans cette voie. La dérogation aux deux conditions simultanément n'a pas encore été étudiée. Les cas de non-réponse non aléatoire dans lesquels la propension à ne pas répondre est liée à la variable y sont encore moins bien compris, bien qu'il faille mettre l'accent, dans ce cas, sur l'estimation de l'erreur quadratique moyenne plutôt que de la variance. Troisièmement, des comparaisons devraient être faites avec des résultats d'imputation multiple. Il faut reconnaître, toutefois, que des méthodes d'imputation appropriées (Rubin 1987) doivent d'abord être établies. Notons qu'aucune des méthodes d'imputation étudiées ici n'est pas appropriée en ce qui a trait à l'imputation multiple.

Il faudrait étendre l'analyse à d'autres méthodes d'imputation et à d'autres paramètres d'intérêt. La présente étude s'est limitée à quatre méthodes d'imputation simples. En pratique, des méthodes beaucoup plus complexes sont utilisées, et elles sont souvent combinées les unes aux autres. L'impact sur l'estimation de la variance du recours à plus d'une méthode d'imputation a été étudié par Rancourt, Lee et Särndal (1993); d'autres travaux sont nécessaires. Pour ce qui est d'autres méthodes d'imputation plus

complexes, l'effet de l'ajout de résidus théoriques aux données imputées pourrait, par exemple, être étudié. Toutefois, cette technique a comme unique objet la sous-estimation de V_{sam} par v_{naive}, et ne tient pas compte de l'effet de V_{imp}. Enfin, d'autres paramètres, par exemple la médiane, et l'effet de l'imputation sur leur variance restent à étudier. Des extensions multidimensionnelles pourraient aussi être envisagées: l'estimation des corrélations, des quotients et des paramètres de régression en présence d'une imputation serait sans doute un sujet intéressant.

REMERCIEMENTS

Les auteurs tiennent à remercier M. J.N.K. Rao pour son appui et ses encouragements constants, ainsi que le rédacteur associé pour ses commentaires utiles.

BIBLIOGRAPHIE

HANSEN, M., HURWITZ, W., et MADOW, W. (1953). *Sample Survey Methods and Theory*. (Volume 2), New York: J. Wiley, 139-141.

KALTON, G., et KASPRZYK, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-17.

LEE, H., RANCOURT, E., et SÄRNDAL, C.-E. (1991). Experiments with variance estimation from survey data with imputed values. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 690-695.

RAO, J.N.K. (1990). Variance estimation under imputation for missing data. Rapport non publié, Statistique Canada.

RAO, J.N.K. (1992). Jackknife variance estimation under imputation for missing survey data. Rapport non publié, Statistique Canada.

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

RANCOURT, E., LEE, H., et SÄRNDAL, C.-E. (1992). Bias corrections for survey estimates from data with imputed values for nonignorable nonresponse. *Proceedings 1992 Annual Research Conference*, Bureau of the Census, 523-539.

RANCOURT, E., LEE, H., et SÄRNDAL, C.-E. (1993). Variance estimation under more than one imputation method. *Proceedings of the International Conference on Establishment Surveys, American Statistical Association*, 374-379.

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

SÄRNDAL, C.-E. (1990). Méthodes pour estimer la précision des estimations d'enquête lorsqu'il y a eu imputation. Conférencier spécial invité. *Recueil: Symposium 90, Mesure et amélioration de la qualité des données*, Statistique Canada, 337-350.