JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

DAVID HAZIZA

Inférence en présence d'imputation simple dans les enquêtes : un survol

Journal de la société française de statistique, tome 146, n° 4 (2005), p. 69-118

<http://www.numdam.org/item?id=JSFS_2005__146_4_69_0>

INFÉRENCE EN PRÉSENCE D'IMPUTATION SIMPLE DANS LES ENQUÊTES : UN SURVOL*

David HAZIZA **

RÉSUMÉ

Dans les enquêtes, on est presque toujours confronté à une non-réponse partielle qui est habituellement traitée en imputant les valeurs manquantes. Bien que l'imputation présente plusieurs avantages, elle comporte néanmoins certains risques. L'article a pour but de passer en revue plusieurs aspects de l'inférence en présence de données imputées : les méthodes d'imputation, l'estimation ponctuelle, la réduction du biais de non-réponse, la construction des classes d'imputation, l'estimation de la variance et la distorsion des relations entre variables. L'article conclut avec quelques perspectives de recherche dans différents domaines de l'inférence en présence de données imputées.

Mots clés : Imputation déterministe, imputation aléatoire, estimation de la variance, domaines, classes d'imputation, estimateur imputé, non-réponse uniforme, non-réponse ignorable, non-réponse non-ignorable, jackknife, bootstrap, imputation non-pondérée.

ABSTRACT

Partial nonresponse, which is usually treated by imputing missing values, is virtually certain to occur in surveys. Eventhough imputation offers several advantages, it presents however some risks. This paper reviews several aspects of inference in the presence of imputed data : the imputation methods, point estimation, the reduction of the nonresponse bias, the construction of imputation classes, variance estimation and the distortion of the relationships between variables. The paper concludes with some research perspectives in different areas of inference in the presence of imputed data.

Keywords : deterministic imputation, random imputation, variance estimation, domains, imputation classes, imputed estimator, uniform nonresponse, ignorable nonresponse, nonignorable nonresponse, jackknife, bootstrap, unweighted imputation.

---

* Cet article a été présenté dans le cadre des Journées de Méthodologie Statistique 2002, Paris (France). L'auteur remercie Anne Ruiz-Gazen, Éric Gautier ainsi que les arbitres et l'Éditeur pour leurs commentaires et suggestions qui ont grandement contribué à améliorer la qualité de l'article.

** Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Tunney's Pasture, Ottawa, ON, Canada, K1A 0T6. Courriel : david.haziza@statcan.ca

Introduction

Dans les enquêtes, il faut se résigner au fait qu'il y aura inévitablement un certain taux de non-réponse. On distingue essentiellement deux types de non-réponse : la non-réponse totale (qui est l'absence complète d'information sur une unité) et la non-réponse partielle (qui est une absence d'information limitée à certaines variables seulement). La non-réponse est un problème important dans les enquêtes car elle mène, en général, à des estimateurs ponctuels biaisés. Il s'avère donc crucial de réduire le biais de non-réponse, ce qui requiert généralement une utilisation judicieuse de l'information auxiliaire disponible. Dans cet article, nous appellerons *information auxiliaire* un ensemble de variables disponibles pour toutes les unités échantillonnées ou pour toutes les unités de la population.

Pour traiter la non-réponse, il existe plusieurs méthodes. Dans le cas d'une non-réponse totale, on fait appel, la plupart du temps, à une méthode de repondération. La repondération consiste à hausser le poids de sondage des unités répondantes dans l'échantillon afin de compenser pour les non-répondants. Quant à la non-réponse partielle, elle est habituellement traitée en utilisant l'imputation qui consiste à affecter une (ou plusieurs) « valeur(s) artificielle(s) » pour remplacer une valeur manquante.

Le mot *imputation* vient du mot latin *imputare* : porter en compte. La première utilisation du mot imputation dans le contexte des enquêtes est vraisemblablement due à Hansen, Hurvitz et Madow (1953) dans le contexte de « l'American survey of retail and shares » de 1948. Rancourt (2001) présente un historique de l'utilisation de l'imputation dans les enquêtes. L'imputation est un domaine pour lequel « la pratique a longtemps été en avance sur la théorie ». Bien que l'utilisation de l'imputation ait proliféré dans les enquêtes au début des années 70 (avec l'avènement des systèmes automatisés de vérification et d'imputation), les résultats théoriques n'ont commencé à apparaître que très récemment.

On distingue l'imputation simple de l'imputation multiple : l'imputation simple consiste à créer une valeur unique pour « boucher le trou » laissé par la valeur manquante, ce qui mènera à la création d'un fichier complet. L'imputation multiple, suggérée par Rubin (1978), consiste à créer $M \geqslant 2$ valeurs imputées pour chaque valeur manquante, ce qui mènera à la création de M fichiers de données complets. L'idée de base de l'imputation multiple est de combiner les estimations issues de chacun des fichiers complétés de manière adéquate afin d'obtenir un estimateur ponctuel et un estimateur de variance qui tiennent compte de la non-réponse. Pour certaines raisons, les statisticiens d'enquête emploient presque toujours l'imputation simple. C'est pourquoi, dans ce qui suit, nous discutons exclusivement de l'imputation simple.

Les statisticiens d'enquête tentent généralement d'éviter l'utilisation de modèles à des fins d'inférence. Il est coutume de mener une inférence basée uniquement sur le plan de sondage. Ceci n'est cependant pas possible en présence de non-réponse et l'emploi de modèles devient incontournable. La validité des estimateurs (ponctuels et de variance) dépend alors en grande partie de la

validité de certaines hypothèses (ou modèles) à propos du mécanisme de non-réponse et/ou du modèle d'imputation. Ainsi, l'imputation est avant tout un exercice de modélisation. La qualité des estimations repose donc sur la disponibilité d'une information auxiliaire de qualité qui servira à construire des valeurs imputées et/ou à construire des classes d'imputation.

Les estimations produites lors d'une enquête sont sujettes à de nombreuses erreurs dont les erreurs d'échantillonnage, les erreurs de mesure, les erreurs de codification, les erreurs de couverture et les erreurs de non-réponse. Dans cet article, nous nous limitons aux erreurs d'échantillonnage et aux erreurs de non-réponse et nous supposons que les autres erreurs sont négligeables. Par ailleurs, bien que nous mettions l'accent sur les enquêtes par sondage, les grands principes décrits dans cet article n'en sont pas moins pertinents dans le cas de recensements ou de données administratives [1].

L'article est construit de la façon suivante : dans la section 1, nous présentons quelques méthodes pouvant être utilisées pour le traitement de la non-réponse partielle dans les enquêtes. Dans la section 2, nous présentons l'estimateur imputé d'un total et nous décrivons plusieurs méthodes d'imputation fréquemment utilisées dans les enquêtes. Nous y discutons également de la notion de mécanisme de non-réponse et des cadres de travail proposés dans la littérature pour mener une inférence. Le biais de l'estimateur imputé fait l'objet de la section 3. Dans la section 4, nous montrons que certaines méthodes d'imputation déterministes ont tendance à distordre la distribution des variables d'intérêt (c'est-à-dire que la fonction de répartition après imputation pour une variable donnée est un estimateur biaisé de la fonction de répartition prévalant au niveau de la population) alors que les méthodes d'imputation aléatoires ont tendance à la préserver. L'estimation de la variance en présence de valeurs imputées sera traitée dans la section 5. D'une part, nous montrons par un exemple pourquoi il ne faut pas traiter les valeurs imputées comme si elles avaient été observées, et d'autre part, nous présentons quelques méthodes qui permettent d'estimer la variance correctement en présence de valeurs imputées. Dans la section 6, nous discutons du problème de la distorsion des relations entre les variables. Nous traitons des classes d'imputation et de leur construction dans la section 7. Nous y présentons également une étude par simulation. Finalement, nous concluons dans la section 8 et présentons certaines questions qui nécessiteront des études théoriques et empiriques dans le futur.

1. Traitement de la non-réponse partielle

En présence de non-réponse partielle, plusieurs options s'offrent au statisticien d'enquête quant au traitement des valeurs manquantes. En plus de l'imputation, deux options méritent d'être soulignées :

1. NDLR : pour d'autres situations où la présence de données manquantes requiert des mécanismes d'imputation, voir le numéro spécial de ce Journal « Données longitudinales Incomplètes », vol. 145, n°2 (2004).

1.1. Utilisation des répondants complets seulement

Cette option équivaut à éliminer les unités pour lesquelles il y a au moins une valeur manquante. Les estimations requises sont alors basées seulement sur l'ensemble des répondants complets. Bien qu'elle soit simple et qu'elle permette d'utiliser un fichier de données complet, cette option présente certains risques. En effet, elle mène généralement à des estimateurs fortement biaisés pour l'estimation de totaux et de moyennes à moins que la non-réponse ne soit indépendante de toutes les variables (variables d'intérêt et variables auxiliaires), ce qui survient, par exemple, lorsque le mécanisme de non-réponse est uniforme (voir section 2.4). En rejetant tous les répondants partiels, le statisticien se prive également d'une information de grande valeur. Finalement, les poids de sondage ne peuvent être utilisés pour faire l'inférence (à moins qu'ils ne soient ajustés) si bien que cette dernière doit être conditionnelle à l'échantillon de répondants complets. Cette option ne doit donc pas être sérieusement considérée à d'autres fins qu'une description sommaire du fichier.

1.2. Méthodes de repondération

Les méthodes de repondération peuvent s'avérer une solution de choix dans certains cas pour résoudre le problème de la non-réponse partielle. Les méthodes de repondération sont généralement simples et l'information auxiliaire disponible peut être utilisée à bon escient pour former les classes de repondération. Le principal inconvénient de la repondération est que celle-ci force le statisticien à créer un poids ajusté pour chacune des variables mesurées par l'enquête. Ainsi, dans une enquête comprenant plus d'une centaine de variables (par exemple l'Enquête sur la Population Active Canadienne – EPA –), il faudrait créer autant de poids ajustés, ce qui explique en grande partie la raison pour laquelle cette option est généralement rejetée dans le cas de la non-réponse partielle.

1.3. Pourquoi impute-t-on ?

L'imputation présente plusieurs avantages pour traiter la non-réponse partielle parmi lesquels :

(1) L'imputation mène à la création d'un fichier de données complet.

(2) Les résultats issus de différentes analyses seront vraisemblablement cohérents.

(3) Contrairement aux méthodes de repondération, l'imputation permet l'utilisation d'un poids de sondage unique.

(4) L'information disponible sur les répondants partiels peut être utilisée comme information auxiliaire pour améliorer la qualité des valeurs imputées.

Le choix entre la repondération et l'imputation est relativement aisé. Il existe toutefois certaines situations pour lesquelles le choix s'avère plus nébuleux. Par exemple, à l'EPA, les répondants qui décident de mettre fin prématurément à l'entretien sont identifiés comme non-répondants complets, et ce, malgré le fait

que certaines informations ont été collectées sur lesdites unités. La quantité d'information recueillie est alors jugée insuffisante.

Toutefois, nous insistons sur le fait que l'imputation comporte également plusieurs risques dont les plus importants sont les suivants :

(1) Bien que l'imputation mène à la création d'un fichier de données complet, l'inférence, en particulier l'estimation ponctuelle, n'est valide que si les hypothèses sous-jacentes (concernant le mécanisme de non-réponse et/ou le modèle d'imputation) sont satisfaites.

(2) Certaines méthodes d'imputation ont tendance à distordre la distribution des variables d'intérêt.

(3) L'imputation a tendance à distordre les relations entre les variables.

(4) Le fait de traiter les valeurs imputées comme des valeurs observées peut entraîner une sous-estimation substantielle de la variance de l'estimateur, surtout si le taux de non-réponse n'est pas négligeable.

Dans cet article, nous décrirons en détail chacun des risques (1)-(4) et nous donnerons quelques exemples en guise d'illustration.

2. Contexte et définitions

2.1. Un estimateur imputé

Considérons une population finie U de taille N. L'objectif est d'estimer des paramètres simples (les paramètres plus complexes feront l'objet des sections 6.1 et 6.2) tels un total ou une moyenne d'une variable d'intérêt y, donnés respectivement par $Y = \sum_{i \in U} y_i$ et $\overline{Y} = \frac{1}{N} \sum_{i \in U} y_i$. Pour cela, on tire un échantillon aléatoire s, de taille n, selon un plan de sondage $p(.)$. Soit I_i la variable indicatrice de sélection dans l'échantillon s définie par

$$I_i = \begin{cases} 1 & \text{si l'unité } i \text{ appartient à } s \\ 0 & \text{sinon} \end{cases}$$

Dans ce qui suit, nous supposerons que le plan de sondage est ignorable. Un plan de sondage est dit ignorable si $P(I_i = 1|y, \mathbf{z}) = P(I_i = 1|\mathbf{z})$ où $\mathbf{z}$ est un vecteur de variables auxiliaires disponible, à l'étape du plan de sondage, pour toutes les unités dans la population. On montre aisément qu'un plan de sondage est ignorable si la distribution de la variable d'intérêt dans l'échantillon est identique à celle qui prévaut dans la population après avoir pris en compte l'information auxiliaire appropriée. En l'absence de non-réponse, un estimateur de Y est donné par

$$\widehat{Y} = \sum_{i \in s} w_i y_i \,, \tag{2.1}$$

où $w_i = 1/\pi_i$ désigne le poids de sondage de l'unité i et $\pi_i = P(i \in s)$ est la probabilité d'inclusion de l'unité i dans l'échantillon. L'estimateur (2.1) est

sans biais pour Y, c'est-à-dire, $E_p(\widehat{Y}) = Y$ où $E_p(.)$ désigne l'espérance par rapport au plan de sondage $p(.)$.

En présence de non-réponse à la variable y, il est impossible de calculer l'estimateur (2.1) puisque certaines valeurs sont manquantes. On définit plutôt un estimateur imputé de Y selon

$$\widehat{Y}_I = \sum_{i \in s_r} w_i y_i + \sum_{i \in s_m} w_i y_i^* , \tag{2.2}$$

où s_r est l'ensemble des r unités qui ont répondu à l'item y, s_m est l'ensemble des m unités qui n'ont pas répondu à l'item $y (r + m = n)$, et y_i^* est la valeur imputée pour remplacer la valeur manquante y_i. Notons que $\widehat{Y}_I$ est simplement le total pondéré des valeurs observées et des valeurs imputées dans l'échantillon. L'estimateur imputé d'une moyenne, $\overline{Y}$, est obtenu en divisant $\widehat{Y}_I$ par N ou par la taille estimée de la population $\widehat{N} = \sum_{i \in s} w_i$.

2.2. Méthodes d'imputation

On distingue généralement les méthodes d'imputation dites déterministes de celles dites aléatoires. Les méthodes déterministes sont celles qui fournissent une valeur fixe étant donné l'échantillon si le processus d'imputation est répété (voir section 2.2.1). Les méthodes aléatoires sont celles qui ont une composante aléatoire ; par conséquent, ces méthodes ne fournissent pas nécessairement la même valeur si l'on répète le processus d'imputation pour un échantillon donné (voir section 2.2.2). L'article de Kovar et Whitridge (1995) fournit une bonne revue des méthodes d'imputation. La majorité des méthodes d'imputation peut être représentée par le modèle (Kalton et Kasprzyk, 1986),

$$\begin{gathered} y_i = f(\mathbf{z_i}) + \varepsilon_i , \\ E(\varepsilon_i) = 0 , \quad E(\varepsilon_i \varepsilon_j) = 0 , \; i \neq j , \quad E(\varepsilon_i^2) = \sigma_i^2 , \end{gathered} \tag{2.3}$$

où $\mathbf{z}$ est un vecteur de variables auxiliaires disponible pour toutes les unités dans l'échantillon s. Dans le cas des méthodes déterministes, la valeur imputée y_i^* est obtenue en estimant la fonction f par $\widehat{f}_r$ au moyen des unités répondantes $i \in s_r$, et en posant $y_i^* = \widehat{f}_r(\mathbf{z_i})$. L'imputation aléatoire peut être vue comme une imputation déterministe à laquelle on a ajouté un résidu aléatoire e^*. Ce résidu peut être tiré, par exemple, à partir d'une loi normale de moyenne 0 et de variance u. En pratique, on préfère plutôt utiliser un résidu aléatoire tiré au hasard parmi les résidus observés dans l'ensemble s_r des répondants. Soit $e_j = \dfrac{1}{\widehat{\sigma}_j}[y_j - \widehat{f}_r(\mathbf{z_j})]$ le résidu standardisé pour le répondant $j \in s_r$ où $\widehat{\sigma}_j$ est un estimateur de σ_j. La valeur manquante pour la i^{e} unité, y_i, est alors remplacée par

$$y_i^* = \widehat{f}_r(\mathbf{z_i}) + \widehat{\sigma}_i e_i^* , \tag{2.4}$$

où e_i^* est tiré au hasard (habituellement avec remise) dans l'ensemble des résidus standardisés correspondant aux répondants.

Notons que le modèle (2.3) peut également servir à représenter une méthode d'imputation à l'intérieur de classes d'imputation (par exemple, l'imputation par la moyenne à l'intérieur de classes).

En pratique, on peut recourir à l'imputation pondérée ou l'imputation non-pondérée. Dans le cas de l'imputation pondérée (déterministe ou aléatoire), on utilise les poids de sondage, w_i, dans la procédure d'imputation. Notons que les méthodes d'imputation pondérée et non-pondérée donnent des résultats identiques pour des plans de sondage autopondérés (plans avec poids de sondage égaux). Dans ce qui suit, nous considérons le cas de l'imputation pondérée seulement. Le problème de l'imputation non-pondérée, qui consiste à poser $w_i = 1$ sera traité dans la section 6.3. Dans les sections 2.2.1 et 2.2.2, nous décrivons quelques méthodes d'imputation fréquemment utilisées dans les enquêtes.

