INFÉRENCE EN PRÉSENCE D'IMPUTATION : UN SURVOL

David HAZIZA

Statistique Canada, DMEM

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 qui sont disponibles pour toutes les unités échantillonnées ou pour toutes les unités de la population.

Pour traiter la non-réponse, il existe plusieurs méthodes. La plupart du temps, la non-réponse totale est traitée en utilisant une méthode de repondération et la non-réponse partielle est traitée en utilisant l'imputation. La repondération consiste à hausser le poids de sondage des unités répondantes dans l'échantillon pour compenser pour les non-répondants. L'imputation est une technique qui consiste à affecter une (ou plusieurs) "valeur(s) artificielle(s)" pour remplacer une valeur manquante.

Le mot *imputation* vient du mot latin *imputare* qui peut être traduit en anglais par "guesstimate inside" (Rancourt, 2001). 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". En effet, bien qu'il y eut prolifération de l'utilisation de l'imputation dans les enquêtes au début des années 70 avec l'avènement des systèmes automatisés de vérification et d'imputation, ce n'est que récemment que les résultats théoriques ont commencé à apparaître.

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 complété. L'imputation multiple, suggérée par Rubin (1978), consiste à créer $M \geq 2$ valeurs imputées pour boucher le trou laissé par la valeur manquante, ce qui mènera à la création de M fichiers de données complétés. L'idée de base de l'imputation multiple est d'adéquatement combiner les

estimations issues de chacun des fichiers complétés 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 généralement l'imputation simple. En effet, l'utilisation de l'imputation multiple dans le contexte des enquêtes est encore relativement rare. C'est pourquoi, dans ce qui suit, nous discuterons exclusivement de l'imputation simple.

Les statisticiens d'enquête tentent généralement d'éviter l'utilisation de modèles à des fins d'inférence. En effet, 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. En effet, dans ce cas, l'emploi de modèles devient alors inévitable. La validité des estimateurs (ponctuels et de variance) dépendra 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. C'est pourquoi, l'imputation est avant tout un exercice de modélisation. La qualité des estimations reposera donc sur la disponibilité d'information auxiliaire de qualité. Cette information auxiliaire servira à construire des valeurs imputées et/ou à construire des classes d'imputation.

Le plan de l'article est comme suit : Dans la section 1, nous présentons quelques méthodes servant au traitement de la non-réponse partielle dans les enquêtes. Dans la section 2, nous présentons l'estimateur imputé d'une moyenne ainsi que différentes méthodes d'imputation. Nous y discutons également de la notion de mécanisme de non-réponse et des approches proposées 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 alors que les méthodes d'imputation aléatoires ont tendance à la préserver. L'estimation de la variance en présence de valeurs imputées sera traitée à 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. Finalement, nous traitons des classes d'imputation et de leur construction dans la section 7. Nous y présentons également une étude par simulation.

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 :

1.1. 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 que simple et bien 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 soit indépendante de toutes les variables d'intérêt (par exemple, non-réponse uniforme). En rejetant tous les répondants partiels, le statisticien se prive aussi d'information de grande valeur. Finalement, les poids de sondage ne peuvent être utilisés pour faire l'inférence (à moins qu'ils soient ajustés) si bien que celle-ci doit être conditionnelle à l'échantillon de répondants complets. C'est pourquoi cette option ne doit pas être sérieusement considérée à d'autres fins qu'une description sommaire du fichier.

1.2. Méthodes de repondération

Les méthodes de repondération, quant à elles, 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. Par exemple, dans une enquête comprenant plus d'une centaine de variables telle l'Enquête sur la Population Active Canadienne (EPA), il faudrait créer le même nombre de poids ajustés. Ce désavantage important 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 certains avantages pour traiter la non-réponse partielle dont les suivants:

(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.

Dans ce contexte, le choix entre la repondération et l'imputation est relativement facile à faire. 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.

Il est toutefois important de souligner que l'imputation comporte également certains 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 (à propos du mécanisme de non-réponse et/ou du modèle d'imputation) sont satisfaites.
2. Certaines méthodes d'imputation ont tendance à distordre la distribution des variables d'intérêt.
3. L'imputation a comme effet d'atténuer 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 est appréciable.

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

2. Quelques généralités

2.1. Un estimateur imputé

Considérons une population finie *U* de taille *N*. L'objectif est d'estimer des paramètres simples tels un total ou une moyenne de la population donnés respectivement par $Y = \sum_{i \in U} y_i$ et $\bar{Y} = \frac{1}{N} \sum_{i \in U} y_i$.

L'estimation de paramètres plus complexes sera considérée à la section 6. Pour cela, on tire un échantillon aléatoire *s*, de taille *n*, selon un plan de sondage *p*(.). En l'absence de non-réponse, un estimateur de $\bar{Y}$ est donné par

$$\bar{y} = \frac{1}{\sum_{i \in s} w_i} \sum_{i \in s} w_i y_i , \tag{2.1}$$

où $w_i = 1/p_i$ dénote le poids de sondage de l'unité *i* et $p_i = P(i \in s)$ est la probabilité d'inclusion de l'unité *i* dans l'échantillon. L'estimateur (2.1) est approximativement sans biais pour $\bar{Y}$, c'est-à-dire, $E_p(\bar{y}) \approx \bar{Y}$, où $E_p(.)$ dénote l'espérance par rapport au plan de sondage *p*(.).

En présence de non-réponse à la variable *y*, on définit plutôt un estimateur imputé de $\bar{Y}$ donné par

$$\bar{y}_I = \frac{1}{\sum_{i \in s} w_i} \left[\sum_{i \in s_r} w_i y_i + \sum_{i \in s_m} w_i y_i^* \right], \tag{2.2}$$

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 $\bar{y}_I$ est simplement la moyenne pondérée des valeurs observées et des valeurs imputées dans l'échantillon.

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 (par exemple, imputation par la moyenne, par le ratio, par régression et par plus proche voisin). Les méthodes aléatoires sont celles qui ont une composante aléatoire; par conséquent, ces méthodes ne fournissent pas nécessairement la même valeur étant donné l'échantillon si le processus d'imputation est répété (par exemple, imputation par hot-deck aléatoire). 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),

$$\begin{gathered} y_i = f(\mathbf{z_i}) + e_i , \\ E(e_i) = 0, \quad E(e_i e_j) = 0, i \neq j, \quad E(e_i^2) = s_i^2 , \end{gathered} \tag{2.3}$$

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 $f(\mathbf{z_i})$ par $\hat{f}_r(\mathbf{z_i})$ au moyen des unités répondantes, $i \in s_r$, c'est-à-dire, $y_i^* = \hat{f}_r(\mathbf{z_i})$. L'imputation aléatoire

peut être vue comme une imputation déterministe à laquelle on a ajouté un résidu aléatoire e^*, c'est-à-dire, $y_i^* = \hat{f}_r(\mathbf{z_i}) + e_i^*$. Ce résidu peut être tiré, par exemple, d'une distribution normale avec moyenne 0 et variance *u*. En pratique, on préfère plutôt utiliser un résidu aléatoire qui correspond aux résidus observés dans l'ensemble s_r des répondants, c'est-à-dire,

$$e_i^* = \left[y_j - \hat{f}_r(z_j)\right]\frac{\hat{s}_i}{\hat{s}_j}, \quad j \in s_r . \tag{2.4}$$

Notons que le modèle (2.3) peut aussi servir à représenter une méthode d'imputation à l'intérieur de classes. Pour cela, il suffit d'ajouter des variables indicatrices d'appartenance aux classes au modèle (2.3).

2.2.1. Quelques méthodes déterministes

(1) Imputation par la régression : Dans ce cas, $f(\mathbf{z_i}) = \mathbf{z_i'ß}$ et $s_i^2 = \mathbf{?'z_i}$ pour un certain vecteur de constantes **?** . On a alors

$$y_i^* = \mathbf{z_i'\hat{B}_r}, \tag{2.5}$$

où $\mathbf{\hat{B}_r} = \left(\sum_{i\in s_r} \frac{w_i}{s_i^2}\mathbf{z_i z_i'}\right)^{-1}\left(\sum_{i\in s_r} \frac{w_i}{s_i^2}\mathbf{z_i} y_\mathbf{i}\right)$ est l'estimateur obtenu par la méthode des moindres carrés pondérée. Un cas particulier de (2.5) est l'imputation par la régression linéaire simple. Dans ce cas, une seule variable auxiliaire *z* est disponible, $f(\mathbf{z_i}) = ß_0 + ß_1 z_i$ et $s_i^2 = s^2$. On a alors

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

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

(2) Imputation par le ratio : Dans ce cas, une seule variable auxiliaire *z* est disponible, $f(\mathbf{z_i}) = ß z_i$ et $s_i^2 = s^2 z_i$. On a alors

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

où $\hat{B}_r = \frac{\bar{y}_r}{\bar{z}_r}$.

(3) Imputation par la valeur précédente (ou historique) : $f(\mathbf{z_i}) = y_{i,t-1} \equiv z_i$ et $s_i^2 = s^2$. Ici, $y_{i,t-1}$ représente la valeur de la variable *y* pour l'unité *i* au temps $t-1$. On a alors

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

(4) Imputation par la moyenne : $z_i = 1 \forall i \in U$, $f(\mathbf{z_i}) = ß$ et $s_i^2 = s^2$. On a alors

$$y_i^* = \hat{B}_r = \bar{y}_r . \tag{2.9}$$

(5) Imputation par le plus proche voisin (PPV): 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 s_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, Euclidienne).

2.2.2. Quelques méthodes aléatoires

(1) Imputation par hot-deck aléatoire non-pondérée : Cette méthode consiste à tirer un répondant au hasard (en général, avec remise) dans l'ensemble s_r des répondants, c'est-à-dire,

$$y_i^* = y_j,\ \ j \in s_r \text{, tel que } P(y_i^* = y_j) = 1/r. \tag{2.11}$$

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

(2) Imputation par le ratio avec résidus: La valeur imputée utilisée pour remplacer la valeur manquante y_i est donnée par

$$y_i^* = \hat{B}_r z_i + e_i^*, \tag{2.12}$$

où $e_i^* = \left[y_j - \hat{B}_r z_j\right]\sqrt{\frac{z_i}{z_j}}$, $j \in s_r$, est un cas particulier de (2.4) dans le cas de l'imputation par le ratio.

2.3. Non-réponse vs échantillonnage à deux phases

La situation prévalant en présence de non-réponse est souvent comparée à la situation prévalant dans le cas d'échantillonnage à deux phases. L'échantillonnage à deux phases est fréquemment utilisé dans les enquêtes lorsque la base de sondage contient peu ou pas d'information. Dans ce cas, il est coutume de préalablement tiré un échantillon s_1 de première phase de taille généralement grande 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 à la première phase, on tire un échantillon s_2 de s_1 selon un plan de sondage $p_2(.|s_1)$. Dans le cas d'échantillonnage à deux phases, un estimateur (qui n'utilise pas d'information auxiliaire) de $\bar{Y}$ est donné par

$$\bar{y}_{DP} = \frac{1}{\sum_{i \in s_2} w_{1i} w_{2i}} \sum_{i \in s_2} w_{1i} w_{2i} y_i, \tag{2.13}$$

où $w_{1i} = 1/p_{1i}$ et $p_{1i} = P(i \in s_1)$ est la probabilité d'inclusion de l'unité i dans l'échantillon de première phase s_1, $w_{2i} = 1/p_{2i}$ et $p_{2i} = P(i \in s_2 \mid s_1, i \in s_1)$ est la probabilité d'inclusion

conditionnelle de l'unité i dans l'échantillon de deuxième phase s_2. L'estimateur $\bar{y}_{DP}$ en (2.13) est approximativement sans biais pour $\bar{Y}$, c'est-à-dire, $E(\bar{y}_{DP}) \equiv E_{p_1} E_{p_2}(\bar{y}_{DP} \mid s_1) \approx \bar{Y}$, où $E_{p_1}(.)$ et $E_{p_2}(.)$ dénotent respectivement l'espérance par rapport au plan de sondage $p_1(.)$ et $p_2(.|s_1)$. Notons que dans le cas d'échantillonnage à deux phases, le statisticien contrôle le mécanisme de sélection des deux échantillons. En d'autres mots, les probabilités d'inclusion p_{1i} et p_{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, les probabilités d'inclusion dans s_r (c'est-à-dire les probabilités de réponse) ne sont pas connues. Puisque les probabilités de réponse ne sont pas connues, nous n'avons d'autres choix que d'établir certaines hypothèses à propos du mécanisme de non-réponse.

2.4. Mécanisme de non-réponse

Comme mentionné à la section 2.3, les probabilités de réponse ne sont pas connues, ce qui nous amène à établir certaines hypothèses à propos du mécanisme de non-réponse. Soit a_i la variable indicatrice de réponse définie par

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

Soit $p_i = P(a_i = 1 \mid s, i \in s)$ la probabilité de réponse pour l'unité i. Nous supposerons que les unités répondent indépendamment les unes des autres, c'est-à-dire,
$p_{ij} = P(a_i = 1, a_j = 1 \mid s, i \in s, j \in s, i \neq j) = p_i p_j$. 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. Dans ce cas, on peut faire appel à un type de mécanisme plus complexe du type beta-binomial (Haziza et Rao, 2001a).
On distinguera trois types de mécanismes de non-réponse :

(1) **Mécanisme uniforme**

Un mécanisme est dit uniforme si la probabilité de réponse est la même pour toutes les unités dans la population, c'est-à-dire, $p_i = p \ \ \forall\, i \in U$. Donc, 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,

$$P(s_r \mid s, r) = \frac{1}{\binom{N}{r}}.$$

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 est dit ignorable si $P(a_i = 1 \mid y, \mathbf{z}) = P(a_i = 1 \mid \mathbf{z})$, (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). Lorsque le mécanisme est ignorable, on dit que les données sont *Missing At Random* (MAR). Dans la littérature, on emploie aussi l'expression *mécanisme non-confondu*.

(3) **Mécanisme non-ignorable**

Lorsque la probabilité de réponse dépend de la variable d'intérêt, on dit que le mécanisme 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 dû à la non-réponse. L'élimination de ce biais va généralement requérir des techniques sophistiquées (par exemple, Lin, Quin et Shao, 2002).