2.2.1. Quelques méthodes déterministes

(1) Imputation par la régression : Cette méthode consiste à remplacer une valeur manquante par la valeur prédite au moyen d'un modèle de régression linéaire. Dans ce cas, $f(\mathbf{z_i}) = \mathbf{z_i'\beta}$ où $\boldsymbol{\beta}$ est un vecteur de paramètres inconnus et $\sigma_i^2 = \sigma^2 \boldsymbol{\lambda}' \mathbf{z_i}$ pour un certain vecteur de constantes $\boldsymbol{\lambda}$. On a alors

$$y_i^* = \mathbf{z_i'}\widehat{\mathbf{B}}_\mathbf{r}, \tag{2.5}$$

où $\widehat{\mathbf{B}}_\mathbf{r} = \left(\sum_{i \in s_r} w_i \mathbf{z_i} \mathbf{z_i'}/(\boldsymbol{\lambda}' \mathbf{z_i})\right)^{-1} \cdot \left(\sum_{i \in s_r} w_i \mathbf{z_i} y_i/(\boldsymbol{\lambda}' \mathbf{z_i})\right)$ est l'estimateur obtenu par la méthode des moindres carrés. Un cas particulier de (2.5) est l'imputation par la régression linéaire simple. Dans ce cas, une seule variable auxiliaire z est utilisée, $f(\mathbf{z_i}) = \beta_0 + \beta_1 z_i$ et $\sigma_i^2 = \sigma^2$. On a alors

$$y_i^* = \widehat{B}_{0r} + \widehat{B}_{1r} z_i, \tag{2.6}$$

où $\widehat{B}_{1r} = \dfrac{\sum_{i \in s_r} w_i (z_i - \overline{z}_r)(y_i - \overline{y}_r)}{\sum_{i \in s_r} w_i (z_i - \overline{z}_r)^2}$, $\widehat{B}_{0r} = \overline{y}_r - \widehat{B}_{1r} \overline{z}_r$ et $(\overline{y}_r, \overline{z}_r) = \dfrac{1}{\sum_{i \in s_r} w_i} \sum_{i \in s_r} w_i (y_i, z_i)$ désignent respectivement la moyenne des répondants pour les variables y et z.

(2) Imputation par le ratio : Le ratio est un cas particulier de l'imputation par la régression. Dans ce cas, une seule variable auxiliaire z est utilisée, $f(\mathbf{z_i}) = \beta z_i$ et $\sigma_i^2 = \sigma^2 z_i$. On a alors

$$y_i^* = \widehat{B}_r z_i, \tag{2.7}$$

où $\widehat{B}_r = \dfrac{\overline{y}_r}{\overline{z}_r}$.

(3) Imputation par la valeur précédente (ou historique) : Cette méthode consiste à utiliser la valeur observée sur la même unité à une occasion précédente. Soit $y_{i,t-1}$ la valeur observée pour l'unité i au temps $t-1$. On a $f(\mathbf{z_i}) = y_{i,t-1}$, $\sigma_i^2 = \sigma^2$ et

$$y_{i,t}^* = y_{i,t-1}. \tag{2.8}$$

(4) Imputation par la moyenne : Cette méthode consiste à affecter la moyenne des répondants aux non-répondants. On a $\mathbf{z_i} = 1\ \forall i \in U$, $f(\mathbf{z_i}) = \beta$, $\sigma_i^2 = \sigma^2$ et

$$y_i^* = \overline{y}_r. \tag{2.9}$$

(5) Imputation par le plus proche voisin (PPV) : Cette méthode consiste à remplacer une valeur manquante par la valeur d'un donneur choisi de telle manière qu'une certaine distance entre le répondant (donneur) et le non-répondant (receveur), par rapport à un ensemble de variables auxiliaires, est minimum. L'imputation par PPV est une méthode d'imputation non-paramétrique. En effet, on ne spécifie pas la forme de $f(\mathbf{z_i})$, pas plus que la structure de variance σ_i^2. Dans le cas de l'imputation par PPV,

$$y_i^* = y_j\ ,\ j \in s_r\ ,\ \text{tel que dist}(\mathbf{z_i}, \mathbf{z_j}) \text{ est minimum}, \tag{2.10}$$

où dist(.,.) est une fonction de distance donnée (par exemple, on peut prendre une distance euclidienne).

2.2.2. Quelques méthodes aléatoires

(1) Imputation par hot-deck aléatoire : Cette méthode consiste à tirer un répondant (donneur) au hasard avec remise dans l'ensemble s_r des répondants, et à « donner » la valeur du répondant au non-répondant (receveur), c'est-à-dire,

$$y_i^* = y_j\ ,\ j \in s_r\ ,\ \text{tel que } P(y_i^* = y_j) = w_j / \sum_{k \in s_r} w_k. \tag{2.11}$$

L'imputation par hot-deck aléatoire peut être vue comme de l'imputation par la moyenne à laquelle on a ajouté un résidu $e_i^* = y_j - \overline{y}_r$, tel que décrit en (2.4).

(2) Imputation par la régression avec résidus : L'imputation par la régression avec résidu est équivalente à l'imputation par la régression à laquelle on a ajouté un résidu aléatoire tiré (avec remise) dans l'ensemble des résidus standardisés correspondant aux répondants. La valeur imputée utilisée pour remplacer la valeur manquante y_i est alors donnée par

$$y_i^* = \mathbf{z_i'}\widehat{\mathbf{B}}_\mathbf{r} + (\boldsymbol{\lambda}'\mathbf{z_i})^{1/2} e_i^*\ , \tag{2.12}$$

où $e_i^* = e_j$, $j \in s_r$, tel que $P(e_i^* = e_j) = w_j / \sum_{k \in s_r} w_k$ et $e_j = (\boldsymbol{\lambda}'\mathbf{z_j})^{-1/2} \lfloor y_j - \mathbf{z_j'}\widehat{\mathbf{B}}_\mathbf{r} \rfloor$.

2.3. Non-réponse et échantillonnage à deux phases

La situation prévalant en présence de non-réponse est souvent comparée à celle prévalant dans le cas de l'échantillonnage à deux phases. Ce dernier est fréquemment utilisé dans les enquêtes lorsque la base de sondage contient peu ou pas d'information. Dans ce cas, il est d'usage de tirer préalablement un échantillon s_1 de première phase généralement de grande taille selon un plan de sondage $p_1(.)$, ce qui permettra de recueillir de l'information auxiliaire peu coûteuse. À l'aide de l'information recueillie en première phase, on tire un échantillon s_2 de s_1 selon un plan de sondage $p_2(.|s_1)$. Un estimateur (qui n'utilise pas d'information auxiliaire) de Y est alors donné par

$$\widehat{Y}_{DP} = \sum_{i \in s_2} w_{1i} w_{2i} y_i \,, \tag{2.13}$$

où $w_{1i} = 1/\pi_{1i}$ et $\pi_{1i} = P(i \in s_1)$ est la probabilité d'inclusion de l'unité i dans l'échantillon de première phase s_1, $w_{2i} = 1/\pi_{2i}$ et $\pi_{2i} = P(i \in s_2 | s_1, i \in s_1)$ est la probabilité d'inclusion conditionnelle de l'unité i dans l'échantillon de deuxième phase s_2. L'estimateur $\widehat{Y}_{DP}$ en (2.13) est sans biais par rapport au plan de sondage pour Y, c'est-à-dire, $E(\widehat{Y}_{DP}) = E_{p_1} E_{p_2}(\widehat{Y}_{DP}|s_1) = Y$ où $E_{p_1}(.)$ et $E_{p_2}(.)$ désignent respectivement l'espérance par rapport aux plans de sondage $p_1(.)$ et $p_2(.|s_1)$. Notons que, dans le cas de l'échantillonnage à deux phases, le statisticien contrôle le mécanisme de sélection des deux échantillons. Autrement dit, les probabilités d'inclusion π_{1i} et π_{2i} sont connues. En présence de non-réponse, l'ensemble des répondants s_r est souvent vu comme un échantillon de deuxième phase. Cependant, dans ce cas, le mécanisme (appelé mécanisme de non-réponse) qui a mené à s_r n'est généralement pas connu et par conséquent les probabilités d'inclusion dans s_r (c'est-à-dire les probabilités de réponse) ne sont pas connues. Nous n'aurons donc d'autres choix que d'établir certaines hypothèses à propos du mécanisme de non-réponse.

2.4. Mécanisme de non-réponse

Commençons par définir le concept de mécanisme de non-réponse. Pour être rigoureux, il convient de parler *des* mécanismes de non-réponse puisque les causes et raisons qui mènent à des valeurs manquantes sont généralement nombreuses. Tenter de décrire de manière précise toutes ces raisons serait toutefois irréaliste. C'est pourquoi, dans ce qui suit, nous parlerons *du* mécanisme de non-réponse. Soit a_i la variable indicatrice de réponse définie par

$$a_i = \begin{cases} 1 & \text{si l'unité } i \text{ appartient à } s_r \\ 0 & \text{si l'unité } i \text{ appartient à } s_m \end{cases}$$

La distribution des variables indicatrices, $p(a_i|s)$, est appelée mécanisme de non-réponse. Cette distribution n'étant généralement pas connue, on n'a d'autre choix que d'établir certaines hypothèses à propos du mécanisme de non-réponse. Soit $p_i = P(a_i = 1|s, i \in s)$ la probabilité de réponse pour l'unité i. Nous supposerons que les unités répondent indépendamment les unes des autres, c'est-à-dire, $p_{ij} = P(a_i = 1, a_j = 1|s, i \in s, j \in s, i \neq j) = p_i p_j$.

L'hypothèse d'indépendance est fréquemment satisfaite en pratique bien que l'on puisse facilement concevoir certaines situations où elle ne l'est pas. Par exemple, dans le cas d'un sondage par grappes, les unités à l'intérieur d'une grappe pourraient ne pas répondre indépendamment les unes des autres. On peut alors faire appel à un type de mécanisme plus complexe du type beta-binomial (Haziza et Rao, 2003a). Les variables indicatrices a_i sont donc des variables aléatoires indépendantes tirées d'une distribution de Bernoulli de paramètre p_i. On distinguera trois types de mécanismes de non-réponse.

(1) Mécanisme uniforme

Un mécanisme de non-réponse est dit uniforme si la probabilité de réponse est la même pour toutes les unités dans la population, c'est-à-dire, $p_i = p$ $\forall i \in U$. Dans le cas d'un mécanisme uniforme, la probabilité de réponse est indépendante de toutes les variables d'une enquête (variables auxiliaires et variables d'intérêt). Ce mécanisme est, bien sûr, très peu réaliste. En pratique, on supposera plutôt un mécanisme uniforme à l'intérieur de classes. Lorsque le mécanisme est uniforme, on dit alors que les données sont *Missing Completely At Random* (MCAR). La proposition suivante est due à Oh et Scheuren (1983).

PROPOSITION 1. — *Supposons que l'on tire un échantillon aléatoire simple sans remise, s, de taille n, d'une population U de taille N. Si le mécanisme de non-réponse est uniforme, alors, étant donné l'échantillon s et le nombre de répondants r, l'ensemble des répondants s_r est un échantillon aléatoire simple sans remise tiré de la population U, c'est-à-dire,*

$$P(s_r|s,r) = \frac{r!(N-r)!}{N!}$$

Ce résultat s'avèrera utile pour illustrer certains résultats lorsque nous discuterons de l'estimation ponctuelle et de l'estimation de la variance en présence de données imputées.

(2) Mécanisme ignorable

Un mécanisme de non-réponse est dit ignorable si $P(a_i = 1|y, \mathbf{z}) = P(a_i = 1|\mathbf{z})$, (Rubin, 1976). Dans ce cas, la probabilité de réponse peut dépendre de variables auxiliaires mais pas de la variable d'intérêt (celle que l'on impute). Autrement dit, après avoir pris en compte l'information auxiliaire appropriée, la distribution de la variable d'intérêt dans l'ensemble des répondants est identique à celle qui prévaudrait dans l'échantillon s si on n'avait pas observé de non-réponse. Lorsque le mécanisme est ignorable, on dit que les données sont *Missing At Random* (MAR). Dans la littérature, on emploie également l'expression *mécanisme non-confondu.* Notons qu'un mécanisme uniforme est un cas particulier d'un mécanisme ignorable.

(3) Mécanisme non-ignorable

Lorsque la probabilité de réponse dépend de la variable d'intérêt, on dit que le mécanisme de non-réponse est non-ignorable. Lorsque le mécanisme est non-ignorable, on dit que les données sont *Not Missing At Random* (NMAR). Dans

la littérature, on utilise aussi l'expression *mécanisme confondu*. En présence d'un mécanisme non-ignorable, il y aura inévitablement un biais de non-réponse. L'élimination de ce biais va généralement requérir des techniques sophistiquées (par exemple, Qin, Leung et Shao, 2002).

Dans certaines situations, il est connu que le mécanisme de non-réponse est ignorable. Ce type de situation survient en présence de non-réponse « planifiée ». Par exemple, dans le cas de l'échantillonnage à deux-phases, la variable d'intérêt est observée pour les unités appartenant à s_2 mais pas pour celles appartenant à $s_1 \setminus s_2$. Un autre exemple survient dans le cas de questionnaires modulaires pour lesquels le questionnaire est divisé en sections. Des ensembles de sections sont alors aléatoirement assignés aux unités échantillonnées. Dans ces deux exemples, le mécanisme de non-réponse est contrôlé par le statisticien d'enquête, ce qui facilite grandement le problème de l'estimation.

Lorsque la non-réponse est hors de notre contrôle, nous ne pouvons affirmer avec certitude que le mécanisme de non-réponse est bien ignorable, à moins d'effectuer un suivi des non-répondants et de recueillir les variables que nous n'avions pas réussi à obtenir lors de tentatives antérieures. Nous pourrions alors comparer les caractéristiques des répondants avec celles des non-répondants. L'ignorabilité ou non du mécanisme de non-réponse relève donc de l'hypothèse sur laquelle nous nous baserons à des fins d'inférence.

2.5. Cadres de travail pour l'inférence

Afin d'étudier les propriétés (biais et variance) de l'estimateur imputé (2.2), deux cadres de travail ont été utilisés dans la littérature : le cadre de travail basé sur le plan de sondage (BP) proposé par Rao (1990) et celui basé sur un modèle (BM) proposé par Särndal (1990, 1992). Avant d'imputer, il est coutume de former des classes et d'imputer indépendamment à l'intérieur de chaque classe. Par souci de simplicité, nous considérons le cas d'une seule classe d'imputation. Les classes d'imputation seront discutées en détail dans la section 7.

Sous le cadre de travail BP, on fait l'hypothèse suivante :

HYPOTHÈSE BP. — *À l'intérieur de la classe, on suppose que le mécanisme de non-réponse est uniforme.*

Sous le cadre de travail BM, on fait l'hypothèse suivante :

HYPOTHÈSE BM. — *À l'intérieur de la classe, on suppose que le mécanisme de non-réponse est ignorable. On fait alors appel à un modèle d'imputation, généralement de la forme*

$$\begin{gathered} m : y_i = \mathbf{z}_i'\boldsymbol{\beta} + \varepsilon_i\,, \\ E_m(\varepsilon_i) = 0, \qquad E_m(\varepsilon_i\varepsilon_j) = 0,\ i \neq j, \qquad E_m(\varepsilon_i^2) = \sigma_i^2 = \sigma^2\boldsymbol{\lambda}'\mathbf{z}_i\,, \end{gathered} \tag{2.14}$$

où $E_m(.)$ désigne l'espérance par rapport au modèle d'imputation.

Le cadre de travail BP consiste à décrire complètement le mécanisme de non-réponse, alors que le cadre de travail BM suppose un mécanisme plus général que nous n'essayons pas de décrire, d'où l'emploi d'un modèle d'imputation décrivant la relation qui lie la variable d'intérêt aux variables auxiliaires. Il s'ensuit que, sous le cadre de travail BP, la validité des estimateurs dépend de la validité du mécanisme de non-réponse tandis que, sous le cadre de travail BM, la validité des estimateurs dépend de la validité du modèle d'imputation. En pratique, les classes d'imputation sont formées de manière à satisfaire l'une des deux hypothèses BP ou BM. Une question se pose alors naturellement : quelle approche choisir ? Est-il préférable de modéliser la probabilité de réponse à une variable donnée (à l'aide d'un modèle logistique, par exemple) et de former des classes homogènes par rapport aux probabilités de réponse estimées, auquel cas on satisfait l'hypothèse BP, ou vaut-il mieux modéliser la variable d'intérêt (celle que l'on impute) et s'assurer que le modèle (2.14) soit valide à l'intérieur des classes, auquel cas on satisfait l'hypothèse BM ? Le but premier de l'imputation étant de réduire le biais de non-réponse, le choix de l'approche (hypothèse BP ou hypothèse BM) doit être dicté par la qualité des modèles sous-jacents. Dans le contexte de l'imputation, il semble intuitivement plus attrayant de modéliser la variable d'intérêt et de mener une inférence sous l'hypothèse BM. Cependant, il existe certaines situations pratiques pour lesquelles le modèle pour la variable d'intérêt n'est pas très satisfaisant (parce qu'il n'ajuste pas les données adéquatement ou parce que peu prédictif) auquel cas modéliser les probabilités de réponse peut s'avérer un choix judicieux. Par exemple, l'Enquête sur les Dépenses en Immobilisations menée à Statistique Canada produit des données sur les investissements qui se font au Canada et dans tous les types d'industries Canadiennes. Pour cette enquête, les variables d'intérêt sont les capitaux immobilisés pour de la nouvelle construction (CC) ainsi que les capitaux immobilisés pour de la nouvelle machinerie et du nouvel équipement (CM). Pour une année donnée, un grand nombre d'entreprises n'ont investi aucun montant pour la nouvelle construction ou du nouvel équipement si bien que le fichier de données avant imputation contient un grand nombre de valeurs égales à zéro. Dans ce cas, la modélisation de la variable d'intérêt (CC ou CM) peut s'avérer ardue et/ou le modèle sera relativement pauvre.

3. Biais de l'estimateur imputé

Nous avons vu dans la section 2.1 qu'en l'absence de non-réponse, l'estimateur $\widehat{Y}$ en (2.1) est sans biais pour Y. Qu'en est-il de l'estimateur imputé $\widehat{Y}_I$ en (2.2) ? Le biais de l'estimateur imputé dépendra de la validité des hypothèses à propos du mécanisme de non-réponse et/ou du modèle d'imputation.