2.5. Approches pour l'inférence

Pour étudier les propriétés (biais et variance) de l'estimateur imputé (2.2), deux approches ont été proposées dans la littérature : l'approche basée sur le plan de sondage (BP) proposée par Rao (1990) et l'approche basée sur un modèle (BM) proposée 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 à la section 7.

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

Approche 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

$$\begin{aligned} m: y_i &= \mathbf{z}_i' \beta + e_i, \\ E_m(e_i) = 0, \quad E_m(e_i e_j) &= 0,\ i \neq j, \quad E_m(e_i^2) = s_i^2 = s^2 \mathbf{z}_i, \end{aligned} \tag{2.14}$$

où $E_m(.)$ dénote l'espérance par rapport au modèle d'imputation. L'approche BP consiste à décrire complètement le mécanisme de non-réponse alors que l'approche BM suppose un mécanisme plus général que nous n'essayerons pas de décrire, d'où l'emploi d'un modèle d'imputation qui décrit la relation qui lie la variable d'intérêt aux variables auxiliaires. Il s'ensuit que, sous l'approche BP, la validité des estimateurs dépendra de la validité du mécanisme de non-réponse alors que sous l'approche BM, la validité des estimateurs dépendra de la validité du modèle d'imputation.

3. Propriétés de l'estimateur imputé

Nous avons vu dans la section 2.1 qu'en l'absence de non-réponse, l'estimateur $\bar{y}$ en (2.1) est approximativement sans biais pour la moyenne de la population $\bar{Y}$. Qu'en est-il de l'estimateur imputé $\bar{y}_I$ 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 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, lestimateur imputé $\bar{y}_I$ sera vraisemblablement approximativement sans biais pour $\bar{Y}$. Considérons le cas de l'imputation par régression. Dans ce cas, les valeurs imputées sont données par (2.5), ce qui mène à

$$\bar{y}_I = \frac{1}{\sum_{i\in s} w_i}\left[\hat{Y}_r + \left(\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_\mathbf{r}\right)' \hat{\mathbf{B}}_\mathbf{r}\right], \tag{3.1}$$

où $\hat{Y}_r = \sum_{i\in s_r} w_i y_i$, $\hat{\mathbf{Z}}_\mathbf{r} = \sum_{i\in s_r} w_i \mathbf{z_i}$ et $\hat{\mathbf{Z}} = \sum_{i\in s} w_i \mathbf{z_i}$. Notons que l'expression entre crochet dans (3.1) est similaire à un estimateur par la régression généralisée pour un total dans le cas d'échantillonnage à deux phases. On peut facilement montrer que, sous l'approche BP, l'estimateur imputé $\bar{y}_I$ (3.1) est approximativement sans biais pour $\bar{Y}$, c'est-à-dire, $E\left(\bar{y}_I - \bar{Y}\right) \equiv E_p E_r\left(\bar{y}_I - \bar{Y} \mid s\right) \approx 0$ où $E_r(.)$ dénote l'espérance par rapport au mécanisme de non-réponse. De façon similaire, on peut facilement montrer que, sous l'approche BM et le modèle (2.14), l'estimateur imputé $\bar{y}_I$ en (3.1) est sans biais pour $\bar{Y}$, c'est-à-dire, $E\left(\bar{y}_I - \bar{Y}\right) \equiv E_r E_p E_m\left(\bar{y}_I - \bar{Y} \mid s\right) \approx 0$.

3.2. 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 $\bar{y}_I$ sera vraisemblablement biaisé. Nous illustrons maintenant ce point en donnant deux exemples : l'un sous l'approche BP et l'autre sous l'approche BM.

3.2.1. Approche BP et imputation par la moyenne

Sous cette approche, 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 $\bar{y}_I = \bar{y}_r$ qui 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 $P(i \in s_r) = p_i$). On peut montrer que, dans ce cas, l'estimateur imputé $\bar{y}_I$ est biaisé et que le biais est donné par

$$Biais(\bar{y}_I) = E_p E_r(\bar{y}_I \mid s) - \bar{Y} \approx \frac{1}{N\bar{P}} \sum_{i\in U} \left(p_i - \bar{P}\right)\left(y_i - \bar{Y}\right), \tag{3.2}$$

où $\bar{P} = \frac{1}{N}\sum_{i\in U} p_i$ est la moyenne des probabilités dans la population. Notons que le biais (3.2) 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 dans le cas où la probabilité de réponse dépend de la variable d'intérêt (mécanisme non-ignorable), le biais en (3.2) ne peut être nul. En fait, l'expression (3.2) justifie la formation de classes d'imputation (voir section 7.1).

3.2.2. Approche 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

$$\bar{y}_I = \frac{\bar{y}_r}{\bar{z}_r}\bar{z}, \tag{3.3}$$

où $\bar{z} = \frac{1}{\sum_{i\in s} w_i}\sum_{i\in s} w_i z_i$. L'imputation par le ratio suggère naturellement l'emploi d'un modèle de la forme

$$y_i = ßz_i + e_i. \tag{3.4}$$

Comme nous l'avons vu à la section 3.1, l'estimateur imputé (3.3) est sans biais pourvu que le modèle (3.4) soit valide pour les unités dans la population. Qu'en est-il si le modèle (3.4) 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 (3.4) mais plutôt

$$y_i = ß_0 + ß_1 z_i + e_i. \tag{3.5}$$

De plus supposons que les probabilités de réponse sont telles que $P(i \in s_r) = p_i$. On peut alors montrer que, sous le modèle (3.5), l'estimateur imputé (3.3) est biaisé et que le biais est égal à

$$Biais(\bar{y}_I) = E_r E_p E_m(\bar{y}_I - \bar{Y}) \approx ß_0\left[\frac{\bar{Z}}{\bar{Z}_p} - 1\right], \tag{3.6}$$

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

Notons que le biais (3.6) est égal à 0 si

(a) $ß_0 = 0$
ou

(b) $\bar{Z} = \bar{Z}_p \Leftrightarrow \frac{1}{N\bar{P}}\sum_{i\in U}(p_i - \bar{P})(z_i - \bar{Z}) = 0.$

La condition (a) est équivalente à utiliser le modèle ratio (3.4). 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 (3.6) est différent de zéro.

3.2.3. Exemples numériques

Pour illustrer les résultats obtenus à la section 3.2.2, nous avons effectué deux études par simulation.

Étude 1 : Nous avons d'abord généré une population de taille $N = 1000$ comprenant deux variables *y* et z_1. De cette population nous avons tiré $R = 10000$ échantillons aléatoires simples sans remise. Dans chaque échantillon tiré, nous avons généré de la non-réponse selon la fonction logistique

$$p_i = \frac{\exp(g_0 + g_1 z_{1i})}{1 + \exp(g_0 + g_1 z_{1i})}. \tag{3.7}$$

Les paramètres g_0 et g_1 ont été choisis de manière à obtenir un taux de réponse 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 régression linéaire simple. Les mesures Monte Carlo suivantes ont été calculées :

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

$$BR(\bar{y}_I) = \frac{1}{R}\sum_{i=1}^{R}\frac{\left(\bar{y}_I^{(i)} - \bar{Y}\right)}{\bar{Y}} \times 100, \qquad (3.9)$$

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

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

$$EQM(\bar{y}_I) = \frac{1}{R}\sum_{i=1}^{R}\left(\bar{y}_I^{(i)} - \bar{Y}\right)^2.$$

La figure 1 indique que la relation entre les variables y et z_1 dans la population est bien linéaire et qu'elle passe par l'origine, 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 bais n'est pas négligeable (environ 4%). Ce résultat s'explique facilement par le fait que la probabilité de réponse et la variables 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. En d'autres mots, 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 dans le cas de l'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.