3.1. Biais de non-réponse

Dans cette section, nous définissons le concept de biais de non-réponse. Pour cela, nous utilisons la décomposition suivante de l'erreur totale, $\widehat{Y}_I - Y$ comme point de départ :

$$\widehat{Y}_I - Y = (\widehat{Y} - Y) + (\widehat{Y}_I - \widehat{Y}), \tag{3.1}$$

où $\widehat{Y}$ est donné par (2.1). Les termes $\widehat{Y} - Y$ et $\widehat{Y}_I - \widehat{Y}$ en (3.1) représentent l'erreur due à l'échantillonnage et l'erreur due à la non-réponse, respectivement. Rappelons qu'en l'absence de non-réponse, $\widehat{Y}$ est un estimateur sans biais de Y, d'où $E_p(\widehat{Y} - Y) = 0$.

3.1.1. Biais de non-réponse sous le cadre de travail BP

Sous le cadre de travail BP, le biais de non-réponse de l'estimateur imputé $\widehat{Y}_I$ est donné par

$$\begin{aligned} \text{Biais}(\widehat{Y}_I) = E_p E_r(\widehat{Y}_I - Y|s) &= E_p E_r(\widehat{Y} - Y|s) + E_p E_r(\widehat{Y}_I - \widehat{Y}|s) \\ &= E_p(\widehat{Y} - Y) + E_p E_r(\widehat{Y}_I - \widehat{Y}|s) \\ &= E_p(B_r), \end{aligned} \tag{3.2}$$

où $B_r = E_r(\widehat{Y}_I - \widehat{Y}|s)$ est le biais de non-réponse conditionnel, étant donné l'échantillon s et $E_r(.)$ désigne l'espérance par rapport au mécanisme de non-réponse. Sous le cadre de travail BP, l'estimateur imputé est sans biais lorsque $E_p(B_r) = 0$. Dans le cas de l'imputation aléatoire, il faut tenir compte du mécanisme d'imputation qui consiste à tirer des résidus dans l'ensemble des résidus correspondant aux répondants. Dans ce cas, le biais de l'estimateur imputé est

$$\text{Biais}(\widehat{Y}_I) = E_p E_r E_I(\widehat{Y}_I - Y|s) = E_p(B_{rI}),$$

où $B_{rI} = E_r E_I(\widehat{Y}_I - \widehat{Y}|s)$ et $E_I(.)$ désigne l'espérance par rapport au mécanisme d'imputation.

3.1.2. Biais de non-réponse sous le cadre de travail BM

Sous le cadre de travail BM, le biais de non-réponse de l'estimateur imputé $\widehat{Y}_I$ est donné par

$$\begin{aligned} \text{Biais}(\widehat{Y}_I) &= E_m E_p E_r(\widehat{Y}_I - Y|s) \\ &= E_r E_p E_m(\widehat{Y} - Y|s) \\ &= E_r E_p E_m(\widehat{Y} - Y|s) + E_r E_p E_m(\widehat{Y}_I - \widehat{Y}|s) \\ &= E_p E_m(\widehat{Y} - Y) + E_r E_p E_m(\widehat{Y}_I - \widehat{Y}|s) \\ &= E_r E_p(B_m), \end{aligned} \tag{3.3}$$

où $B_m = E_m(\widehat{Y}_I - \widehat{Y}|s)$ est le biais de non-réponse conditionnel, étant donné l'échantillon s. Notons qu'il a été possible d'interchanger l'ordre des espérances puisque le plan de sondage et le mécanisme de non-réponse sont ignorables. Sous le cadre de travail BM, l'estimateur imputé est sans biais lorsque $E_r E_p(B_m) = 0$. Dans le cas de l'imputation aléatoire, on a

$$\text{Biais}(\widehat{Y}_I) = E_m E_p E_r E_I(\widehat{Y}_I - Y|s) = E_r E_p(B_{mI}),$$

où $B_{mI} = E_m E_I(\widehat{Y}_I - \widehat{Y}|s)$.

En pratique, il n'est pas possible de déterminer si le biais de non-réponse est nul puisque le mécanisme de non-réponse n'est pas connu. Il est alors d'usage de supposer que le biais est suffisamment petit. Cependant, cette hypothèse n'est généralement justifiée que lorsque l'information auxiliaire appropriée disponible est utilisée de manière adéquate dans la construction des valeurs imputées ou dans la construction des classes d'imputation.

3.2. Biais de l'estimateur imputé lorsque les hypothèses sont satisfaites

Lorsque les hypothèses à propos du mécanisme de non-réponse et/ou du modèle d'imputation sont satisfaites, l'estimateur imputé $\widehat{Y}_I$ sera approximativement sans biais [2] pour Y. Considérons le cas de l'imputation par la régression. Les valeurs imputées sont données par (2.5), ce qui mène à

$$\widehat{Y}_I = \widehat{Y}_r + (\widehat{\mathbf{Z}} - \widehat{\mathbf{Z}}_\mathbf{r})'\widehat{\mathbf{B}}_\mathbf{r}, \tag{3.4}$$

où $\widehat{Y}_r = \sum_{i \in s_r} w_i y_i$, $\widehat{\mathbf{Z}}_\mathbf{r} = \sum_{i \in s_r} w_i \mathbf{z_i}$ et $\widehat{\mathbf{Z}} = \sum_{i \in s} w_i \mathbf{z_i}$. Notons que l'expression (3.4) est similaire à celle d'un estimateur par la régression généralisée pour un total dans le cas de l'échantillonnage à deux phases. On peut facilement montrer (Rao, 1990) que, sous le cadre de travail BP, l'estimateur imputé $\widehat{Y}_I$ en (3.4) est approximativement sans biais pour Y. En vertu de la décomposition (3.2), on a donc $E_p E_r(\widehat{Y}_I - \widehat{Y}|s) \approx 0$. De façon similaire, on peut montrer (Särndal, 1992) que, sous le cadre de travail BM et le modèle (2.14), l'estimateur imputé $\widehat{Y}_I$ en (3.4) est approximativement sans biais pour Y. En vertu de la décomposition (3.3), on a donc $E_r E_p E_m(\widehat{Y}_I - \widehat{Y}|s) \approx 0$. Ces résultats sont également valides dans le cas de l'imputation par la moyenne et l'imputation par le ratio. Dans le cas de l'imputation par PPV, Chen et Shao (2000) ont montré que, sous certaines conditions, l'estimateur imputé (2.2) est approximativement sans biais sous le cadre de travail BM. Dans le cas de l'imputation par la régression avec résidus (2.12), l'estimateur imputé est approximativement sans biais (sous les cadres de travail BP et BM) puisque $E_I(e_i^*) = 0$ ce qui entraîne que $E_I(y_i^*) = \mathbf{z_i'}\widehat{\mathbf{B}}_\mathbf{r}$ qui est la valeur imputée utilisée dans le cas de l'imputation par la régression.

3.3. Biais de l'estimateur imputé lorsque les hypothèses ne sont pas satisfaites

Lorsque les hypothèses à propos du mécanisme de non-réponse et/ou du modèle d'imputation ne sont pas valides, l'estimateur imputé $\widehat{Y}_I$ sera vraisemblablement biaisé. Nous illustrons maintenant ce point en donnant deux exemples : l'un sous le cadre de travail BP et l'autre sous le cadre de travail BM.

---

2. Dans tout l'article, on dira que A est approximativement égal à B, et on écrira $A \approx B$ lorsque A est égal à B jusqu'à un terme qui est de grandeur négligeable par rapport à celles de A et B quand la taille de l'échantillon tend vers l'infini.

3.3.1. Cadre de travail BP et imputation par la moyenne

Sous ce cadre de travail, on suppose que le mécanisme de non-réponse est uniforme. Dans le cas de l'imputation par la moyenne, l'utilisation des valeurs imputées (2.9) dans (2.2) mène à $\widehat{Y}_I = \widehat{N}\overline{y}_r$. Dans ce cas, l'estimateur imputé $\widehat{Y}_I$ est approximativement sans biais sous réponse uniforme. Qu'arrive-t-il si l'on suppose que le mécanisme de non-réponse est uniforme alors qu'en réalité, il n'est pas uniforme? Considérons un mécanisme pour lequel la probabilité de répondre à l'item y varie d'une unité à l'autre (c'est-à-dire que $P(i \in s_r | s, i \in s) = p_i$). On peut alors montrer que l'estimateur imputé $\widehat{Y}_I$ est biaisé et que le biais, $\text{Biais}(\widehat{Y}_I) = E_p E_r(\widehat{Y}_I - Y|s)$ est donné par

$$\text{Biais}(\widehat{Y}_I) \approx \frac{1}{\overline{P}} \sum_{i \in U} (p_i - \overline{P})(y_i - \overline{Y}), \tag{3.5}$$

où $\overline{P} = \dfrac{1}{N} \sum_{i \in U} p_i$ est la moyenne des probabilités dans la population. Notons que le biais (3.5) est égal à 0 si la covariance entre la probabilité de réponse et la variable d'intérêt est nulle, ce qui est le cas, par exemple, pour un mécanisme de non-réponse uniforme ($p_i = p$). De plus, notons que, lorsque la probabilité de réponse dépend de la variable d'intérêt (mécanisme non-ignorable), le biais en (3.5) ne peut être nul. L'expression (3.5) justifie la formation de classes d'imputation (voir section 7.1).

3.3.2. Cadre de travail BM et imputation par le ratio

Dans le cas de l'imputation par le ratio, l'utilisation des valeurs imputées (2.7) dans l'estimateur imputé (2.2) mène à

$$\widehat{Y}_I = \frac{\overline{y}_r}{\overline{z}_r} \widehat{Z}, \tag{3.6}$$

où $\widehat{Z} = \sum_{i \in s} w_i z_i$. L'imputation par le ratio suggère naturellement l'emploi d'un modèle de la forme

$$y_i = \beta z_i + \varepsilon_i. \tag{3.7}$$

Comme nous l'avons vu dans la section 3.1, l'estimateur imputé (3.6) est sans biais pourvu que le modèle (3.7) soit valide. Qu'en est-il si le modèle (3.7) n'est pas valide? Encore une fois, l'estimateur imputé sera vraisemblablement biaisé. En effet, supposons que le vrai modèle qui lie les variables y et z n'est pas (3.7) mais plutôt

$$y_i = \beta_0 + \beta_1 z_i + \varepsilon_i. \tag{3.8}$$

De plus supposons que les probabilités de réponse sont telles que $P(i \in s_r | s; i \in s) = p_i$. On peut alors montrer que, sous le modèle (3.8), l'estimateur imputé (3.6) est biaisé et que le biais, $\text{Biais}(\widehat{Y}_I) = E_r E_p E_m(\widehat{Y}_I - Y|s)$

est

$$\text{Biais}(\widehat{Y}_I) \approx N\beta_0 \left[\frac{\overline{Z}}{\overline{Z}_p} - 1\right], \tag{3.9}$$

où $\overline{Z} = \frac{1}{N}\sum_{i\in U} z_i$ et $\overline{Z}_p = \sum_{i\in U} p_i z_i / \sum_{i\in U} p_i$.

Notons que le biais (3.9) est égal à 0 si

(a) $\quad \beta_0 = 0$

ou

(b) $\quad \overline{Z} = \overline{Z}_p$, c'est-à-dire $\frac{1}{N\overline{P}} \sum_{i\in U}(p_i - \overline{P})(z_i - \overline{Z}) = 0.$

La condition (a) est équivalente à satisfaire le modèle ratio (3.7). La condition (b) est satisfaite lorsque la covariance entre la probabilité de réponse et la variable auxiliaire z est nulle, ce qui survient, par exemple, dans le cas d'un mécanisme de non-réponse uniforme. En général cependant, le biais (3.9) est différent de zéro.

3.4. Exemples numériques

Pour illustrer les résultats présentés dans les sections 3.2 et 3.3, nous avons effectué deux études par simulation.

Étude 1. — Nous avons d'abord généré une population de taille $N = 1\,000$ comprenant deux variables y et z_1. De cette population nous avons tiré $R = 10\,000$ échantillons aléatoires simples sans remise. Dans chaque échantillon tiré, nous avons généré la variable indicatrice de réponse a_i à partir d'une distribution de Bernoulli de paramètre p_i où p_i est défini selon la fonction logistique

$$p_i = \frac{\exp(\gamma_0 + \gamma_1 z_{1i})}{1 + \exp(\gamma_0 + \gamma_1 z_{1i})}.$$

Les paramètres γ_0 et γ_1 ont été choisis de manière à obtenir un taux de réponse global approximativement égal à 70 %. Finalement, pour remplacer les valeurs manquantes, nous avons tour à tour utilisé l'imputation par la moyenne, l'imputation par le ratio et l'imputation par la régression linéaire simple. Les mesures Monte Carlo suivantes ont été calculées :

(1) Le biais relatif de l'estimateur imputé donné par

$$BR(\widehat{Y}_I) = \frac{1}{R}\sum_{i=1}^{R} \frac{(\widehat{Y}_I^{(i)} - Y)}{Y} \times 100,$$

où $\widehat{Y}_I^{(i)}$ représente l'estimateur imputé dans le i^{e} échantillon, $i = 1, \ldots, R$.

(2) L'erreur quadratique moyenne de l'estimateur imputé donnée par

$$EQM(\widehat{Y}_I) = \frac{1}{R}\sum_{i=1}^{R} (\widehat{Y}_I^{(i)} - Y)^2.$$

La figure 1 indique que la relation entre les variables y et z_1 dans la population est bien linéaire avec une ordonnée à l'origine nulle, ce qui est confirmé par la p-valeur (0.7333) dans le tableau 1. Le tableau 2 montre que, dans le cas de l'imputation par la moyenne, le biais n'est pas négligeable (environ 4 %). Ce résultat s'explique facilement par le fait que la probabilité de réponse et la variable d'intérêt sont corrélées avec la variable auxiliaire z_1. Or, en imputant par la moyenne, nous n'avons pas utilisé z_1 pour construire les valeurs imputées. Autrement dit, l'information auxiliaire appropriée n'a pas été incluse dans le modèle d'imputation. Pour nous en convaincre, il suffit de remarquer que l'inclusion de z_1 dans le modèle d'imputation (imputation par le ratio) a suffi pour réduire le biais à un niveau négligeable (environ 0.04 %). Les résultats très semblables obtenus à l'aide de l'imputation par le ratio et ceux obtenus par l'imputation par la régression s'expliquent par le fait que, l'ordonnée à l'origine n'étant pas significative, son inclusion dans le modèle d'imputation n'a donc pas un grand impact sur les résultats.

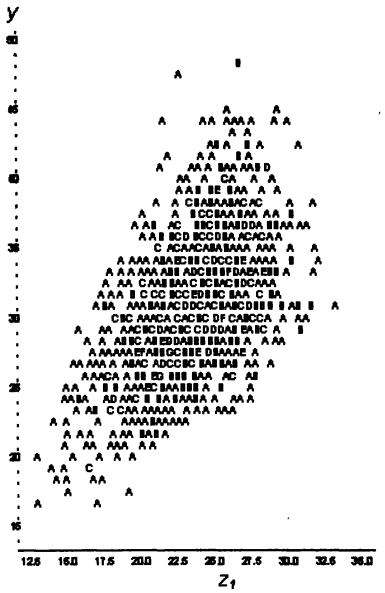


FIG 1. — Relation entre y et z_1

TABLEAU 1. – Analyse de régression

Variable	DL	Estimation des paramètres	Écart-type	t	$Pr > \|t\|$
Ordonnée à l'origine	1	0.249	0.732	0.34	0.7333
z_1	1	1.303	0.030	42.33	< .0001

TABLEAU 2. – Biais relatif (en %) et erreur quadratique moyenne des estimateurs

	Moyenne	Ratio	Régression
Biais Relatif (%)	3.99	0.038	−0.098
EQM	1.94×10^6	0.31×10^6	0.32×10^6

Étude 2. — Dans les mêmes conditions que celles de l'étude 1, nous avons effectué une autre étude par simulation en remplaçant la variable z_1 par la variable z_2.

La figure 2 indique que la relation entre les variables y et z_2 dans la population est bien linéaire mais que l'ordonnée à l'origine n'est pas nulle, ce qui est confirmé par la p-valeur (< 0.0001) dans le tableau 3. Le tableau 4 indique que, dans le cas de l'imputation par la moyenne, le biais n'est pas négligeable (environ 6.6 %). L'explication de ce biais est similaire à celle dans l'étude 1. On constate cependant que l'inclusion de la variable z_2 dans le modèle d'imputation dans le cas de l'imputation par le ratio n'a pas résolu le problème. En effet, dans ce cas, le biais relatif est encore plus grand en valeur absolue (environ 14 %). Ce biais substantiel s'explique par le fait que l'ordonnée à l'origine est très différente de zéro alors qu'elle n'est pas prise en compte pour construire les valeurs imputées puisque l'imputation par le ratio force la droite de régression à passer par l'origine. L'inclusion de l'ordonnée à l'origine dans le modèle d'imputation (imputation par la régression) a suffi pour réduire le biais à un niveau négligeable (environ 0.1 %).

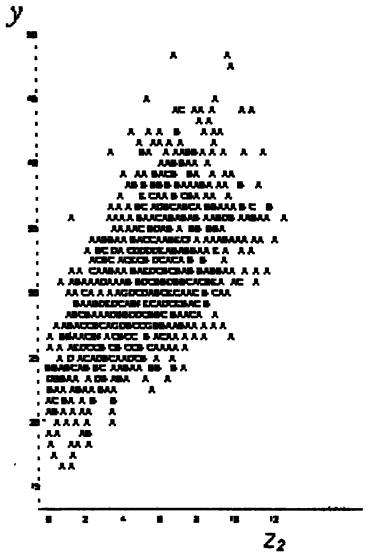


FIG 2. — Relation entre y et z_2

TABLEAU 3. – Analyse de régression

Variable	DL	Estimation des paramètres	Écart-type	t	$Pr > \|t\|$
Ordonnée à l'origine	1	22.526	0.200	112.28	0.0001
z_1	1	2.241	0.046	47.88	$< .0001$

TABLEAU 4. – Biais relatif (en %) et erreur quadratique moyenne des estimateurs

	Moyenne	Ratio	Régression
Biais Relatif (%)	6.58	-13.96	0.12
EQM	4.54×10^6	19.2×10^6	0.33×10^6

En conclusion, les exemples précédents montrent clairement que l'imputation est avant tout un exercice de modélisation. Le choix des variables auxiliaires est donc crucial. Il est important d'inclure toutes les variables auxiliaires appropriées, surtout si celles-ci sont corrélées avec la probabilité de réponse. La validation des modèles s'avèrera donc une étape importante du processus d'imputation. Celle-ci comprend, par exemple, la détection des valeurs aberrantes ou encore l'examen de certains graphiques comme :

- graphiques des résidus en fonction des valeurs prédites,
- graphiques des résidus en fonction des variables auxiliaires choisies dans le modèle,
- graphiques des résidus en fonction des variables non choisies dans le modèle.

4. Distorsion des distributions

Il est bien connu que certaines méthodes d'imputation déterministes tendent à distordre la distribution des variables d'intérêt (celles que l'on impute) alors que les méthodes d'imputation aléatoires tendent à la préserver. Dans cette section, nous illustrons ce phénomène. Considérons une population finie U de taille N et soit y une variable d'intérêt. L'objectif est d'estimer la variance dans la population de la variable y, $S_y^2 = \frac{1}{N-1}\sum_{i\in U}(y_i - \overline{Y})^2$. Pour cela, tirons un échantillon aléatoire simple sans remise, s, de taille n. En l'absence de non-réponse, la variance dans l'échantillon, $s_y^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \overline{y})^2$, est

un estimateur sans biais de S_y^2, c'est-à-dire $E_p(s_y^2) = S_y^2$. En présence de non-réponse à la variable y, on définit un estimateur imputé pour S_y^2 par

$$s_{yI}^2 = \frac{1}{n-1}\left[\sum_{i\in s_r}(y_i - \overline{y}_I)^2 + \sum_{i\in s_m}(y_i^* - \overline{y}_I)^2\right], \tag{4.1}$$

où

$$\overline{y}_I = \frac{1}{n}\left[\sum_{i\in s_r} y_i + \sum_{i\in s_m} y_i^*\right].$$

Notons que s_{yI}^2 représente simplement la variabilité des valeurs observées et des valeurs imputées dans l'échantillon. Pour illustrer le phénomène de distorsion, considérons les deux exemples suivants.

Exemple 1. — Cadre de travail BP et imputation par la régression linéaire simple

L'imputation par la régression linéaire simple utilise les valeurs imputées (2.6). La variance s_{yI}^2 en (4.1) devient alors

$$s_{yI}^2 = \frac{1}{n-1}\left[(r-1)s_{yr}^2 + (m-1)\widehat{B}_{1r}^2 s_{zm}^2 + \frac{mr}{n}\widehat{B}_{1r}^2(\overline{z}_m - \overline{z}_r)^2\right],$$

où $s_{yr}^2 = \frac{1}{r-1}\sum_{i\in s_r}(y_i - \overline{y}_r)^2$, $s_{zm}^2 = \frac{1}{m-1}\sum_{i\in s_m}(z_i - \overline{z}_m)^2$ et $\overline{z}_m = \frac{1}{m}\sum_{i\in s_m} z_i$.

Par la Proposition 1 (voir section 2.4), on a

$$E_r(s_{yI}^2 | s, r) \approx \frac{r-1}{n-1}S_y^2 + B_1^2\frac{m-1}{n-1}S_z^2, \tag{4.2}$$

où $B_1 = \dfrac{\sum_{i\in U}(z_i - \overline{Z})(y_i - \overline{Y})}{\sum_{i\in U}(z_i - \overline{Z})^2}$, $S_z^2 = \frac{1}{N-1}\sum_{i\in U}(z_i - \overline{Z})^2$ et $\overline{Z} = \frac{1}{N}\sum_{i\in U} z_i$.

Le biais relatif de s_{yI}^2 est donc donné par

$$BR(s_{yI}^2) = \frac{E(s_{yI}^2) - S_y^2}{S_y^2} \approx -\left[1 - E_p\left(\frac{r}{n}\right)\right](1 - \rho_{yz}^2) \leqslant 0, \tag{4.3}$$

où $\rho_{yz} = \dfrac{1}{N-1}\dfrac{\sum_{i\in U}(z_i - \overline{Z})(y_i - \overline{Y})}{S_y S_z}$ est le coefficient de corrélation entre les variables y et z. Le biais relatif en (4.3) est nul quand le taux de réponse espéré, $E_p(r/n)$ est égal à 1 ou quand $|\rho_{xy}| = 1$.

L'expression (4.3) montre que l'imputation par la régression ne préserve pas la variance S_y^2 de la population. L'imputation par la régression a donc tendance à sous-estimer la variabilité « naturelle » que l'on aurait observée s'il n'y avait

pas eu de non-réponse. Par conséquent, l'imputation par la régression distord la distribution de la variable d'intérêt. La distorsion relative dépend de la corrélation entre les variables y et z. Une forte corrélation assurera que la variance de la variable d'intérêt après imputation ne sera pas trop affectée. Dans le cas de l'imputation par la moyenne, l'expression (4.3) devient

$$BR(s_{yI}^2) \approx -\left[1 - E_p\left(\frac{r}{n}\right)\right] \leqslant 0. \tag{4.4}$$

L'expression (4.4) montre que dans le cas de l'imputation par la moyenne, la distorsion relative dépend uniquement du taux de réponse espéré, $E_p(r/n)$.

Exemple 2. — Cadre de travail BP et imputation par hot-deck aléatoire

Dans le cas de l'imputation par hot-deck aléatoire, les valeurs imputées sont tirées selon (2.11). On peut alors montrer (Rao, 1990) que

$$E_p E_r E_I(s_{yI}^2|s,r) \approx \frac{r-1}{r}\left[1 + \frac{r}{n(n-1)}\right] S_y^2 \approx S_y^2, \tag{4.5}$$

pour r grand. L'expression (4.5) montre que l'imputation par hot-deck aléatoire préserve la variance S_y^2.

En conclusion, nous avons montré que certaines méthodes déterministes, contrairement aux méthodes aléatoires, ne préservent pas la variance S_y^2. Notons que préserver la variance n'entraîne évidemment pas que la fonction de répartition $F(.)$ de la variable d'intérêt y est préservée. Chen, Rao et Sitter (2000) ont cependant montré que la distribution est préservée dans le cas de l'imputation par hot-deck aléatoire.

5. Estimation de la variance

Dans cette section, nous décrivons quelques méthodes permettant d'estimer correctement la variance des estimateurs en présence de valeurs imputées. Mais d'abord, énumérons quelques raisons pour lesquelles il est important d'estimer correctement la variance :

- Cela permet de mesurer la qualité (précision) des estimateurs.
- Cela aide à tirer les bonnes conclusions, plus particulièrement lorsque l'on construit des tests d'hypothèse et des intervalles de confiance.
- Cela permet d'informer correctement les utilisateurs.
- En présence de valeurs imputées, cela permet de fournir l'heure juste et de connaître l'impact de l'imputation sur la qualité des estimateurs.
- Cela permet de mieux répartir les ressources entre l'échantillon et les procédures d'imputation et de suivi.

Afin d'estimer la variance en présence de données imputées, les chercheurs ont traditionnellement utilisé l'approche deux phases. Sous cette approche,

on suppose le processus suivant :

Population $U \longrightarrow$ Échantillon s

$\longrightarrow$ Échantillon avec répondants et non-répondants (s_r, s_m)

Dans ce cas, la variance de l'estimateur imputé (2.2) est donnée par

$$V(\widehat{Y}_I - Y) = V_p E_r(\widehat{Y}_I - Y|s) + E_p V_r(\widehat{Y}_I - Y|s),$$

où $V_p(.)$ et $V_r(.)$ désignent respectivement la variance par rapport au plan de sondage et au mécanisme de non-réponse.

Dans le cas de méthodes d'imputation aléatoires, il faut également tenir compte du mécanisme d'imputation dans le calcul de variance. La variance de l'estimateur imputé (2.2) est alors donnée par

$$V(\widehat{Y}_I - Y) = V_p E_r E_I(\widehat{Y}_I - Y|s) + E_p V_r E_I(\widehat{Y}_I - Y|s) + E_p E_r V_I(\widehat{Y}_I - Y|s),$$

où $V_I(.)$ désigne la variance par rapport au mécanisme d'imputation. Sous l'approche deux phases, Rao (1990) a proposé d'estimer la variance sous le cadre de travail BP (voir section 5.1) alors que Särndal (1990) a proposé d'estimer la variance sous le cadre de travail BM (voir section 5.2).

Fay (1991) a proposé une approche alternative qui consiste à renverser l'ordre du mécanisme d'échantillonnage et du mécanisme de non-réponse (nous l'appellerons donc « approche renversée »). Sous cette approche, on suppose le processus suivant :

Population $U \longrightarrow$ Population avec répondants et non-répondants (U_r, U_m)

$\longrightarrow$ Échantillon avec répondants et non-répondants (s_r, s_m)

Dans ce cas, la variance de l'estimateur imputé (2.2) est donnée par

$$V(\widehat{Y}_I - Y) = E_r V_p(\widehat{Y}_I - Y|\mathbf{a}) + V_r E_p(\widehat{Y}_I - Y|\mathbf{a}),$$

(Shao et Steel, 1999), où $\mathbf{a} = (a_1, \ldots, a_N)'$ désigne le vecteur des variables indicatrices de réponse. Dans le cas de méthode d'imputation aléatoire, la variance de l'estimateur imputé (2.2) est donnée par

$$V(\widehat{Y}_I - Y) = E_r V_p E_I(\widehat{Y}_I - Y|\mathbf{a}) + E_r E_p V_I(\widehat{Y}_I - Y|\mathbf{a}) + V_r E_p E_I(\widehat{Y}_I - Y|\mathbf{a}).$$

Notons que, dans le cas de l'approche deux phases, les espérances et variances internes sont évaluées étant donné l'échantillon s alors que, dans le cas de l'approche renversée, les espérances et variances internes sont évaluées étant donné le vecteur des variables indicatrices de réponse $\mathbf{a}$. Dans le cas de l'imputation déterministe, estimer la variance totale revient à estimer séparément les deux composantes $V_1 = E_r V_p(\widehat{Y}_I - Y|\mathbf{a})$ et $V_2 = V_r E_p(\widehat{Y}_I - Y|\mathbf{a})$.

5.1. Deux phases : cadre de travail BP

Cette approche est due à Rao (1990) et Rao et Sitter (1995). Nous supposons ici que le mécanisme de non-réponse est uniforme. Dans le cas de l'échantillonnage aléatoire simple sans remise, il est bien connu qu'un estimateur sans biais de la variance de $\widehat{Y}$, en l'absence de non-réponse, est donné par

$$\widehat{V}(\widehat{Y}) = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) s_y^2, \tag{5.1}$$

où $s_y^2 = \dfrac{1}{n-1}\sum_{i\in s}(y_i - \overline{y})^2$ et $\overline{y} = \dfrac{1}{n}\sum_{i\in s} y_i$. En présence de non-réponse à la variable y et l'imputation par la moyenne, on a $\widehat{Y}_I = N\overline{y}_r$. Par la Proposition 1 (voir section 2.4), on détermine aisément la variance de $\widehat{Y}_I$ donnée par

$$V(\widehat{Y}_I - Y) \approx N^2 \left(\frac{1}{E_p(r)} - \frac{1}{N}\right) S_y^2. \tag{5.2}$$

Sous le cadre de travail BP, on peut montrer (en utilisant la Proposition 1) que la variabilité de la variable d'intérêt y dans l'ensemble des répondants, $s_{yr}^2 = \dfrac{1}{r-1}\sum_{i\in s_r}(y_i - \overline{y}_r)^2$, est un estimateur sans biais de S_y^2 c'est-à-dire $E_p E_r(s_{yr}^2|s,r) = S_y^2$. Un estimateur correct de la variance (5.2) est donc obtenu en remplaçant $E_p(r)$ par r et S_y^2 par s_{yr}^2 ce qui mène à

$$\widehat{V}_{cor}(\widehat{Y}_I) = N^2 \left(\frac{1}{r} - \frac{1}{N}\right) s_{yr}^2. \tag{5.3}$$

Traiter les valeurs imputées comme si elles avaient été observées revient à appliquer l'expression (5.1) aux données après imputation, auquel cas on obtiendrait un estimateur incorrect de la variance donné par

$$\widehat{V}_{inc}(\widehat{Y}_I) = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \frac{r-1}{n-1} s_{yr}^2. \tag{5.4}$$

On a alors

$$\frac{\widehat{V}_{cor}(\widehat{Y}_I)}{\widehat{V}_{inc}(\widehat{Y}_I)} = \frac{(1-r/N)}{(1-n/N)}\left(\frac{n}{r}\right)\left(\frac{n-1}{r-1}\right) \approx \left(\frac{n}{r}\right)^2, \tag{5.5}$$

si n/N est négligeable et $\dfrac{n}{r} \approx \dfrac{n-1}{r-1}$. Par exemple, si le taux de non-réponse est 50 %, le ratio (5.5) est égal à 4. Le fait de traiter les valeurs imputées comme si elles avaient été observées mène donc à un estimateur de variance quatre fois plus petit que celui qui tient compte de la non-réponse.

Nous montrons maintenant que l'imputation par hot-deck aléatoire augmente la variabilité de l'estimateur imputé. Dans ce cas, la variance de l'estimateur

imputé $\widehat{Y}_I$ en (2.2) est donnée par

$$\begin{aligned} V(\widehat{Y}_I) &= V_p E_r E_I(\widehat{Y}_I|s,r) + E_p V_r E_I(\widehat{Y}_I|s,r) + E_p E_r V_I(\widehat{Y}_I|s,r) \\ &\approx N^2 \left[\frac{1}{E_p(r)} - \frac{1}{N} + \frac{1 - E_p(r/n)}{n}\right] S_y^2. \end{aligned} \tag{5.6}$$

Désignons par $V_{hot\text{-}deck}(\widehat{Y}_I)$ la variance de l'estimateur imputé $\widehat{Y}_I$ dans le cas de l'imputation par hot-deck aléatoire et par $V_{moyenne}(\widehat{Y}_I)$ la variance de l'estimateur imputé $\widehat{Y}_I$ dans le cas de l'imputation par la moyenne. En supposant que la fraction de sondage n/N est négligeable, une comparaison de (5.2) et (5.6) montre que

$$\frac{V_{hot\text{-}deck}(\widehat{Y}_I)}{V_{moyenne}(\widehat{Y}_I)} = 1 + p(1-p) \geqslant 1, \tag{5.7}$$

où $p = E_p\left(\frac{r}{n}\right)$ est le taux de réponse espéré. Notons que le ratio (5.7) est maximum lorsque $p = 1/2$, auquel cas il vaut 1.25.

5.2. Deux phases : cadre de travail BM

Cette méthode d'estimation de la variance, développée sous le cadre de travail BM, est due à Särndal (1990, 1992). La méthode se sert de la décomposition (3.1) comme point de départ. Si le plan de sondage et le mécanisme de non-réponse sont ignorables, on peut interchanger l'ordre des espérance et alors la variance de l'estimateur imputé $\widehat{Y}_I$ est donnée par

$$\begin{aligned} V_{tot} &= V(\widehat{Y}_I - Y) = E(\widehat{Y}_I - Y)^2 = E_m E_p E_r(\widehat{Y}_I - Y)^2 = E_r E_p E_m(\widehat{Y}_I - Y)^2 \\ &= E_m V_p(\widehat{Y} - Y|s,s_r) + E_r E_p V_m(\widehat{Y}_I - \widehat{Y}|s,s_r) \\ &\qquad + 2E_m E_p\left[(\widehat{Y} - Y)E_r(\widehat{Y}_I - \widehat{Y}|s,s_r)\right] \\ &= V_{éch} + V_{imp} + 2V_{mix}, \end{aligned} \tag{5.8}$$

où $V_{éch}$ désigne la variance due à l'échantillonnage, V_{imp} désigne la variance due à la non-réponse et V_{mix} désigne un terme mixte. Un estimateur de la variance totale $V(\widehat{Y}_I - Y)$ est alors obtenu en estimant chaque composante séparément, ce qui mène à

$$\widehat{V}_{tot} = \widehat{V}_{éch} + \widehat{V}_{imp} + 2\widehat{V}_{mix}$$

L'estimation de $V_{éch}$, V_{imp} et V_{mix} peut être effectuée comme suit :

Estimation de $V_{éch}$: Cette composante représente la variance due à l'échantillonnage. Soit $\widehat{V}_{ord}$ l'estimateur de la variance obtenu en traitant les valeurs imputées comme si elles avaient été observées. Il est bien connu que, pour plusieurs méthodes d'imputation (en particulier pour les méthodes déterministes telles que l'imputation par la moyenne, le ratio ou la régression), $\widehat{V}_{ord}$

sous-estime $V_{éch}$. Afin de compenser pour cette sous-estimation, nous évaluons l'espérance suivante :

$$V_{dif} = E_m(\widehat{V}(\widehat{Y}) - \widehat{V}_{ord}|s, s_r),$$

où $\widehat{V}(\widehat{Y})$ désigne l'estimateur de la variance de $\widehat{Y}$ obtenu en cas de réponse complète.

Ensuite, nous déterminons un estimateur, $\widehat{V}_{dif}$ de V_{dif}, sans biais sous le modèle, c'est-à-dire $E_m(\widehat{V}_{dif}|s, s_r) = V_{dif}$. Cela requiert habituellement l'estimation de certains paramètres du modèle d'imputation. Alors,

$$\widehat{V}_{éch} = \widehat{V}_{ord} + \widehat{V}_{dif}$$

est sans biais sous le modèle pour $V_{éch}$.

Estimation de V_{imp} : Il suffit de déterminer $V_m(\widehat{Y}_I - \widehat{Y}|s, s_r)$ qui dépendra vraisemblablement de paramètres inconnus du modèle d'imputation. Pour obtenir $\widehat{V}_{imp}$, il suffira d'estimer correctement ces paramètres.