Figure 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	.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	0.31	0.32

Étude 2 : Dans les mêmes conditions utilisées dans 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 qu'elle ne passe pas par l'origine, 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 dans le cas de l'imputation par le ratio est encore plus grand en valeur absolue (–14%). Ce biais substantiel s'explique par le fait que l'ordonnée à l'origine est fortement significative. Or l'imputation par le ratio suppose que la droite de régression passe par l'origine. Cette dernière n'est donc pas prise en compte pour construire les valeurs imputées. L'inclusion de l'ordonnée à l'origine dans le modèle d'imputation suffit pour réduire le biais à un niveau négligeable (environ 0.1%).

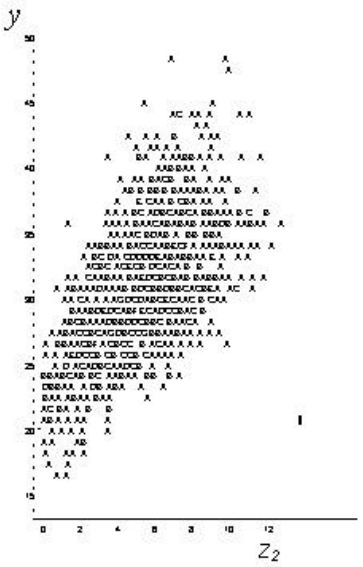


Figure 2

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	<.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	19.22	0.33

En conclusion, les deux 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 tels :

- 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 (c'est-à-dire, les variables que l'on impute) alors que les méthodes d'imputation aléatoires tendent à les 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, $S_y^2 = \frac{1}{N-1}\sum_{i\in U}\left(y_i - \bar{Y}\right)^2$ Pour cela, tirons un échantillon aléatoire simple sans remise, s, de taille n. En l'absence de non-réponse, la variance dans l'échantillon, $s_y^2 = \frac{1}{n-1}\sum_{i\in s}\left(y_i - \bar{y}\right)^2$, est un estimateur sans biais de S_y^2; c'est-à-dire, $E_p\left(s_y^2\right) = S_y^2$.

En présence de non-réponse à la variable y, on définit un estimateur imputé pour S_y^2 par

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

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

(1) Approche BP et imputation par régression linéaire simple

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

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

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

Par la Proposition 1, on a

$$E_r\left(s_{yI}^2 \mid s, r\right) \approx \frac{r-1}{n-1}S_y^2 + \mathrm{B}_1^2\frac{m-1}{n-1}S_z^2, \tag{4.2}$$

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

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

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

où $?_{yz} = \frac{1}{N-1}\frac{\sum_{i\in U}(z_i - \bar{Z})(y_i - \bar{Y})}{S_y S_z}$ 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 $\left|r_{xy}\right| = 1$.

L'expression (4.3) montre que l'imputation par régression ne préserve pas la variance S_y^2 de la population. L'imputation par régression a donc tendance à sous-estimer la variabilité "naturelle" que l'on aurait observé s'il n'y avait pas eu de non-réponse. On dit alors que l'imputation par régression distord la distribution de la variable d'intérêt. La distorsion relative dépend de la corrélation entre les variable *y* et *z*. Une forte corrélation assurera que la variance après imputation des *y* ne sera pas trop affectée. Dans le cas de l'imputation par la moyenne, l'expression (4.3) devient

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

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, $E_p(r/n)$.

(2) Approche 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 que

$$E_r E_* \left(s_{yI}^2 \mid s, r\right) \approx \frac{r-1}{r}\left[1+\frac{r}{n(n-1)}\right] S_y^2 \approx S_y^2, \qquad (4.5)$$

pour r grand. Ici, $E_*(.)$ dénote l'espérance par rapport au mécanisme d'imputation aléatoire.

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 préservent la moyenne $\bar{Y}$ (section 3.1) mais pas la variance S_y^2, alors que les méthodes aléatoires préservent les deux. Notons que préserver la moyenne et la variance n'entraîne pas systématiquement que la fonction de répartition $F(.)$ de la variable 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. Notons que toutes les méthodes présentées dans cette section supposent que l'estimateur imputé est approximativement sans biais. Voici quelques raisons pour lesquelles il est important de calculer correctement la variance en présence de valeurs imputées :

- Cela permet de mesurer la qualité (précision) des estimateurs.
- Cela aide à tirer les bonnes conclusions, plus particulièrement avec les tests d'hypothèse et 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.

Traditionnellement, les chercheurs ont utilisé l'approche deux phases pour estimer la variance. Sous cette approche, on suppose le processus suivant :

Population $U \rightarrow$ Échantillon s

$\rightarrow$ Échantillon avec répondants et non-répondants $\left(s_r, s_m\right)$

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

$$V\left(\bar{y}_I - \bar{Y}\right) = V_p E_r\left(\bar{y}_I - \bar{Y} \mid s\right) + E_p V_r\left(\bar{y}_I - \bar{Y} \mid s\right),$$

où $V_p(.)$ et $V_r(.)$ dénotent 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 tenir compte du mécanisme d'imputation (qui sert à tirer les valeurs imputées) dans le calcul de variance. Dans ce cas, la variance de l'estimateur imputé (2.2) est donnée par

$$V(\bar{y}_I) = V_p E_r E_* (\bar{y}_I - \bar{Y} \mid s) + E_p V_r E_* (\bar{y}_I - \bar{Y} \mid s) + E_p E_r V_* (\bar{y}_I - \bar{Y} \mid s),$$

où $V_*(.)$ dénote la variance par rapport au mécanisme d'imputation. Sous l'approche deux phases, Rao (1990) a proposé d'estimer la variance sous l'approche BP (section 5.1) alors que Särndal (1990) a proposé d'estimer la variance sous l'approche BM (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 $U \rightarrow$ Population avec répondants et non-répondants (U_r, U_m)

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

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

$$V(\bar{y}_I - \bar{Y}) = E_r V_p (\bar{y}_I - \bar{Y} \mid a_i) + V_r E_p (\bar{y}_I - \bar{Y} \mid a_i),$$

(Shao et Steel, 1999). Dans le cas de méthode d'imputation aléatoire, la variance de l'estimateur imputé (2.2) est donnée par

$$V(\bar{y}_I - \bar{Y}) = E_r V_p E_* (\bar{y}_I - \bar{Y} \mid a_i) + E_r E_p V_* (\bar{y}_I - \bar{Y} \mid a_i) + V_r E_p E_* (\bar{y}_I - \bar{Y} \mid a_i).$$

Notons que dans le cas de l'approche deux phases, les espérances et variances internes sont conditionnelles sur l'échantillon s alors que dans le cas de l'approche renversée, les espérances et variances internes sont conditionnelles sur les indicateurs de réponse a_i.