Estimation de V_{mix} : Il suffit de déterminer $E_m\lfloor(\widehat{Y} - Y)(\widehat{Y}_I - \widehat{Y})|s, s_r\rfloor$ qui dépendra vraisemblablement de paramètres inconnus du modèle d'imputation. Pour obtenir $\widehat{V}_{mix}$, il suffira d'estimer correctement ces paramètres.

Afin d'illustrer la méthode, nous présentons les trois composantes dans le cas de l'échantillonnage aléatoire simple sans remise et de l'imputation par la moyenne, auquel cas on a $\widehat{Y}_I = N\overline{y}_r$. Notons que l'imputation par la moyenne suggère l'emploi du modèle d'imputation

$$\begin{gathered} m \,:\, y_i = \beta + \varepsilon_i\,, \\ E_m(\varepsilon_i) = 0, \qquad E_m(\varepsilon_i\varepsilon_j) = 0,\ i \neq j, \qquad E_m(\varepsilon_i^2) = \sigma^2. \end{gathered} \tag{5.9}$$

De plus, notons que, sous le modèle (5.9), la variabilité de la variable d'intérêt y dans l'ensemble des répondants, s_{yr}^2, est un estimateur sans biais pour σ^2 c'est-à-dire $E_m(s_{yr}^2|s, s_r) = \sigma^2$.

Composante $\widehat{V}_{ord}$: La variance obtenue en traitant les valeurs imputées comme si elles avaient été observées est donnée par (5.4).

Composante $\widehat{V}_{dif}$: En vertu de (5.1) et (5.4), on a $\widehat{V}(\widehat{Y}) - \widehat{V}_{ord} = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\left[s_y^2 - \frac{r-1}{n-1}s_{yr}^2\right]$. Il s'ensuit que

$$V_{dif} = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\frac{n-r}{n-1}\sigma^2. \tag{5.10}$$

Un estimateur de V_{dif} est obtenu en estimant σ^2 par s_{yr}^2 dans (5.10), ce qui mène à

$$\widehat{V}_{dif} = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \frac{n-r}{n-1} s_{yr}^2. \tag{5.11}$$

Composante $\widehat{V}_{imp}$: On peut facilement montrer que

$$V_m(\widehat{Y}_I - \widehat{Y}|s, s_r) = N^2 \left(\frac{1}{r} - \frac{1}{n}\right) \sigma^2. \tag{5.12}$$

Un estimateur de V_{imp} est obtenu en estimant σ^2 par s_{yr}^2 dans (5.12), ce qui mène à

$$\widehat{V}_{imp} = N^2 \left(\frac{1}{r} - \frac{1}{n}\right) s_{yr}^2. \tag{5.13}$$

Composante $\widehat{V}_{mix}$: Dans le cas de l'échantillonnage aléatoire simple sans remise, on peut facilement montrer que $E_m\lfloor(\widehat{Y} - Y)(\widehat{Y}_I - \widehat{Y})|s, s_r\rfloor = 0$. Il s'ensuit que $V_{mix} = 0$.

La somme de (5.4), (5.11) et (5.13) donne l'estimateur de la variance totale, $\widehat{V}_{tot}$. On obtient

$$\widehat{V}_{tot} = N^2 \left(\frac{1}{r} - \frac{1}{N}\right) s_{yr}^2. \tag{5.14}$$

Notons que l'estimateur (5.14) coïncide avec l'estimateur correct de la variance (5.3) obtenu sous le cadre de travail BP (section 5.1).

Caractéristiques de la méthode :

(i) La méthode développée par Särndal peut être utilisée pour un ensemble de méthodes d'imputation telles que : imputation par la moyenne, par le ratio, par la régression, historique et par la régression aléatoire. Deville et Särndal (1994) ont généralisé la méthode au cas de plans de sondage arbitraires et de l'imputation par la régression.

(ii) Contrairement au jackknife et au bootstrap, la méthode n'est pas intensive au point de vue informatique.

(iii) La méthode peut être utilisée pour des estimateurs non-linéaires (tels un ratio de deux totaux) et pour des paramètres complexes (tel que le coefficient de corrélation entre deux variables).

(iv) L'application de la méthode pour des quantiles n'a pas encore été étudiée.

(v) Puisque les composantes $\widehat{V}_{dif}$, $\widehat{V}_{imp}$ et $\widehat{V}_{mix}$ sont développées en utilisant un modèle d'imputation, leur validité dépend de la validité de ce modèle. Par exemple, dans le cas d'un plan à deux degrés, le modèle d'imputation (5.9) ne tient pas compte de la corrélation intra-grappe. Dans ce cas, l'utilisation du modèle (5.9) dans le développement des estimateurs de variance peut mener à des estimateurs de variance biaisés tel qu'illustré dans Haziza et Rao (2003a) lorsque la corrélation intra-grappe est

non négligeable. Un modèle d'imputation plus approprié est le modèle ANOVA à un effet aléatoire.

(vi) Dans le cas de plans de sondage auto-pondérés, la composante V_{mix} est égale à zéro dans le cas de l'imputation par la régression (qui englobe l'imputation par la moyenne et par le ratio comme cas particuliers) et l'imputation par hot-deck aléatoire. Brick, Kalton et Kim (2004) ont montré que dans le contexte d'un plan de sondage stratifié aléatoire simple et de l'imputation par hot-deck aléatoire, la contribution (positive ou négative) de V_{mix} à la variance totale V_{tot} peut être aussi élevée que 45 % en valeur absolue, variant selon les taux de sondage et les taux de réponse observés dans les strates. Il est donc préférable d'inclure la composante V_{mix} dans le cas de plans de sondage à probabilités inégales. De plus, dans le cas de l'imputation historique, la contribution de la composante V_{mix} à la variance totale est importante et toujours négative (Beaumont, Haziza et Rancourt, 2005).

(vii) Le développement de la composante V_{dif} requiert habituellement de simples mais fastidieuses manipulations algébriques. Dans le cas de plans de sondage à un seul degré et de méthodes d'imputation aléatoires, l'importance de V_{dif} tend à être très petite et celle-ci peut donc être omise. En effet, les méthodes d'imputation aléatoires préservent la variabilité des valeurs imputées qui correspondent aux valeurs pour les répondants; il s'ensuit que la composante $V_{éch}$ peut être directement estimée par l'estimateur naïf de la variance $\widehat{V}_{ord}$. Par conséquent, lorsque qu'une méthode déterministe est utilisée (par exemple, moyenne ou ratio), la variance due à l'échantillonnage peut être correctement estimée à l'aide de la formule naïve si des résidus aléatoires ont été ajoutés aux valeurs imputées pour le calcul de la variance.

5.3. L'approche renversée

Pour illustrer cette méthode, considérons le cas d'un échantillon aléatoire simple sans remise, s, de taille n, tiré d'une population U de taille N. Dans le cas de l'imputation par la moyenne, on a $\widehat{Y}_I = N\overline{y}_r$.

Estimation de $V_1 = E_r V_p(\widehat{Y}_I - Y|\mathbf{a})$: Il suffit d'estimer $V_p(\widehat{Y}_I - Y|\mathbf{a})$. Pour cela, écrivons d'abord l'estimateur imputé en fonction des variables indicatrices de réponse a_i. On a alors

$$\widehat{Y}_I = N\frac{\sum_{i\in s} a_i y_i}{\sum_{i\in s} a_i}.$$

On est donc ramené à estimer la variance due à l'échantillonnage pour un ratio de deux totaux, $\sum_{i\in s} t_i$ et $\sum_{i\in s} a_i$ où $t_i = a_i y_i$, ce que l'on sait faire. Soit $\widehat{V}_1$ un estimateur de V_1. On peut, par exemple, utiliser la linéarisation en série

de Taylor. Dans ce cas, on obtient $\widehat{Y}_I \approx \frac{N}{n}\sum_{i\in s}\widehat{\xi}_i$ où $\widehat{\xi}_i = \frac{N}{\sum_{i\in s} a_i} a_i[y_i - \overline{y}_r]$ avec $\overline{y}_r = \sum_{i\in s} a_i y_i / \sum_{i\in s} a_i$. La composante $\widehat{V}_1$ est obtenue en remplaçant y_i dans (5.1) par $\widehat{\xi}_i$ ce qui mène à

$$\widehat{V}_1 \approx N^2\left(1 - \frac{n}{N}\right)\frac{s_{yr}^2}{r}. \tag{5.15}$$

Estimation de $V_2 = V_r E_p(\widehat{Y}_I - Y|\mathbf{a})$ sous le cadre de travail BP : D'abord, notons que $E_p(\widehat{Y}_I - Y|\mathbf{a}) \approx N\sum_{i\in U} a_i y_i / \sum_{i\in U} a_i$. De plus, notons que sous le cadre de travail BP, on a $V_r(a_i) = p(1-p)$. On peut alors approcher $V_r E_p(\widehat{Y}_I - Y|\mathbf{a})$ par linéarisation en série de Taylor, ce qui mène à

$$V_r E_p(\widehat{Y}_I - Y|\mathbf{a}) \approx p(1-p)\frac{N^2}{E_r\left(\sum_{i\in U} a_i\right)^2}\sum_{i\in U}(y_i - \overline{Y})^2. \tag{5.16}$$

Un estimateur de $V_r E_p(\widehat{Y}_I - Y|\mathbf{a})$ est obtenu en remplaçant les quantités inconnues dans (5.16) par des estimateurs appropriés, ce qui mène à

$$\widehat{V}_{2BP} = N^2\left(\frac{n}{N} - \frac{r}{N}\right)\frac{s_{yr}^2}{r}. \tag{5.17}$$

Estimation de $V_2 = V_r E_p(\widehat{Y}_I - Y|\mathbf{a})$ sous le cadre de travail BM : D'abord, notons que nous faisons appel au modèle d'imputation (5.9). On a alors

$$\begin{aligned} V_r E_p(\widehat{Y}_I - Y|\mathbf{a}) &= E_r V_m E_p(\widehat{Y}_I - Y|\mathbf{a}) + V_r E_m E_p(\widehat{Y}_I - Y|\mathbf{a}) \\ &= E_r V_m\left(\textstyle\sum_{i\in U} c_i y_i|\mathbf{a}\right), \end{aligned}$$

puisque $E_p E_m(\widehat{Y}_I - Y|\mathbf{a}) = 0$, où $c_i = \left(N a_i / \sum_{i\in U} a_i\right) - 1$. On a alors

$$V_r E_p(\widehat{Y}_I - Y|\mathbf{a}) = E_r\left[\sum_{i\in U} c_i^2 V_m(y_i)\right] = \sigma^2 E_r\left[\sum_{i\in U} c_i^2\right]. \tag{5.18}$$

Un estimateur de $V_r E_p(\widehat{Y}_I - Y|\mathbf{a})$ est obtenu en remplaçant les quantités inconnues dans (5.18) par des estimateurs appropriés, ce qui mène à

$$\widehat{V}_{2BM} = N^2\left(\frac{n}{N} - \frac{r}{N}\right)\frac{s_{yr}^2}{r}. \tag{5.19}$$

Notons que, dans le cas de l'imputation par la moyenne, $\widehat{V}_{2BP}$ en (5.17) coïncide avec $\widehat{V}_{2BM}$ en (5.19). En général cependant, les estimateurs de $V_r E_p(\widehat{Y}_I - Y|\mathbf{a})$ obtenus sous le cadre de travail BP sont différents de ceux obtenus sous le cadre de travail BM. La somme de (5.15) et (5.17) ou (5.19) donne un estimateur de la variance totale que l'on désigne par $\widehat{V}_{tot}^{AR}$. On obtient

$$\widehat{V}_{tot}^{AR} = N^2 \left(\frac{1}{r} - \frac{1}{N}\right) s_{yr}^2$$

qui coïncide avec les estimateurs (5.3) et (5.14). Finalement, notons que le ratio

$$\frac{\widehat{V}_{2BP}}{\widehat{V}_I} = \frac{n}{N} \frac{\left(1 - \frac{r}{n}\right)}{\left(1 - \frac{n}{N}\right)}$$

est proche de 0 lorsque $n/N \approx 0$.

Caractéristiques de l'approche renversée :

(i) L'approche renversée permet d'obtenir des estimateurs de variance sous les cadres de travail BP et BM.

(ii) L'estimateur $\widehat{V}_1$ de V_1 ne dépend pas du mécanisme de non-réponse et/ou du modèle d'imputation. L'estimateur $\widehat{V}_1$ est donc robuste à une mauvaise spécification du modèle, mais la composante $\widehat{V}_2$ dépend du mécanisme de non-réponse et/ou du modèle d'imputation.

(iii) L'estimation de V_1 peut être effectuée en utilisant les méthodes connues telles la linéarisation en série de Taylor, le jackknife, le bootstrap, etc. En fait, le jackknife ajusté de Rao-Shao (Rao et Shao, 1992), présenté dans la section 5.4, et le bootstrap de Shao-Sitter (Shao et Sitter, 1996), présenté dans la section 5.5, trouvent leur justification dans l'approche renversée puisque ces deux techniques permettent d'obtenir un estimateur de V_1. Elles ne permettent toutefois pas d'obtenir un estimateur de la composante V_2.

(iv) Le ratio $\dfrac{V_2}{V_1}$ est d'ordre $O(n/N)$. Donc, lorsque la fraction de sondage n/N est négligeable, la composante V_2 est négligeable par rapport à la composante V_1, auquel cas on peut omettre la composante V_2 dans le calcul de la variance et on a $\widehat{V}_{tot}^{AR} \approx \widehat{V}_1$.

5.4. Le jackknife

Le jackknife proposé par Quenouille (1949) est une méthode de ré-échantillonnage qui permet d'estimer la variance dans le cas de plans et/ou de paramètres complexes. Supposons que l'on veuille estimer un paramètre θ. Pour cela, on tire un échantillon aléatoire, s, de taille n, selon un plan de sondage $p(.)$. Soit $\widehat{\theta}$ un estimateur sans biais (ou approximativement sans biais) de θ obtenu en utilisant les unités observées dans s. En l'absence de non-réponse, le jackknife fonctionne comme suit :

(1) Enlever une unité (ou groupe d'unités).

(2) Ajuster les poids de sondages.

(3) Calculer l'estimateur $\widehat{\theta}$ avec les poids ajustés.

(4) Replacer l'unité enlevée à l'étape (i), enlever la prochaine unité.

(5) Répéter les étapes (1)-(4) jusqu'à ce que toutes les unités aient été enlevées chacune à leur tour.

La variance jackknife de $\widehat{\theta}$ est donnée par

$$\widehat{V}_J(\widehat{\theta}) = \frac{n-1}{n}\sum_{j\in s}(\widehat{\theta}_{(j)} - \widehat{\theta})^2,$$

où $\widehat{\theta}_{(j)}$ est calculé de la même manière que $\widehat{\theta}$ lorsque la j^{e} unité a été enlevée, $j = 1, \ldots, n$. Notons que $\widehat{\theta}_{(j)}$ est calculé avec les poids de sondage ajustés $w_{i(j)}$ définis par

$$w_{i(j)} = \begin{cases} \dfrac{n}{n-1}w_i & \text{si } i \neq j \\ 0 & \text{si } i = j \end{cases}$$

En présence de non-réponse, le jackknife, tel que décrit précédemment (que nous appellerons jackknife « traditionnel »), mène à une sous estimation de la variance de l'estimateur imputé. Considérons le cas d'un échantillon aléatoire simple sans remise, s, de taille n, tiré d'une population U de taille N. La variance de l'estimateur imputé (2.2) obtenue par jackknife « traditionnel » est donnée par

$$\widehat{V}_J(\widehat{Y}_I) = \frac{n-1}{n}\sum_{j\in s}(\widehat{Y}_{I(j)} - \widehat{Y}_I)^2 = N^2\frac{s^2_{yI}}{n}, \tag{5.20}$$

où

$$\widehat{Y}_{I(j)} = \begin{cases} \widehat{Y}_I - y_j & \text{si } j \in s_r \\ \widehat{Y}_I - y_j^* & \text{si } j \in s_m \end{cases}$$

et s^2_{yI} est donné par (4.1).

Dans le cas de l'imputation par la moyenne, on a $\widehat{Y}_I = N\overline{y}_r$ auquel cas l'estimateur de variance (5.20) devient

$$\widehat{V}_J(\widehat{Y}_I) = N^2\frac{r-1}{n-1}\frac{s^2_{yr}}{n},$$

qui coïncide avec l'estimateur incorrect de la variance (5.4) lorsque la fraction de sondage n/N est négligeable. Utiliser le jackknife « traditionnel » revient donc à traiter les valeurs imputées comme si elles avaient été observées. C'est pourquoi, Rao et Shao (1992) ont proposé un jackknife ajusté qui mène à un estimateur de variance qui tient compte de la non-réponse. Le jackknife ajusté se calcule de la même façon que le jackknife « traditionnel »

sauf que, lorsque qu'une unité répondante, $j \in s_r$, est enlevée, les valeurs imputées y_i^* sont ajustées par une quantité qui tient compte de l'impact de l'élimination de j sur les valeurs imputées. Lorsque qu'une unité non-répondante, $j \in s_m$, est enlevée, les valeurs imputées y_i^* sont laissées telles quelles. Dans le cas de méthodes d'imputation déterministes, l'ajustement de Rao-Shao est équivalent à réimputer dans chacune des répliques. Par exemple, dans le cas de l'imputation par la moyenne, les valeurs imputées ajustées, y_i^{*a}, sont données par

$$y_i^{*a} = \begin{cases} \overline{y}_{r(j)} & \text{si } j \in s_r \\ \overline{y}_r & \text{si } j \in s_m \end{cases}$$

où $\overline{y}_{r(j)} = \dfrac{1}{r-1} \sum_{i \neq j} y_i$. L'utilisation des valeurs imputées ajustées mène à l'estimateur jackknife de Rao-Shao, donné par

$$\widehat{V}_{JRS}(\widehat{Y}_I) = N^2 \frac{s_{yr}^2}{n},$$

qui coïncide avec l'estimateur correct de la variance (5.3) lorsque la fraction de sondage n/N est négligeable. Il est important de noter que l'estimateur de Rao-Shao est un estimateur de la composante V_1 de l'approche renversée. Donc, si la fraction de sondage n/N est importante, l'estimateur jackknife de Rao-Shao est biaisé puisqu'il n'inclut pas la composante non-négligeable V_2 de l'approche renversée.

Caractéristiques du jackknife ajusté

(i) Le jackknife ajusté peut être appliqué à plusieurs méthodes d'imputation (hot-deck aléatoire, moyenne, ratio, régression, etc.).

(ii) Le jackknife ajusté peut être utilisé pour des plans complexes (stratifié à degrés multiples).

(iii) La méthode est intensive du point de vue informatique. Un programme efficace permettra toutefois de réduire le temps d'exécution de manière considérable.

(iv) La méthode ne peut être utilisée dans le cas de paramètres « non-lisses » (quantiles, etc.).

(v) La méthode suppose que l'échantillon est tiré avec remise. Les résultats seront donc valides si la fraction de sondage n/N est petite.

(vi) Dans le cas de l'imputation par PPV, Chen et Shao (2001) ont montré que l'utilisation de la technique de Rao-Shao mène à une sur-estimation importante de la variance de l'estimateur. Pour contrer ce problème, ils ont proposé un jackknife « partiellement ajusté ».

5.5. Le bootstrap

Le boostrap (Efron, 1979) est, comme le jackknife, une autre méthode de ré-échantillonnage qui peut être utilisée afin d'estimer la variance de paramètres complexes. L'utilisation du bootstrap en présence de valeurs imputées mène

généralement à une sous-estimation de la variance de l'estimateur imputé. L'adaptation du bootstrap en présence de valeurs imputées a été proposée par Shao et Sitter (1996). Supposons que l'on veuille estimer un paramètre θ. Pour cela, on tire un échantillon aléatoire simple sans remise, s, de taille n, d'une population U de taille N. Soit $\widehat{\theta}_I$ un estimateur imputé basé sur les valeurs imputées et observées. Le bootstrap de Shao et Sitter peut être décrit comme suit :

(1) Tirer un échantillon bootstrap s^* (échantillon aléatoire simple avec remise) de taille $n^* = n - 1$ de l'échantillon original s après imputation.

(2) Soit a_i^* l'indicateur de réponse pour l'unité i dans s^*. Soit $s_r^* = \{i \in s^* : a_i^* = 1\}$ et $s_m^* = \{i \in s^* : a_i^* = 0\}$. Réimputer les non-répondants dans s^* (i.e., les unités dans s_m^*) en utilisant la même méthode d'imputation qui a été utilisée pour obtenir l'estimateur ponctuel $\widehat{\theta}_I$, au moyen des unités répondantes dans s^* (i.e., les unités dans s_r^*).

(3) Calculer l'estimateur imputé $\widehat{\theta}_I^*$ dans l'échantillon bootstrap s^* de la même façon que l'on a calculé l'estimateur imputé $\widehat{\theta}_I$.

(4) Répéter les étapes (1)-(3) B fois.

L'estimateur de variance bootstrap de $\widehat{\theta}_I$ est donné par

$$\widehat{V}_B(\widehat{\theta}_I) = \frac{1}{B-1}\sum_{b=1}^{B}(\theta_{I(b)}^* - \widehat{\bar{\theta}}_I)^2, \tag{5.21}$$

où

$$\widehat{\bar{\theta}}_I = \frac{1}{B}\sum_{b=1}^{B}\theta_{I(b)}^*.$$

Notons que l'estimateur (5.21) est un estimateur de la composante V_1 de l'approche renversée. Donc, si la fraction de sondage n/N est importante, l'estimateur (5.21) est biaisé puisqu'il n'inclut pas la composante non-négligeable V_2 de l'approche renversée.

Caractéristiques du bootstrap

(i) Le bootstrap peut être appliqué à plusieurs méthodes d'imputation (hot-deck aléatoire, moyenne, ratio, régression, etc.).

(ii) Le bootstrap peut être utilisé pour des plans complexes (stratifié à degrés multiples) mais certaines études théoriques et empiriques restent à faire.

(iii) La méthode est très intensive du point de vue informatique.

(iv) Contrairement au jackknife, le bootstrap peut être utilisé dans le cas de paramètres « non-lisses » (quantiles, etc.)

(v) La méthode suppose que l'échantillon est tiré avec remise. Les résultats seront donc valides si la fraction de sondage n/N est petite.

(vi) Lorsque la taille n de l'échantillon est petite, la procédure proposée par Shao et Sitter peut mener à des estimateurs de variance considérablement

biaisés. Saigo, Shao et Sitter (2001) ont modifié la procédure de Shao-Sitter pour contrer ce problème.

6. Distorsion des relations entre variables

Jusqu'à maintenant, nous avons discuté de l'inférence en présence de valeurs imputées pour des paramètres simples tels un total ou une moyenne. En pratique, il est souvent requis d'estimer des paramètres plus complexes tels la moyenne d'un domaine, un coefficient de régression, un coefficient de corrélation, etc. La moyenne d'un domaine d, $\overline{Y}_d$ peut s'écrire comme

$$\overline{Y}_d = \sum_{i \in U} x_i y_i / \sum_{i \in U} x_i, \tag{6.1}$$

où x_i est un indicateur de domaine tel que $x_i = 1$ si l'unité i appartient au domaine d et $x_i = 0$, sinon. Un coefficient de régression, $\mathbf{B_N}$ peut être exprimé comme

$$\mathbf{B_N} = \left(\sum_{i \in U} \mathbf{x_i x_i'} \right)^{-1} \sum_{i \in U} \mathbf{x_i} y_i, \tag{6.2}$$

où $\mathbf{x}$ est un vecteur de variables auxiliaires disponible pour toutes les unités dans l'échantillon. Notons que la moyenne d'un domaine $\overline{Y}_d$ en (6.1) est un cas particulier de (6.2) quand x_i est un indicateur de domaine. Un coefficient de corrélation entre deux variables x et y est donné par

$$\rho_{xy} = \frac{1}{N-1} \left[\frac{\sum_{i \in U} x_i y_i - N \overline{XY}}{S_x S_y} \right]. \tag{6.3}$$

Considérons le cas du coefficient de corrélation (6.3). En présence de valeurs imputées, l'obtention d'un estimateur imputé approximativement sans biais pour ρ_{xy} passe par l'obtention d'un estimateur approximativement sans biais pour chacune des composantes dans (6.3). En présence de valeurs imputées, ceci peut facilement être accompli pour les paramètres $\overline{X}$, $\overline{Y}$, S_x et S_y au moyen des méthodes d'imputation décrites dans la section 2.2. L'estimation de la composante $\sum_{i \in U} x_i y_i$ s'avère cependant problématique. Notons que cette composante est commune à tous les paramètres (6.1)-(6.3). L'obtention d'un estimateur sans biais de $\sum_{i \in U} x_i y_i$ n'est pas aisée puisque cette composante est, en quelque sorte, une mesure de la relation entre les variables x et y. Or, l'imputation a comme effet de modifier les relations entre les variables, ce qui explique la difficulté rencontrée. La littérature à propos de l'inférence pour des paramètres complexes en présence de valeurs imputées est peu abondante (on peut citer Santos, 1981; Shao et Wang, 2002; Skinner et Rao, 2002; Haziza

et Rao, 2004, 2005). Un estimateur imputé approximativement sans biais du terme $\sum_{i \in U} x_i y_i$ peut être obtenu d'au moins deux façons :

(1) utilisation d'une méthode d'imputation sophistiquée qui mène directement à un estimateur approximativement sans biais.

(2) utilisation d'une méthode d'imputation simple (moyenne, ratio, hot-deck aléatoire, etc.) et d'un estimateur plus sophistiqué.

Le problème de l'estimation pour des domaines est considéré dans la section 6.1. Dans la section 6.2, nous mentionnons quelques articles traitant du cas d'un coefficient de corrélation. Finalement, l'utilisation de méthodes d'imputation pondérées par opposition à l'utilisation de méthodes d'imputation non-pondérées fera l'objet de la section 6.3. En effet, nous montrerons que la non-utilisation des poids de sondage dans la construction des valeurs imputées peut mener à un problème de distorsion de la relation entre la variable d'intérêt et la probabilité d'inclusion dans l'échantillon.

6.1. Les domaines

En pratique, des estimations sont très fréquemment requises pour des sous-groupes (domaines) de la population. Par exemple, pour l'Enquête sur la Population Active Canadienne, des estimations du taux de chomâge sont produites au niveau national mais également par province, par industrie ou par groupe d'âge-sexe. Cependant, les domaines d'intérêt ne sont pas toujours connus à l'étape de l'imputation. En effet, une fois le fichier complété après imputation, il est souvent envoyé à de nombreux chargés d'étude qui s'intéressent potentiellement à des domaines différents. Ces domaines ne sont donc pas toujours pris en compte à l'étape de l'imputation. L'estimateur imputé de la moyenne $\overline{Y}_d$ est donné par

$$\overline{y}_{dI} = \frac{1}{\sum_{i \in s} w_i x_i} \left[\sum_{i \in s_r} w_i x_i y_i + \sum_{i \in s_m} w_i x_i y_i^* \right]. \tag{6.4}$$

Nous supposons ici que l'indicateur de domaine x_i est connu pour toutes les unités échantillonnées. Pour illustrer la problématique, nous avons effectué une étude par simulation. Nous avons créé une population de taille $N = 11\,270$ individus à partir d'un échantillon de l'Enquête sur la Population Active Canadienne pour le mois de Janvier 2001. La population créée contient deux variables : *Revenu hebdomadaire* (RH) et *âge de l'individu*. La moyenne de la variable RH dans la population est 555 dollars. Le tableau 5 exhibe la moyenne de la variable RH par groupe d'âge dans la population. Un simple coup d'œil au tableau 5 révèle qu'il y a une relation entre les variables RH et âge.

TABLEAU 5. – Moyenne du *Revenu Hebdomadaire* par groupe d'âge

Âge	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60+
RH	139.7	343.6	513.9	587.2	625.6	661.5	704.5	692.4	629.6	515.4

L'objectif est d'estimer la moyenne de la variable RH pour deux domaines d'intérêt : le groupe des 15-19 ans et celui des 30-34 ans. Notons que le premier domaine est celui pour lequel la moyenne est la plus éloignée de la moyenne de la population alors que le deuxième est celui pour lequel la moyenne en est le plus près. De cette population, $R = 5\,000$ échantillons aléatoires simples sans remise, de taille $n = 500$, ont été tirés. Dans chaque échantillon, de la non-réponse à la variable RH a été générée selon un mécanisme de non-réponse uniforme. Le taux de réponse a été fixé à 70 %. Pour imputer les valeurs manquantes, nous avons utilisé :

(a) $y_i^* = \overline{y}_r$ la moyenne globale des répondants qui ne tient pas compte des domaines d'intérêt.

(b) $y_i^* = \overline{y}_{dr}$ la moyenne des répondants à l'intérieur du domaine d'intérêt (qui tient donc compte du domaine en question).

TABLEAU 6. – Biais relatif (%) de l'estimateur imputé (6.4)

	$y_i^* = \overline{y}_{dr}$	$y_i^* = \overline{y}_r$
15-19	0.5	88
30-34	0.4	−2.5

Le tableau 6 exhibe le biais relatif de l'estimateur imputé (6.4). Les résultats montrent que, lorsque l'on tient compte des domaines d'intérêt dans le modèle d'imputation, le biais relatif des estimateurs imputés est négligeable (0.5 % pour les 15-19 ans et 0.4 % pour les 30-34 ans). Par contre, lorsque l'on ne tient pas compte des domaines d'intérêt dans le modèle d'imputation, le biais relatif des estimateurs imputés est considérable pour le domaine des 15-19 ans (environ 88 %) alors qu'il est petit mais pas négligeable pour le domaine des 30-34 ans. Il est donc clair que, dans ce cas, le biais relatif des estimateurs imputés dépendra de la différence entre la moyenne du domaine $\overline{Y}_d$ et la moyenne globale de la population $\overline{Y}$ ce qui est confirmé par le résultat suivant :

PROPOSITION 2. — *Sous le cadre de travail BP et l'imputation par la moyenne, $y_i^* = \overline{y}_r$ le biais de l'estimateur imputé (6.4) est donné par*

$$\text{Biais}(\overline{y}_{dI}) = E_p E_r(\overline{y}_{dI}|s) - \overline{Y}_d \approx (1-p)(\overline{Y} - \overline{Y}_d). \tag{6.5}$$

Le biais en (6.5) est nul dans le cas de réponse complète ($p = 1$) ou lorsque $\overline{Y} = \overline{Y}_d$. Si l'on ne tient pas compte des domaines à l'étape de l'imputation, il est quand même possible d'obtenir un estimateur approximativement sans

biais de la moyenne $\overline{Y}_d$. En effet, Haziza et Rao (2005) ont proposé une correction pour éliminer le biais, ce qui mène à l'estimateur ajusté

$$\overline{y}_I^a = \widehat{p}^{-1}\overline{y}_{dI} + (1 - \widehat{p}^{-1})\overline{y}_I \tag{6.6}$$

où $\widehat{p}$ est un estimateur de la probabilité de réponse p, $\overline{y}_{dI}$ est donné par (6.4) et $\overline{y}_I$ est donné par (2.2). L'estimateur ajusté (6.6) est approximativement sans biais sous les cadres de travail BP et BM avec le modèle d'imputation (5.9). Cet estimateur est donc robuste en ce sens qu'il est approximativement sans biais sous les deux cadres de travail. Haziza et Rao (2005) ont généralisé ce résultat au cas de l'imputation par la régression et de l'imputation par la régression aléatoire.

6.2. Coefficient de corrélation

L'estimation d'un coefficient de corrélation est relativement plus complexe. Dans ce cas, les deux variables x et y sont susceptibles d'être manquantes. Shao et Wang (2002) ont proposé une procédure d'imputation jointe qui mène à un estimateur approximativement sans biais. Skinner et Rao (2002) ont proposé un estimateur ajusté dans le cas d'échantillonnage aléatoire simple sans remise, le cadre de travail BM et imputation par hot-deck aléatoire où l'ensemble des donneurs est restreint aux unités qui ont répondu aux deux variables x et y. Haziza et Rao (2004) ont généralisé les résultats de Skinner et Rao (2002) au cas de plans de sondage stratifiés à degrés multiples, les cadres de travail BP et BM et certaines variantes de l'imputation par hot-deck aléatoire.

6.3. Imputation pondérée et imputation non-pondérée

Dans le cas de plans de sondage à probabilités inégales (plan stratifié à degrés multiples, plan proportionnel à la taille, etc.), il est possible d'utiliser soit une méthode d'imputation pondérée, soit une méthode d'imputation non-pondérée. L'imputation pondérée, contrairement à l'imputation non-pondérée, tient compte d'une partie de l'information du plan de sondage (les poids du sondage) dans la construction des valeurs imputées. Bien sûr, les méthodes pondérées et les méthodes non-pondérées mènent à des résultats identiques dans le cas d'un plan de sondage auto-pondéré (par exemple, échantillonnage aléatoire simple sans remise). En pratique, on utilise fréquemment des méthodes d'imputation non-pondérées, et ce, même dans le cas de plans de sondage à probabilités inégales. Les méthodes non-pondérées sont généralement plus attrayantes pour l'utilisateur car elles sont plus simples. Dans ce cas cependant, l'estimateur imputé sera vraisemblablement biaisé sous le cadre de travail BP. Pour illustrer la problématique, considérons le cas d'un échantillon aléatoire, s, de taille n, tiré d'une population U de taille N selon un plan de sondage $p(.)$. Nous étudions le biais de l'estimateur imputé (2.2) dans le cas de l'imputation par hot-deck aléatoire pondéré (HDP) et l'imputation par hot-deck aléatoire non-pondéré (HDNP). L'imputation HDNP utilise les valeurs imputées (2.11) avec $w_i = 1$ alors que l'imputation HDP utilise les valeurs imputées (2.11).

On peut montrer que dans le cas d'imputation HDP, l'estimateur imputé (2.2) est approximativement sans biais sous les cadres de travail BP et BM. Dans le cas d'imputation HDNP cependant, l'estimateur imputé (2.2) est biaisé sous le cadre de travail BP. Le biais relatif est donné par

$$\mathrm{BR}(\overline{y}_I) = \frac{E_p E_r(\widehat{Y}_I|s) - Y}{Y} \approx (1-p) C_y C_\pi \rho_{\pi y}, \tag{6.7}$$

où $C_y = \frac{S_y}{\overline{Y}}$ et $C_\pi = \frac{S_\pi}{\overline{\Pi}}$ désignent respectivement les coefficients de variation de la variable d'intérêt y et de la probabilité d'inclusion π, et $\rho_{\pi y}$ désigne le coefficient de corrélation entre la variable d'intérêt y et de la probabilité d'inclusion π. Si $C_y > 0$ le biais relatif en (6.7) est nul si

(a) $p = 1$ (cas de réponse complète)

ou

(b) $C_\pi = 0$, (cas d'un plan de sondage autopondéré)

ou

(c) $\rho_{\pi y} = 0$.

Dans le cas d'un plan de sondage à probabilités inégales, le biais relatif est nul lorsque la corrélation entre la variable d'intérêt et la probabilité d'inclusion $\rho_{\pi y}$ est nulle. Dans ce cas, l'inclusion des poids de sondage dans la construction des valeurs imputées (ou l'inclusion des poids de sondage dans le modèle d'imputation) est superflue. Même lorsque la corrélation $\rho_{\pi y}$ est grande et que l'on utilise une méthode d'imputation non-pondérée, il est possible d'obtenir un estimateur sans biais du total Y (Haziza et Rao, 2003b).