Remarques sur l'approche renversée :

- Dans le cas de l'imputation déterministe, l'estimation de la variance totale de l'estimateur revient à estimer séparément les deux composantes $V_1 = E_r V_p (\bar{y}_I - \bar{Y} \mid a_i)$ et $V_2 = V_r E_p (\bar{y}_I - \bar{Y} \mid a_i)$.
- L'approche renversée permet d'estimer la variance des estimateurs sous les approches BP et BM.
- L'estimation de $V_1 = E_r V_p (\bar{y}_I - \bar{Y} \mid a_i)$ revient à estimer $V_p (\bar{y}_I - \bar{Y} \mid a_i)$. L'estimateur v_1 de V_1 ne dépend pas du mécanisme de non-réponse et/ou du modèle d'imputation. L'estimateur v_1 est donc robuste à une mauvaise spécification du modèle. La composante V_2 quant à elle dépend du mécanisme de non-réponse ou du modèle d'imputation.
- 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) et le bootstrap de Shao-Sitter (Shao et Sitter, 1996) 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.

- Le ratio $\frac{V_2}{V_1}$ est d'ordre $O(n/N)$. Donc, quand la fraction de sondage n/N est négligeable, la composante V_2 est négligeable par rapport à la composante V_1. Dans ce cas, on peut omettre le calcul de V_2.

5.1. Deux phases : approche 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'imputation par la moyenne, $\bar{y}_I = \bar{y}_r$. Par la Proposition 1, on détermine aisément la variance de $\bar{y}_I$ donnée par

$$\begin{aligned} V(\bar{y}_I) &= V_p E_r(\bar{y}_r \mid s,r) + E_p V_r(\bar{y}_r \mid s,r) \\ &\approx \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{E_p(r)} - \frac{1}{n}\right) S_y^2 \\ &= \left(\frac{1}{E_p(r)} - \frac{1}{N}\right) S_y^2. \end{aligned} \tag{5.1}$$

Un estimateur correct de la variance est obtenu en estimant les quantités inconnues dans (5.1), ce qui mène à

$$v_{cor}(\bar{y}_I) = \left(\frac{1}{r} - \frac{1}{N}\right) s_{yr}^2, \tag{5.2}$$

où $s_{yr}^2 = \frac{1}{r-1}\sum_{i \in s_r} (y_i - \bar{y}_r)^2$. Traiter les valeurs imputées comme si elles avaient été observées revient à n'estimer que la première composante $V_p E_r(\bar{y}_r \mid s,r)$. Dans ce cas, on obtiendrait un estimateur incorrect de la variance donné par

$$v_{inc}(\bar{y}_I) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{r-1}{n-1} s_{yr}^2. \tag{5.3}$$

On a alors

$$\frac{v_{cor}(\bar{y}_I)}{v_{inc}(\bar{y}_I)} = \frac{(1-r/N)}{(1-n/N)}\left(\frac{n}{r}\right)\left(\frac{n-1}{r-1}\right) \approx \left(\frac{n}{r}\right)^2, \tag{5.4}$$

si n/N est négligeable, $n \approx n-1$ et $r \approx r-1$. Par exemple, si le taux de non-réponse est 50%, le ratio (5.4) 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 $\bar{y}_I$ (2.2) est donnée par

$$\begin{aligned} V(\bar{y}_I) &= V_p E_r E_* (\bar{y}_r \mid s,r) + E_p V_r E_* (\bar{y}_r \mid s,r) + E_p E_r V_* (\bar{y}_r \mid s,r) \\ &\approx \left[\frac{1}{E_p(r)} - \frac{1}{N} + \frac{1 - E_p(r/n)}{n} \right] S_y^2. \end{aligned} \tag{5.5}$$

En supposant que la fraction de sondage *n/N* est négligeable, une comparaison de (5.1) et (5.5) montre que

$$\frac{V_{hot-deck}(\bar{y}_I)}{V_{moyenne}(\bar{y}_I)} = 1 + p(1-p) \geq 1, \tag{5.6}$$

où $p = E_p\left(\frac{r}{n}\right)$ est le taux de réponse espéré. Notons que le ratio (5.6) est maximum lorsque $p = 1/2$, auquel cas il vaut 1.25.

5.2. Deux phases : approche BM

Cette méthode d'estimation de la variance, développée sous l'approche BM, est due à Särndal (1990, 1992). La méthode se sert de la décomposition suivante comme point de départ:

$$\underbrace{\bar{y}_I - \bar{Y}}_{\text{erreur totale}} = \underbrace{(\bar{y} - \bar{Y})}_{\substack{\text{erreur due à} \\ \text{l'échantillonnage}}} + \underbrace{(\bar{y}_I - \bar{y})}_{\substack{\text{erreur due à} \\ \text{la non-réponse}}} \tag{5.7}$$

Si le plan de sondage et le mécanisme de non-réponse sont ignorables, la variance de l'estimateur imputé $\bar{y}_I$ est alors donnée par

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

Le but sera donc d'estimer chacune des composantes de (5.8) séparément. Notons que la composante V_{mix} est égale à 0 lorsque, par exemple, le plan de sondage est un plan à probabilités égales. Même lorsque cette composante n'est pas égale à 0, elle est souvent petite par rapport à $V_{éch}$ et V_{imp}. Il est donc raisonnable d'omettre V_{mix} lors de l'estimation de la variance.

Estimation de $V_{éch}$: Cette composante représente la variance due à l'échantillonnage lorsqu'il n'y pas de non-réponse. Il suffira alors de trouver un estimateur $\hat{V}_{éch}$ sans biais pour $V_{éch}$. On acceptera de décomposer $\hat{V}_{éch}$ comme suit:

$$\hat{V}_{éch} = \hat{V}_{ord} + \hat{V}_{dif},$$

où $\hat{V}_{ord}$ est l'estimateur de la variance lorsque l'on traite les valeurs imputées comme si elles avaient été observées. Pour certaines méthodes déterministes, la composante $\hat{V}_{ord}$ sous-estime $V_{éch}$. Cette sous-

estimation est compensée en ajoutant la composante $\hat{V}_{dif}$ qui est obtenue en estimant les paramètres du modèle d'imputation dans

$$V_{dif} = E_m\left(\hat{V}_{éch} - \hat{V}_{ord} \mid s, s_r\right).$$

La composante $\hat{V}_{éch}$ est alors approximativement sans biais pour $V_{éch}$. Notons que dans le cas de méthodes d'imputation aléatoires, $V_{dif} \approx 0$, et alors $\hat{V}_{ord}$ est approximativement sans biais pour $V_{éch}$.

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

Un estimateur de la variance totale est alors donnée par

$$\hat{V}_{tot} = \hat{V}_{éch} + \hat{V}_{imp}.$$

Exemple: 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 $\bar{y}_I = \bar{y}_r$. Notons que l'imputation par la moyenne suggère le modèle d'imputation

$$\begin{aligned} m: y_i &= ß + e_i, \\ E_m(e_i) = 0, \quad E_m(e_i e_j) &= 0,\ i \neq j, \quad E_m\left(e_i^2\right) = s^2. \end{aligned} \tag{5.9}$$

Notons que, sous ce modèle, la variance des répondants s_{yr}^2 est un estimateur sans biais pour s^2, c'est-à-dire, $E_m\left(s_{yr}^2\right) = s^2$. On peut facilement montrer que

$$\hat{V}_{ord} = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{r-1}{n-1} s_{yr}^2,$$

$$\hat{V}_{dif} = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{n-r}{n-1} s_{yr}^2,$$

et

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

Un estimateur de la variance totale est donné par

$$\hat{V}_{tot} = \hat{V}_{ord} + \hat{V}_{dif} + \hat{V}_{imp} = \left(\frac{1}{r} - \frac{1}{N}\right) s_{yr}^2 . \tag{5.10}$$

Notons que cet estimateur coïncide avec l'estimateur correct de la variance (5.2) obtenu sous l'approche BP (section 5.1). Deville et Särndal (1994) ont généralisé la méthode au cas de plans de sondage arbitraires et de l'imputation par régression.

Caractéristiques de la méthode :

- La méthode peut s'appliquer à plusieurs méthodes d'imputation (moyenne, ratio, régression, hot-deck aléatoire, etc).
- Bien que ce ne soit pas mentionné dans la littérature, la méthode peut être généralisée au cas de données multivariées.
- La méthode peut être généralisée au cas d'estimateurs non-linéaires (par exemple, un ratio).
- La méthode n'est pas intensive du point de vue informatique.

5.3. 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 $\bar{y}_I = \bar{y}_r$.

Estimation de $V_1 = E_r V_p\left(\bar{y}_I - \bar{Y} \mid a_i\right)$:

Il suffit d'estimer $V_p\left(\bar{y}_I - \bar{Y} \mid a_i\right)$. Pour cela, écrivons d'abord l'estimateur imputé en fonction des indicateurs de réponse a_i. On a alors

$$\bar{y}_I = \sum_{i \in s} a_i y_i \Big/ \sum_{i \in s} a_i .$$

On est donc ramené à estimer la variance due à l'échantillonnage d'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 v_1 un estimateur approximativement sans biais de V_1. On peut, par exemple, utiliser la linéarisation en série de Taylor, ce qui mène à

$$v_1 \approx \left(1 - \frac{n}{N}\right) \frac{s_{yr}^2}{r}. \tag{5.11}$$

Estimation de $V_2 = V_r E_p\left(\bar{y}_I - \bar{Y} \mid a_i\right) = V_r E_p\left(\bar{y}_I \mid a_i\right)$ **sous l'approche BP** :

D'abord, notons que $E_p\left(\bar{y}_I \mid a_i\right) = \sum_{i \in U} a_i y_i \Big/ \sum_{i \in U} a_i$. De plus, notons que sous l'approche BP, on a $V_r\left(a_i\right) = p(1-p)$. On peut alors déterminer $V_r E_p\left(\bar{y}_I - \bar{Y} \mid a_i\right)$ par linéarisation en série de Taylor, ce qui donne

$$V_r E_p\left(\bar{y}_I - \bar{Y} \mid a_i\right) \approx p(1-p) \frac{1}{E_r\left(\sum_{i \in U} a_i\right)} \sum_{i \in U} \left(y_i - \bar{Y}\right)^2. \tag{5.12}$$

Un estimateur de $V_r E_p\left(\bar{y}_I - \bar{Y} \mid a_i\right)$ est obtenu en estimant correctement les quantités inconnues dans (5.12), ce qui donne

$$v_{2\,BP} = \left(\frac{n}{N} - \frac{r}{N}\right) \frac{s_{yr}^2}{r}. \tag{5.13}$$

Estimation de $V_2 = V_r E_p\left(\bar{y}_I - \bar{Y} \mid a_i\right)$ sous l'approche BM :

D'abord, notons que nous faisons appel au modèle d'imputation (5.9). On a alors

$$V_2 = E_r V_m E_p\left(\bar{y}_I - \bar{Y} \mid a_i\right) + \underbrace{V_r E_m E_p\left(\bar{y}_I - \bar{Y} \mid a_i\right)}_{=0}$$

$$= E_r V_m\left(\sum_{i\in U} c_i y_i \mid a_i\right),$$

où $c_i = \dfrac{N a_i - \sum_{i\in U} a_i}{N \sum_{i\in U} a_i}$. On a alors,

$$V_2 = E_r\left[\sum_{i\in U} c_i^2 V_m\left(y_i\right)\right] = s^{\,2} E_r\left[\sum_{i\in U} c_i^2\right]. \tag{5.14}$$

Un estimateur de $V_2 = V_r E_p\left(\bar{y}_I - \bar{Y} \mid a_i\right)$ est obtenu en estimant correctement les quantités inconnues dans (5.14), ce qui donne

$$v_{2\,BM} = \left(\frac{n}{N} - \frac{r}{N}\right)\frac{s_{yr}^2}{r}. \tag{5.15}$$

Notons que dans le cas de l'imputation par la moyenne, v_{2BP} en (5.13) coïncide avec v_{2BM} en (5.15). Un estimateur de la variance totale, v_{tot}, est donné par

$$v_{tot} = v_1 + v_{2BP} = \left(\frac{1}{r} - \frac{1}{N}\right) s_{yr}^2,$$

qui coïncide avec les estimateurs (5.2) et (5.10).
Finalement, notons que le ratio

$$\frac{v_{2BP}}{v_1} = \frac{n}{N}\frac{\left(1 - \dfrac{r}{n}\right)}{\left(1 - \dfrac{n}{N}\right)} \to 0 \text{ si } n/N \to 0.$$

5.4. Le jackknife

Le jackknife proposée par Quenouille (1949) est une méthode de réplication qui permet d'estimer la variance dans le cas de plans et/ou de paramètres complexes. Supposons que l'on veuille estimer un paramètre q . Pour cela, on tire un échantillon aléatoire, *s*, de taille *n*, selon un plan de sondage *p*(.).

Soit $\hat{q}$ un estimateur sans biais (ou approximativement sans biais) de q basé sur *s*. En l'absence de non-réponse, le jackknife fonctionne comme suit :

(i) Enlever une unité (ou groupe d'unités).
(ii) Ajuster les poids de sondages.
(iii) Calculer l'estimateur $\hat{q}$ avec les poids ajustés.
(iv) Replacer l'unité enlevée à l'étape (i), enlever la prochaine unité.
(v) Répéter (i)-(iv) jusqu'à ce que toutes les unités aient été enlevées chacune à leur tour.

La variance jackknife de $\hat{q}$ est obtenue par

$$v_J(\hat{q}) = \frac{n-1}{n}\sum_{j\in s}(\hat{q}_{(j)} - \hat{q})^2$$

où $\hat{q}_{(j)}$ est calculé de la même manière que $\hat{q}$ lorsque la j^e unité a été enlevée, $j = 1,\ldots,n$.

Notons que $\hat{q}_{(j)}$ est calculé avec les poids de sondage ajustés $w_{i(j)}$ donnés 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. En effet, 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 donné par

$$v_J(\bar{y}) = \frac{n-1}{n}\sum_{j\in s}(\bar{y}_{I(j)} - \bar{y}_I)^2 = \frac{s^2_{yI}}{n} \qquad (5.16)$$

où

$$\bar{y}_{I(j)} = \begin{cases} \dfrac{1}{n-1}\left[n\bar{y}_I - y_j\right] & \text{si } j \in s_r \\ \dfrac{1}{n-1}\left[n\bar{y}_I - y^*_j\right] & \text{si } j \in s_m, \end{cases}$$

et s^2_{yI} est donné par (4.1).