7. Classes d'imputation

En pratique, il est coutume de préalablement former des classes d'imputation et d'imputer à l'intérieur de chaque classe. L'objectif premier visé par la formation des classes est la réduction du biais dû à la non-réponse. La construction des classes repose donc sur l'utilisation d'une information auxiliaire appropriée. Au lieu de former des classes, il est toujours possible d'imputer directement des valeurs à partir d'un modèle de régression en utilisant la même information auxiliaire. Cependant, il y a au moins deux raisons motivant l'utilisation de classes : (a) c'est plus pratique quand il faut imputer plusieurs variables à la fois et (b) les classes apportent une certaine robustesse par rapport à l'utilisation de l'imputation par la régression si le modèle d'imputation est mal spécifié.

7.1. Justification théorique

Nous donnons d'abord une justification théorique pour la formation des classes d'imputation. Considérons une population finie de taille N. L'objectif est d'estimer la moyenne de la population $\overline{Y} = \frac{1}{N}\sum_{i\in U} y_i$. Pour cela, nous tirons un échantillon aléatoire, s, de taille n, selon un plan de sondage $p(.)$. Supposons que les unités répondent à l'item y indépendamment les unes des autres et que la probabilité de réponse pour l'unité i est p_i, $i = 1, \ldots, N$. Un estimateur imputé pour $\overline{Y}$ obtenu à partir d'une seule classe d'imputation (autrement dit, l'échantillon s), est défini par

$$\overline{y}_{I,1} = \frac{1}{\sum_{i\in s} w_i y_i}\left[\sum_{i\in s_r} w_i y_i + \sum_{i\in s_m} w_i y_i^*\right]. \tag{7.1}$$

Dans le cas de l'imputation par hot-deck aléatoire, l'estimateur imputé $\overline{y}_{I,1}$ est biaisé et le biais est donné par

$$\text{Biais}(\overline{y}_{I,1}) = E(\overline{y}_{I,1}) - \overline{Y} \approx \frac{1}{N\overline{P}}\sum_{i\in U}(p_i - \overline{P})(y_i - \overline{Y}). \tag{7.2}$$

où $\overline{P} = \frac{1}{N}\sum_{i\in U} p_i$ est la moyenne des probabilités dans la population. Le biais (7.2) est égal à 0 si la covariance dans la population entre les variables p et y est zéro, ce qui est le cas, par exemple, si toutes les unités dans la population ont la même probabilité de répondre (mécanisme de non-réponse uniforme) et/ou si la valeur de la variable d'intérêt est la même pour toutes les unités dans la population. Ces deux exigences sont bien sûr très rarement satisfaites en pratique. Pour cette raison, on dira de $\overline{y}_{I,1}$, qu'il est un estimateur «non-ajusté». Pour réduire le biais dû à la non-réponse, il est d'usage de diviser la population en C classes d'imputation disjointes U_g de taille N_g; $\left(\bigcup_{g=1}^{C} U_g = U, \sum_{g=1}^{C} N_g = N\right)$, ce qui mène à la partition correspondante dans l'échantillon s en C classes $s_g = s \cap U_g$ de taille n_g; $\left(\bigcup_{g=1}^{C} s_g = s, \sum_{g=1}^{C} n_g = n\right)$. On impute alors de façon indépendante à l'intérieur de chaque classe par hot-deck aléatoire, ce qui mène à l'estimateur imputé «ajusté» obtenu à partir de C classes

$$\overline{y}_{I,C} = \sum_{g=1}^{C} w_g' \overline{y}_g, \tag{7.3}$$

où $w'_g = \dfrac{\sum_{i \in s_g} w_i}{\sum_{i \in s} w_i}$ est une mesure de la taille relative de la classe g et

$$\overline{y}_g = \frac{1}{\sum_{i \in s_g} w_i} \left[\sum_{i \in s_{r_g}} w_i y_i + \sum_{i \in s_{m_g}} w_i y_i^* \right] \tag{7.4}$$

désigne l'estimateur imputé pour la classe g, $g = 1, \ldots, C$, où $s_g = s \cap U_g$ et s_{r_g} et s_{m_g} désignent les ensembles de répondants et de non-répondants dans la classe g, respectivement. Dans le cas de l'imputation par hot-deck aléatoire à l'intérieur des classes, le biais de l'estimateur ajusté (7.3) est donné par

$$\text{Biais}(\overline{y}_{I,C}) \approx \frac{1}{N} \sum_{g=1}^{C} \overline{P}_g^{-1} \sum_{i \in U_g} (p_i - \overline{P}_g)(y_i - \overline{Y}_g), \tag{7.5}$$

où $\overline{P}_g = \dfrac{1}{N_g} \sum_{i \in U_g} p_i$ et $\overline{Y}_g = \dfrac{1}{N_g} \sum_{i \in U_g} y_i$. Le biais en (7.5) est égal à zéro si la covariance entre les variables p et y est zéro dans chacune des classes. En pratique, il est possible de satisfaire à cette exigence en formant des classes d'imputation qui sont homogènes par rapport aux probabilités de réponse p_i et/ou par rapport à la variable d'intérêt y. Notons que les expressions (7.2) et (7.5) sont également valides dans le cas de l'imputation par la moyenne. Finalement, notons que les estimateurs $\overline{y}_{I,1}$ et $\overline{y}_{I,C}$ coïncident lorsque $C = 1$.

7.2. Construction des classes d'imputation

Plusieurs méthodes sont utilisées en pratique pour la formation des classes d'imputation. Nous en mentionnons maintenant quelques unes.

7.2.1. Strates

Dans le cas de plans d'échantillonnage stratifiés, les strates (ou groupes de strates) peuvent être utilisées comme classes d'imputation. La qualité des classes (en terme d'homogénéité) dépend en grande partie de la qualité de l'information auxiliaire utilisée à l'étape du plan de sondage pour former les strates. Cette méthode est fréquemment utilisée dans le cadre des enquêtes auprès des entreprises.

7.2.2. Méthode par croisement

Cette méthode consiste à former les classes en croisant des variables auxiliaires qualitatives. Cette méthode simple se présente sous plusieurs versions. Elle consiste habituellement à former les classes en croisant des variables géographiques (province, ville,...) ou des variables socio-économiques (âge, sexe,..). Une version plus «sophistiquée» de la méthode consiste à d'abord

effectuer un travail de modélisation ; il s'agit, en effet, de préalablement déterminer, parmi les variables auxiliaires disponibles, un ensemble de variables qui sont corrélées avec la ou les variable(s) d'intérêt. Par la suite, les variables sélectionnées sont croisées pour former les classes. Si le modèle est bien spécifié, les classes seront vraisemblablement homogènes par rapport à la variable d'intérêt. Notons cependant que le croisement de variables peut mener à un nombre gigantesque de classes. Par exemple, le croisement de 8 variables, chacune comprenant 5 catégories, mène à la formation de $8^5 = 390\,625$ classes. Par conséquent, un bon nombre de classes pourrait contenir peu ou pas d'unités, ce qui mènerait potentiellement à des estimations instables. En pratique, on spécifie des contraintes pour assurer une certaine stabilité des estimations. On peut ainsi spécifier que le nombre (ou la proportion) de répondants à l'intérieur d'une classe soit supérieur(e) ou égal à un certain seuil. Si les contraintes ne sont pas satisfaites, un regroupement des classes est habituellement effectué (par exemple, en éliminant une des variables auxiliaires et en croisant les variables restantes).

7.2.3. Méthode des scores

Cette méthode permet de former des classes d'imputation homogènes par rapport aux probabilités de réponse et/ou à la variable d'intérêt (Little, 1986 ; Eltinge et Yansaneh, 1997). Les étapes pour la formation des classes peuvent être décrites comme suit :

(i) En utilisant l'information auxiliaire disponible pour toutes les unités dans l'échantillon, construire deux modèles : l'un pour estimer les probabilités de réponse à la variable y et l'autre pour prédire cette variable d'intérêt. L'estimation des probabilités de réponse peut être effectuée, par exemple, au moyen d'une régression logistique. La prédiction de la variable d'intérêt dépend de la nature de celle-ci (continue, discrète,...). Deux scores, $\widehat{p}$ et $\widehat{y}$ sont alors créés pour toutes les unités dans l'échantillon (répondants et non-répondants). Ces scores serviront de critères d'homogénéité pour la formation des classes.

(ii) Choisir un des deux critères, $\widehat{p}$ ou $\widehat{y}$. Ensuite, diviser l'échantillon en classes en utilisant le critère choisi. Pour cela, plusieurs méthodes peuvent être utilisées. On peut utiliser une méthode simple appelée « méthode des quantiles égaux » qui consiste à ordonner les valeurs par rapport au critère choisi puis à diviser l'échantillon en classes de tailles approximativement égales. Une approche alternative est d'utiliser un algorithme de classification pour former les classes.

(iii) Imputer à l'intérieur de chaque classe et calculer l'estimateur imputé intra-classe $\overline{y}_g$ donné en (7.4) pour la classe, $g = 1, ..., C$.

(iv) Calculer l'estimateur (7.3).

7.3. Comparaison des méthodes

Dans cette section, nous présentons certains résultats provenant d'une étude par simulation effectuée par Haziza et Beaumont (2002). Le but de cette étude

est de comparer la performance de la méthode par croisement et de la méthode des scores dans le cas de l'imputation par hot-deck aléatoire.

7.3.1. Création de la population et enjeux de la simulation

Une population de taille $N = 2\,000$ observations a été générée. La population comprend 5 variables : une variable d'intérêt y et 4 variables auxiliaires z_1, z_2, z_3 et z_4. D'abord, les variables z_j, $j = 1, 2, 3, 4$ ont été générées à partir d'une distribution exponentielle de moyenne 30. Ensuite, étant donné z_1 et z_2 la variable y a été générée selon le modèle de régression

$$y_1 = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \varepsilon_i, \tag{7.6}$$

où les erreurs ε_i ont été générées à partir d'une distribution normale de moyenne 0 et de variance σ^2. Les paramètres du modèle β_0, β_1 et β_2 ont respectivement été fixés à 20, 0.5 et 0.5. Finalement, la variance σ^2 a été choisie de telle sorte que le coefficient R^2 du modèle (7.6) soit approximativement égal à 0.8. Notons qu'une comparaison des p-valeurs a montré que la variable z_1 est plus significative que la variable z_2. Par la suite, les variables auxiliaires z_1, z_2, z_3, z_4 ont été catégorisées, ce qui a mené à la création de 4 nouvelles variables z_{1c}, z_{2c}, z_{3c} et z_{4c} chacune de ces variables comprenant 5 catégories. La catégorisation des variables est nécessaire pour la méthode par croisement qui utilise des variables auxiliaires qualitatives seulement (voir section 7.2.2). La moyenne de la population ainsi créée est égale à $\overline{Y} = 49.87$ De cette population, $R = 1\,000$ échantillons de taille $n = N = 2\,000$ (cas d'un recensement) ont été tirés. Dans ce cas, toutes les unités ont le même poids de sondage égal à 1. Dans chacun des échantillons, la non-réponse à la variable y a été générée selon 4 mécanismes de non-réponse. Le taux de réponse a été fixé à 70 %. Les mécanismes de non-réponse sont décrits dans l'Annexe 1. Finalement, dans chaque échantillon (contenant répondants et non-répondants), les classes d'imputation ont été formées selon la méthode par croisement et la méthode des scores. Dans les sections 7.3.2 et 7.3.3, nous décrivons les différents aspects étudiés dans l'étude par simulation.

7.3.2. Méthode par croisement

Les classes sont formées à partir de combinaisons de q variables auxiliaires sélectionnées. La méthode peut être décrite comme suit :

(1) Les q variables auxiliaires sont d'abord classées par ordre d'importance, de la plus significative à la moins significative.

(2) Ces variables sont croisées afin de former les classes.

(3) À l'intérieur de chaque classe, on impose (ou non) deux contraintes :
   a) Le nombre minimal de donneurs par classe est k.
   b) Le nombre de donneurs est supérieur au nombre de receveurs (contrainte PDR).

(4) Si les deux contraintes précédentes sont vérifiées à l'intérieur d'une classe, les valeurs des donneurs de la classe sont imputées aux receveurs de la

classe par hot-deck aléatoire. Les unités imputées sont alors conservées dans un fichier.

(5) Lorsque l'une des contraintes n'est pas satisfaite, la variable la moins significative est éliminée.

(6) Les $q-1$ variables restantes sont alors croisées. Encore une fois, certaines classes respecteront les contraintes, auquel cas l'imputation est effectuée. Certaines classes ne les respecteront pas. Dans ce dernier cas, on ôte la variable la moins significative et on croise les $q-2$ variables restantes. On répète le processus jusqu'à ce que chaque receveur ait trouvé un donneur.

(7) L'estimateur imputé de la moyenne $\overline{Y}$ donné en (7.1) est calculé.

Pour cette méthode, les aspects suivants ont été étudiés :

1. L'impact du mécanisme de non-réponse.
2. Les conséquences d'une mauvaise classification des variables (mauvaise modélisation).

Pour chaque mécanisme, nous avons fait varier différents paramètres :

- Le nombre minimal k de donneurs par classe : $k = 1, 5, 9$.
- La présence de la contrainte «plus de donneurs que de receveurs» (contrainte PDR).
- L'ordre des variables auxiliaires : bon ordre ($z_{1c}z_{2c}z_{3c}z_{4c}$) et mauvais ordre ($z_{4c}z_{3c}z_{2c}z_{1c}$).

Le Tableau 2.1, présenté dans l'Annexe 2, exhibe le biais relatif de l'estimateur imputé (7.1).

7.3.3. Méthode des scores

Les classes d'imputation sont formées suivant les étapes i) - iv) de la section 7.2.3.

Pour cette méthode, nous souhaitons étudier différents aspects :

1. L'impact du mécanisme de non-réponse.
2. L'importance d'avoir une bonne modélisation de la variable d'intérêt.
3. L'impact du choix du critère ($\widehat{p}$ ou $\widehat{y}$) lors de la formation des classes.
4. Comparer la méthode des quantiles égaux vis-à-vis de la classification.

Pour chaque mécanisme, nous avons fait varier différents paramètres :

- Le nombre de classes : 1-10, 15, 20, 30, 40 et 50.
- Le critère servant à former les classes : $\widehat{y}$ et $\widehat{p}$.
- La méthode de formation des classes et le modèle d'imputation : méthode des quantiles égaux et algorithme de classification.
- Le modèle servant à calculer les prédictions de cette variable :
  - $y = \beta_0 + \beta_1 z_1 + \beta_2 z_2$ (bon modèle)
  - $y = \beta_0 + \beta_2 z_2$ (mauvais modèle avec une variable en moins)

Les graphiques 3.1-3.5, présentés dans l'Annexe 3, exhibent le biais relatif de l'estimateur « ajusté » (7.3).

7.3.3. Discussion des résultats

Méthode par croisement : (voir Tableau 2.1, Annexe 2)

(i) Nous constatons que, sous le mécanisme 1 (mécanisme uniforme), l'estimateur imputé est approximativement sans biais dans tous les scénarios.

Les remarques (ii)-(v) qui suivent réfèrent au cas pour lequel les variables z_{1c}, z_{2c}, z_{3c} et z_{4c}) sont classées dans le bon ordre.

(ii) Pour le mécanisme ignorable 3, nous constatons que le biais augmente légèrement à mesure que le nombre minimal de donneurs augmente. Ce résultat s'explique par le fait que, à mesure que les contraintes deviennent plus sévères, il devient de plus en plus difficile de satisfaire lesdites contraintes, au risque d'éliminer des variables importantes du modèle d'imputation. Dans ce cas, la probabilité de réponse dépend de z_2 qui est fortement significative dans le modèle (7.6). Lorsque le nombre minimal de donneurs augmente, la variable z_{2c} est vraisemblablement éliminée de la liste ($z_{1c}z_{2c}z_{3c}z_{4c}$) ce qui explique la présence du biais.

(iii) Pour le mécanisme ignorable 2, nous constatons que le biais reste petit même lorsque le nombre minimal de donneurs augmente. Ce résultat s'explique par le fait que, malgré que la probabilité de réponse dépende de z_1 (qui est fortement significative), la variable z_{1c} n'est jamais éliminée de la liste. Elle est donc présente dans le modèle d'imputation même lorsque les contraintes deviennent plus sévères.

(iv) Pour le mécanisme non-ignorable, l'estimateur est biaisé dans tous les cas comme il fallait s'y attendre. Notons que le biais augmente quand le nombre minimal de donneurs augmente.

(v) Lorsque la contrainte « plus de donneurs que de receveurs » est présente, les estimateurs sont biaisés. Cette contrainte semble, dans ce cas-ci, difficile à satisfaire ce qui mène à l'élimination de plusieurs variables de la liste, d'où la présence de biais.

Les remarques (vi)-(vii) suivantes réfèrent au cas pour lequel les variables z_{1c}, z_{2c}, z_{3c} et z_{4c} sont classées dans le mauvais ordre.

(vi) Mis à part pour le mécanisme uniforme, les estimateurs sont biaisés de manière substantielle. Ceci s'explique facilement par le fait que les variables les plus significatives (z_{1c} et z_{2c}) sont les premières à être éliminées lorsque les contraintes ne sont pas satisfaites. Remarquons également que le biais relatif des estimateurs augmente à mesure que le nombre minimal de donneurs augmente.

(vi) Encore une fois, la contrainte « plus de donneurs que de receveurs » mène à des estimateurs ayant un biais relatif supérieur à celui observé lorsque la contrainte n'est pas appliquée.

Méthode des scores : (voir Annexe 3)

(i) Bien que l'algorithme de classification présente des résultats légèrement supérieurs à ceux obtenus à l'aide de la méthode des quantiles égaux lorsque le nombre de classes est petit (1 à 5), les deux méthodes mènent à des résultats très similaires lorsque le nombre de classes utilisées augmente (voir graphiques 3.1-3.4).

(ii) Les graphiques 3.1-3.4 montrent que le choix du score ($\widehat{p}$ ou $\widehat{y}$) n'est pas un facteur déterminant quant au biais de l'estimateur imputé.

(iii) Nous constatons que sous le mécanisme 1 (mécanisme uniforme), l'estimateur imputé est toujours approximativement sans biais dans tous les scénarios, ce qui n'est pas surprenant en vertu de l'expression du biais (7.2).