Dans le cas de l'imputation par la moyenne, on a $\bar{y}_I = \bar{y}_r$, auquel cas la variance (5.16) devient

$$v_J(\bar{y}_I) = \frac{r-1}{n-1}\frac{s^2_{yr}}{n},$$

qui coïncide avec l'estimateur incorrect de la variance (5.3) 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 chacun des réplicats. Par exemple, dans le cas de l'imputation par la moyenne, les valeurs imputées ajustées, y_i^{*a}, sont données par

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

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

$$v_{JRS}(\bar{y}_I) = \frac{s_{yr}^2}{r}, \tag{5.17}$$

qui coïncide avec l'estimateur correct de la variance (5.2) 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 appréciable, 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

- Le jackknife ajusté peut être appliqué à plusieurs méthodes d'imputation : (hot-deck aléatoire, moyenne, ratio, régression, etc.)
- Le jackknife ajusté peut être utilisé pour des plans complexes (stratifié à degrés multiples).
- La méthode est intensive du point de vue informatique. Un programme efficace permettra toutefois de considérablement réduire le temps d'exécution.
- La méthode ne peut être utilisée dans le cas de paramètres « non-lisses » (quantiles, etc.).
- La méthode suppose que l'échantillon est tiré avec remise. Les résultats seront donc valides si la fraction de sondage n/N est petite.

5.5. Le bootstrap

Le boostrap (Efron, 1979) est, comme le jackknife, une autre méthode de réplication 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 q. Pour cela, on tire un échantillon aléatoire simple sans remise, s, de taille n, d'une population U de taille N. Soit $\hat{q}_I$ un estimateur imputé basé sur les valeurs imputées et observées. Le bootstrap de Shao et Sitter peut être décrit comme suit :

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

(ii) 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 $\hat{q}_I$, au moyen des unités répondantes dans s^* (i.e., les unités dans s_r^*).

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

(iv) Répéter (i)-(iii) *B* fois

L'estimateur de variance bootstrap de $\hat{q}_I$ est donné par

$$v_B\left(\hat{q}_I\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(q_{I(b)}^* - \hat{\bar{q}}_I\right)^2, \tag{5.18}$$

où $\hat{\bar{q}}_I = \frac{1}{B}\sum_{b=1}^{B} q_{I(b)}^*$.

Shao et Sitter (1996) on montré que, sous un mécanisme uniforme, l'estimateur (5.18) est convergent. Notons, que comme le jackknife, l'estimateur (5.18) est un estimateur de la composante V_1 de l'approche renversée. Donc, si la fraction de sondage *n*/*N* est appréciable, l'estimateur (5.18) est biaisé puisqu'il n'inclut pas la composante non-négligeable V_2 de l'approche renversée.

Caractéristiques du bootstrap

- Le bootstrap peut être appliqué à plusieurs méthodes d'imputation : (hot-deck aléatoire, moyenne, ratio, régression, etc.)
- Le bootstrap peut être utilisé pour des plans complexes (stratifié à degrés multiples) mais certaines études théoriques et empiriques restent à faire.
- La méthode est très intensive du point de vue informatique.
- Contrairement au jackknife, le bootstrap peut être utilisée dans le cas de paramètres « non-lisses » (quantiles, etc.)
- La méthode suppose que l'échantillon est tiré avec remise. Les résultats seront donc valides si la fraction de sondage *n*/*N* est petite.
- 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

Jusqu'à maintenant, nous avons discuté de l'inférence en présence de valeurs imputées pour des paramètres simples tels des totaux ou des moyennes. 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*, $\bar{Y}_d$, peut s'écrire comme

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

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

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

où **x** est un vecteur de variables auxiliaires disponible pour toutes les unités dans l'échantillon. Notons que la moyenne d'un domaine $\overline{Y}_d$ 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

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

Considérons le cas du coefficient de corrélation (6.3). En présence de valeurs imputées, l'obtention d'un estimateur imputé approximativement sans biais pour r_{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. 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 $\sum_{i\in U} x_i y_i$ est loin d'être 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. La littérature à propos de l'inférence pour des paramètres complexes en présence de valeurs imputées est peu abondante (c.f., Santos (1981), Shao et Wang (2002), Skinner et Rao (2002)). Un estimateur imputé approximativement sans biais de $\sum_{i\in U} x_i y_i$ peut être obtenu d'au moins deux façons :

- Il est possible d'utiliser une méthode d'imputation sophistiquée qui mène directement à un estimateur approximativement sans biais.

- Il est possible d'utiliser une méthode d'imputation simple (moyenne, ratio, hot-deck aléatoire, etc.) et d'utiliser un estimateur plus sophistiqué.

Dans cette section, nous discutons du problème de l'estimation pour des domaines. Nous considérons brièvement du cas d'un coefficient de corrélation. Finalement, nous discutons de l'utilisation de méthodes d'imputation pondérées par opposition à l'utilisation de méthodes d'imputation non-pondérées.

6.1. Les domaines

Les domaines d'intérêt ne sont pas toujours connus au stade de l'imputation. En effet, une fois le fichier complété après imputation, il est souvent envoyé à plusieurs chargés d'étude qui s'intéressent potentiellement à des domaines différents. Ces domaines ne sont donc pas toujours pris en compte lors de la construction des valeurs imputées. L'estimateur imputé de la moyenne $\overline{Y}_d$, est donné par

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

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 une population de taille N = 11270 individus. Cette population contient deux variables : *Revenu hebdomadaire* (RH) et *âge de l'individu*. Cette population est extraite d'un échantillon de l'Enquête sur la population active Canadienne pour le mois de Janvier 2001. 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. Un simple coup d'oeil 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

Age	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 est le plus près de la moyenne de la population. De cette population, R = 5000 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 :

- $y_i^* = \bar{y}_r$, la moyenne globale des répondants qui ne tient pas compte des domaines d'intérêt.
- $y_i^* = \bar{y}_{dr}$, la moyenne des répondants à l'intérieur du domaine d'intérêt (et donc, qui tient compte du domaine en question).

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 $\bar{Y}_d$ et la moyenne globale de la population $\bar{Y}$, ce qui est confirmé par le résultat suivant :

Proposition 2 : Sous l'approche BP et imputation par la moyenne, $y_i^* = \bar{y}_r$, le biais de l'estimateur imputé (6.4) est donné par

$$\text{Biais}\left(\bar{y}_{dI}\right) = E_p E_r\left(\bar{y}_{dI} \mid s\right) - \bar{Y}_d \approx (1-p)\left(\bar{Y} - \bar{Y}_d\right). \qquad (6.5)$$

Le biais en (6.5) est nul dans le cas de réponse complète (p = 1) ou lorsque $\bar{Y} = \bar{Y}_d$. Si l'on ne tient pas compte des domaines au stade de l'imputation, il est quand même possible d'obtenir un estimateur approximativement sans biais de la moyenne $\bar{Y}_d$. En effet, Haziza et Rao (2001b) ont proposé une correction pour le biais, ce qui mène à l'estimateur ajusté

$$\bar{y}_I^a = \hat{p}^{-1}\bar{y}_{dI} + \left(1 - \hat{p}^{-1}\right)\bar{y}_I, \qquad (6.6)$$

où $\hat{p}$ est un estimateur de la probabilité de réponse p, $\bar{y}_{dI}$ est donné par (6.4) et $\bar{y}_I$ est donné par (2.2). L'estimateur ajusté (6.6) est approximativement sans biais sous les approches BP et BM sous le modèle d'imputation (5.9). Cet estimateur est donc robuste au sens qu'il est approximativement sans biais sous les deux approches. Haziza et Rao (2001b) ont généralisé ce résultat au cas d'un coefficient de régression et imputation par régression.