(iv) Pour les mécanismes ignorables 2 et 3 et le bon modèle d'imputation (ou le bon modèle de non-réponse), le biais relatif de l'estimateur tend vers 0 lorsque le nombre de classes augmente (voir graphiques 3.1-3.4).

(v) Pour le mécanisme non-ignorable et le bon modèle d'imputation, bien que le biais diminue lorsque le nombre de classes augmente, ce dernier ne devient pas négligeable (voir graphiques 3.1 et 3.2). Ce résultat n'est pas surprenant en vertu de l'expression (7.5).

(vi) Nous remarquons que, dans tous les cas, le biais se stabilise très rapidement (autour de 10 classes). En terme de biais, l'utilisation de classes supplémentaires semble donc superflue.

(vii) Lorsque l'on utilise le mauvais modèle d'imputation (sans z_2), l'estimateur imputé est fortement biaisé pour le mécanisme ignorable 2 (pour lequel la probabilité de réponse dépend de z_2). Cela montre bien que l'omission d'une variable corrélée simultanément avec la probabilité de réponse et la variable d'intérêt, résulte en des estimateurs fortement biaisés (voir graphique 3.5).

Remarques générales

(i) L'utilisation d'un modèle d'imputation qui ne contient pas toute l'information auxiliaire appropriée mènera presque toujours à des estimateurs biaisés.

(ii) La méthode par croisement est sensible aux contraintes et au mauvais classement des variables dans la liste. Dans le cas de la méthode des scores qui permet d'utiliser un nombre de classes relativement petit, la question des contraintes n'est généralement pas un facteur sensible puisque les classes contiennent un grand nombre de donneurs. De plus, le problème de l'ordre des variables ne se pose pas puisque les classes sont formées à partir des scores $\widehat{p}$ ou $\widehat{y}$.

(iii) Dans le cas de la méthode par croisement, l'estimation de la variance sera vraisemblablement ardue car les classes ne sont pas nécessairement disjointes. En effet, un donneur peut être utilisé à différentes étapes du processus dans des classes différentes. Pour simplifier le calcul de la variance, il est d'usage de faire comme si les classes étaient disjointes.

Dans le cas de la méthode des scores, l'estimation de la variance sera relativement aisée puisque, dans ce cas, les classes sont toujours disjointes.

(iv) Pour la méthode des scores, un petit nombre de classes (10-20) suffit en général pour stabiliser le biais des estimateurs.

(v) Pour la méthode des scores, il est à prévoir que l'utilisation du score $\widehat{p}$ mènera à des estimateurs avec une plus grande erreur quadratique moyenne que celle des estimateurs obtenus lorsque le score $\widehat{y}$ est utilisé (Haziza et Beaumont, 2005).

8. Conclusion

L'imputation est une technique fréquemment utilisée pour traiter la non-réponse partielle dans les enquêtes. Nous avons montré que l'imputation est avant tout un exercice de modélisation. Il est donc important de considérer toute l'information auxiliaire appropriée surtout si celle-ci est corrélée avec la probabilité de réponse. De plus, la communauté scientifique reconnaît aujourd'hui l'importance de tenir compte de l'impact de la non-réponse sur la variance (ou coefficient de variation) des estimateurs. Depuis le début des années 90, les chercheurs ont surtout privilégié le domaine de l'estimation de la variance en présence de données imputées, ce qui a mené à de nombreux articles scientifiques à ce sujet. Pour plusieurs autres sujets, il reste cependant plusieurs questions sans réponse. Dans le futur, de nombreux défis attendent donc les chercheurs et méthodologues. Voici quelques sujets pour lesquelles des études théoriques et empiriques sont à faire :

(i) L'estimation (ponctuelle et de variance) pour des paramètres complexes (quantiles, coefficients de régression, coefficients de corrélation).

(ii) La construction des classes d'imputation lorsque l'on utilise les deux scores ($\widehat{p}$ et $\widehat{y}$) simultanément.

(iii) La construction des classes d'imputation par rapport à plusieurs variables d'intérêt simultanément.

(iv) L'imputation et l'estimation (ponctuelle et de variance) pour des plans complexes.

(v) La préservation de la structure multivariée pour des données d'enquête.

Références

[1] BEAUMONT J.-F., HAZIZA D., RANCOURT E. (2005). « Variance estimation under auxiliary value imputation », *Technical report, Statistics Canada, Ottawa.*

[2] BRICK J. M., KALTON G., KIM J. K. (2004). « Variance estimation with hot deck imputation using a model », *Survey Methodology*, **31**, 57-66.

[3] CHEN J., SHAO J. (2000). « Nearest-neigbor imputation for survey data », *Journal of Official Statistics* **16**, 583-599.

[4] CHEN J., SHAO J. (2001). « Jackknife variance estimation for nearest-neigbor imputation », *Journal of the American Statistical Association*, **96**, 260-269.

[5] CHEN J., RAO J. N. K., SITTER R. R. (2000). « Efficient random imputation for missing data in complex surveys », *Statistica Sinica*, **10**, 1153-1169.

[6] DEVILLE J. C., SÄRNDAL C. E. (1994). « Variance estimation for the regression imputed Horvitz-Thompson estimator », *Journal of Official Statistics*, **10**, 381-394.

[7] EFRON B. (1979). « Bootstrap methods : another look at the jackknife », *Annals of Statistics*, **7**, 1-26.

[8] ELTINGE J. L., YANSANEH I. S. (1997). « Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U. S. Consumer Expenditure Survey », *Survey Methodology*, **23**, 33-40.

[9] FAY R. E. (1991). « A design-based perspective on missing data variance », *Proceedings of the 1991 Annual Research Conference, U.S. Bureau of the Census*, 420-440.

[10] HANSEN M. H., HURVITZ W. N., MADOW W. G. (1953). « Sample Survey Methods and theory », I et II, New York Wiley.

[11] HAZIZA D., BEAUMONT J.-F. (2002). « Construction des classes d'imputation dans les enquêtes », *Technical report, Statistics Canada, Ottawa.*

[12] HAZIZA D., BEAUMONT J.-F. (2005). « On the construction of imputation classes in surveys », Soumis pour publication.

[13] HAZIZA D., RAO J. N. K. (2003a), « Inference for totals in cluster sampling under mean imputation for missing survey », *Actes du Symposium de Statistique Canada 2003*, CD-Rom.

[14] HAZIZA D., RAO J. N. K. (2003b). « Inference for population means under unweighted imputation for missing survey data », *Survey Methodology*, **23**, 81-90.

[15] HAZIZA D., RAO J. N. K. (2004). « Inférence pour des statistiques bivariées en presence d'imputation dans le cas d'enquêtes stratifiées à degrés multiples », dans Échantillonnage et Méthodes d'Enquêtes, Ardilly. P., (éditeur), 189-196.

[16] HAZIZA D., RAO J. N. K. (2005). « Inference for domains under imputation for missing survey data », *Canadian Journal of Statistics*, **33**, 149-161.

[17] KALTON G., KASPRZYK D. (1986). « The treatment of missing survey data », *Survey Methodology*, **12**, 1-16.

[18] KOVAR J. G., WHITRIDGE P. J. (1995). « Imputation of business survey data », dans *Business Survey Methods*, Cox, B. G., Binder, D. A., Chinnappa, B. N., Christianson, A., Colledge, M.J. and Kott, P. S. (editors), New York : John Wiley and Sons, 403-423.

[19] LITTLE R. J. A. (1986). « Survey nonresponse adjustments », *International Statistical Review*, **54**, 139-157.

[20] OH H. L., SCHEUREN F. J. (1983). « Weighting adjustments for unit nonresponse », dans *Incomplete Data in Sample Survey*, vol2, Madow, W. G., Olkin, I. , Rubin, D. B. (editors), New York : John Wiley and Sons, 143-184.

[21] QIN J., LEUNG D., SHAO J. (2002). « Estimation with survey data under nonignorable nonresponse or informative sampling », *Journal of the American Statistical Association*, **97**, 193-200.

[22] QUENOUILLE M. (1949). « Problems in plane sampling », *Annals of Mathematical Statistics*, **20**, 355-375.

[23] RANCOURT E. (2001). « Edit and imputation : from suspicious to scientific techniques », *Actes de l'Association internationale des statisticiens d'enquête*, 605-633.

[24] RAO J. N. K. (1990). « Variance estimation under imputation for missing data », *Technical report, Statistics Canada, Ottawa.*

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

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

[27] RUBIN D. B. (1976). « Inference and missing data », *Biometrika*, **63**, 581-590.

[28] 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, 20-34.

[29] SAIGO H., SHAO J., SITTER R. R. (2001). « A repeated half-sample bootstrap and balanced repeated replication for randomly imputed data », *Survey Methodology*, 27, 189-196.

[30] SANTOS R. (1981). « Effect of imputation on regression coefficients », *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 140-145.

[31] SÄRNDAL C. E. (1990). « Methods for estimating the precision of survey estimates when imputation has been used », *Proceedings of Symposium 1990*, Measurement and improvement of data quality, 337-347.

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

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

[34] SHAO J., STEEL P. (1999). « Variance estimation for survey data with composite imputation and nonnegligible sampling fractions », *Journal of the American Statistical Association*, **94**, 254-265.

[35] SHAO J., WANG H. (2002). « Sample correlation coefficients based on survey data under regression imputation », *Journal of the American Statistical Association*, **97**, 544-552.

[36] SKINNER C.J., RAO J. N. K. (2002). « Jackknife variance estimation for multivariate statistics under hot deck imputation from common donors », *Journal of Statistical Planning and Inference*, **102**, 149-167.

Annexe 1

Les 4 mécanismes de non-réponse utilisés sont comme suit :

Mécanisme 1. Mécanisme uniforme, c'est-à-dire $p_i = 0.7$ pout tout $i = 1, \ldots, N$.

Mécanisme 2. Mécanisme ignorable qui dépend de z_1 tel que $\log\left(\frac{p_i}{1-p_i}\right) = \lambda_0 + \lambda_1 z_{1i}$ où λ_0 et λ_1 sont choisis de manière à obtenir un taux global de réponse de 70 %.

Mécanisme 3. Mécanisme ignorable : comme le mécanisme 2 mais remplacer z_1 par z_2.

Mécanisme 4. Mécanisme non-ignorable : comme le mécanisme 2 mais remplacer z_1 par y.

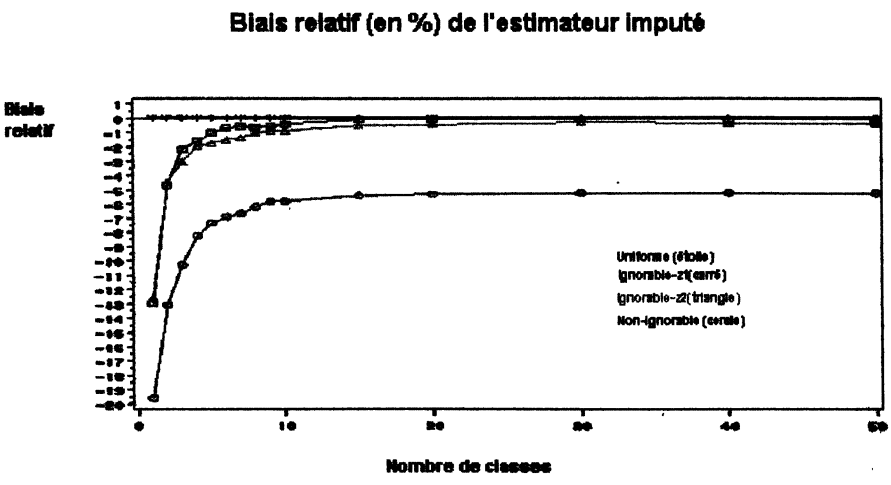
Annexe 2

TABLEAU 2.1. – Biais relatif (en %) de l'estimateur imputé dans le cas de la méthode par croisement

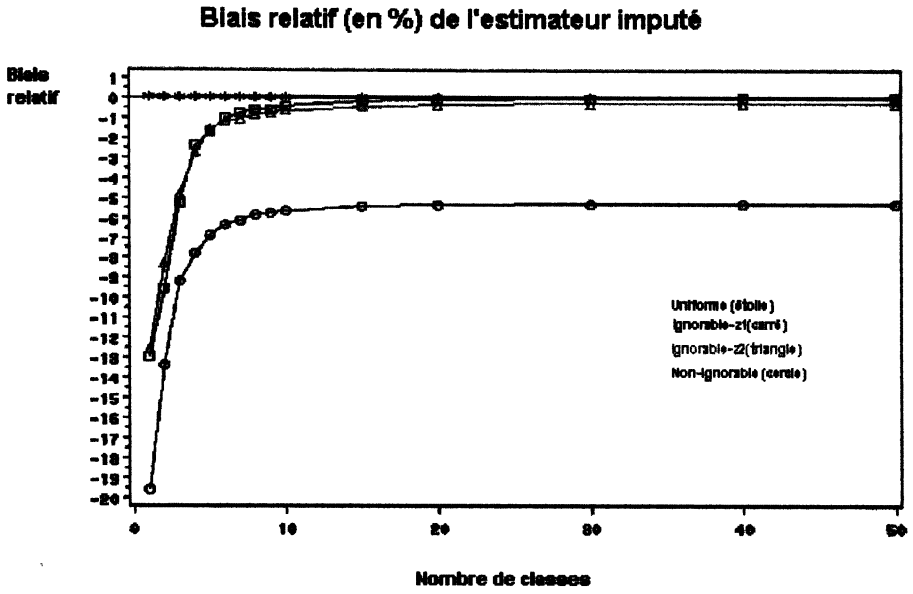
		Nombre minimal de donneurs (sans contrainte PDR)			Nombre minimal de donneurs (avec contrainte PDR)		
Ordre des variables	Mécanisme	1	5	9	1	5	9
$z_{1c}z_{2c}z_{3c}z_{4c}$	1	0.04	0.01	0.00	0.05	0.00	0.02
	2	−0.73	−0.69	−0.59	−0.74	−0.76	−0.76
	3	−0.65	−0.73	−2.11	−11.73	−12.22	−12.37
	4	−8.40	−8.54	−8.73	−12.94	−13.28	−13.40
$z_{4c}z_{3c}z_{2c}z_{1c}$	1	0.01	0.00	−0.01	0.05	0.00	0.01
	2	−7.97	−13.04	−13.27	−13.03	−13.26	−13.21
	3	−1.53	−10.20	−11.57	−11.92	−12.26	−12.35
	4	−11.38	−16.23	−18.04	−18.43	−19.34	−19.63

Annexe 3

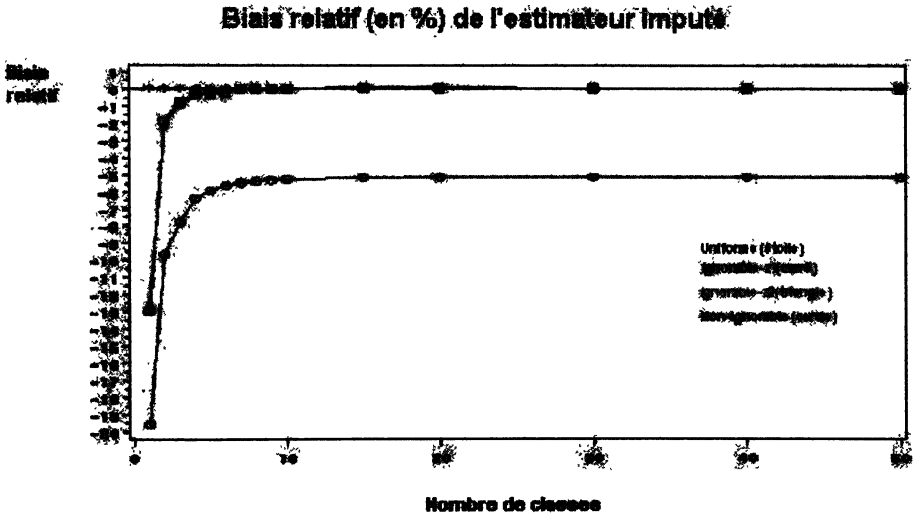
3.1. Graphiques obtenus pour le bon modèle, le critère $\widehat{y}$ et la méthode des quantiles égaux



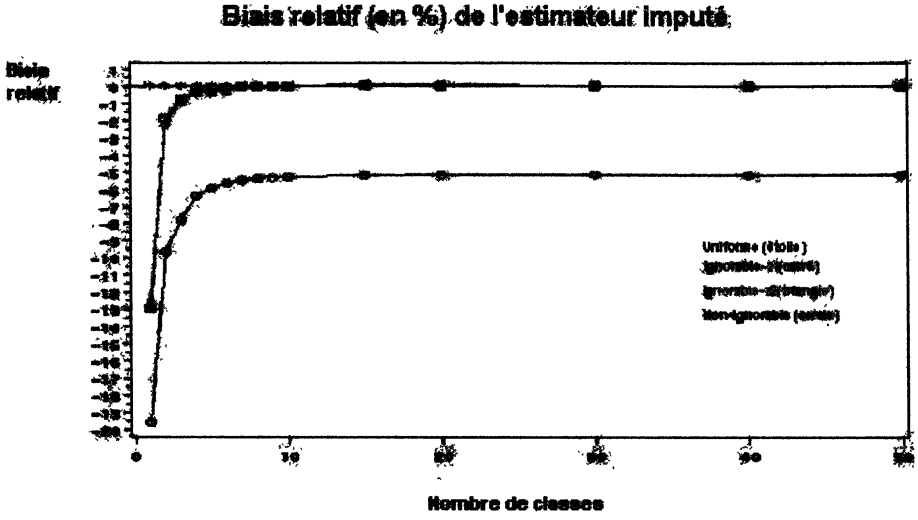
3.2. Graphiques obtenus pour le bon modèle, le critère $\widehat{y}$ et la méthode de classification



3.3. Graphiques obtenus pour le bon modèle, le critère $\widehat{p}$ et la méthode des quantiles égaux



3.4. Graphiques obtenus pour le bon modèle, le critère $\widehat{p}$ et la méthode de classification



3.5. Graphiques obtenus pour le mauvais modèle (z_2 en moins), le critère $\widehat{y}$ et la méthode de classification