Tableau 6 : Biais relatif (%) de l'estimateur imputé (6.4)

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

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, approche BP 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 (2002) ont généralisé les résultats de Skinner et Rao (2002) au cas de plans de sondage stratifiés à degrés multiples, approches BP et BM et certaines variantes de l'imputation par hot-deck aléatoire. L'estimateur ajusté obtenu est approximativement sans biais sous les approches BP et BM.

6.3. Imputation pondérée vs imputation non-pondérée

Dans le cas de plans de sondage à probabilités inégales (stratifié à degrés multiples, proportionnel à la taille, etc.), il est possible d'utiliser une méthode d'imputation pondérée ou d'utiliser une méthode d'imputation non-pondérée. L'imputation pondérée, contrairement à l'imputation non-pondérée, tient compte du plan de 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 à probabilités égales (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, en effet, plus attrayantes pour l'utilisateur car elles ont l'avantage d'être plus simples. Dans ce cas cependant, l'estimateur imputé sera vraisemblablement biaisé. Pour illustrer la problématique, considérons le cas d'un échantillon aléatoire, *s*, de taille *n*, tiré d'une population *U* de taille *N*. Nous étudions le biais de l'estimateur imputé (2.2) dans le cas de l'imputation par hot-deck aléatoire pondéré (HDP) et imputation par hot-deck aléatoire non-pondéré (HDNP). L'imputation HDNP utilise les valeurs imputées (2.11). Dans le cas de l'imputation HDP, la valeur manquante est remplacée par la valeur d'un répondant tiré au hasard (avec remise) de l'ensemble des répondants s_r tel que

$$y_i^* = y_j, \quad j \in s_r \text{ , tel que } P\left(y_i^* = y_j\right) = w_j \Big/ \sum_{i \in s_r} w_i \,. \tag{6.7}$$

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

$$\mathrm{BR}\left(\bar{y}_I\right) = \frac{E_p E_r\left(\bar{y}_I \mid s\right) - \bar{Y}}{\bar{Y}} \approx (1-p) C_y C_p \, r_{py}, \tag{6.8}$$

où $C_y = \dfrac{S_y}{\bar{Y}}$ et $C_p = \dfrac{S_p}{\bar{\Pi}}$ dénotent respectivement les coefficients de variation de la variable d'intérêt *y* et de la probabilité d'inclusion p, et r_{py} dénote le coefficient de corrélation entre la variable d'intérêt *y* et de la probabilité d'inclusion p. Si $C_y > 0$, le biais relatif est nul lorsque :

- $p = 1$ (cas de réponse complète)
ou
- $C_p = 0$ (cas d'un plan de sondage à probabilités égales)
ou
- $r_{py} = 0$.

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 r_{py} 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 si la corrélation r_{py} est grande et que l'on utilise une méthode d'imputation non-pondérée, il est possible d'obtenir un estimateur sans biais de la moyenne $\overline{Y}$, (Haziza et Rao, 2003).

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. Au lieu de former des classes, il est toujours possible d'imputer directement des valeurs à partir d'un modèle de régression. Cependant, il y a au moins deux raisons motivant l'utilisation de classes: (1) c'est plus pratique quand il s'agit d'imputer plusieurs variables à la fois et (2) les classes apportent une certaine robustesse par rapport à l'utilisation de l'imputation par 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 $p_i, i = 1,...,N$. Un estimateur imputé pour $\overline{Y}$ basé sur une seule classe d'imputation (en d'autres mots, l'échantillon *s*), est défini par

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

où s_r est l'ensemble des unités qui ont répondu à l'item *y*, s_m est l'ensemble des unités qui n'ont pas répondu à l'item *y*, et y_i^* est la valeur imputée créée afin de "boucher le trou" de la valeur manquante y_i. 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

$$Biais(\overline{y}_I) = E(\overline{y}_I) - \overline{Y} \approx \frac{1}{N\overline{P}}\sum_{i\in U}(p_i - \overline{P})(y_i - \overline{Y}) \qquad (7.2)$$

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 appellera $\bar{y}_{I,1}$, un estimateur « non-ajusté » . Pour réduire le biais dû à la non-réponse, il est coutume 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 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 » basé sur C classes

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

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,

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

dénote l'estimateur imputé pour la classe g, g = 1,…, C. Dans le cas de l'imputation par hot-deck aléatoire à l'intérieur des classes, le biais de l'estimateur ajusté est donné par

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

où $\bar{P}_g = \dfrac{1}{N_g}\sum_{i \in U_g} p_i$ et $\bar{Y}_g = \dfrac{1}{N_g}\sum_{i \in U_g} y_i$. 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 $\bar{y}_{I,1}$ et $\bar{y}_{I,C}$ 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 au stade 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 catégoriques. 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 $8^5 = 390625$ classes. Par conséquent, un bon nombre de classes pourrait contenir peu ou pas d'unités, ce qui mènera potentiellement à des estimations instables. En pratique, on spécifie certaines contraintes pour assurer une certaine stabilité des estimations. D'une part, on peut spécifier que le nombre de répondants à l'intérieur d'une classe soit plus grand ou égal à un certain seuil. D'autre part, on peut spécifier qu'à l'intérieur d'une classe, la proportion de répondants à l'intérieur d'une classe soit supérieure ou égale 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 et 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 et l'autre pour prédire la variable d'intérêt. L'estimation des probabilités de réponse peut être effectuée, par exemple, en utilisant la régression logistique. La prévision de la variable d'intérêt dépend de la nature de celle-ci (continue, catégorique,…). Deux scores, $\hat{p}$ et $\hat{y}$, sont alors disponibles 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, $\hat{p}$ ou $\hat{y}$. 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 d'abord ordonner les valeurs par rapport au critère choisi et à diviser l'échantillon en classes de tailles approximativement égales. Une approche alternative est d'utiliser un algorithme de classification pour former les classes. Notons qu'il est également possible d'utiliser simultanément les deux critères d'homogénéité pour former les classes.

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

(iv) Combiner les estimateurs imputés intra-classes tel que décrit en (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 en présence d'imputation par hot-deck aléatoire.

7.3.1. Création de la population et enjeux de la simulation

Une population de taille $N = 2000$ 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. Le modèle utilisé pour générer la variable y est le modèle de régression

$$y_i = b_0 + b_1 z_{1i} + b_2 z_{2i} + e_i, \qquad (7.6)$$

avec $e_i \sim \mathrm{N}(0, s^2)$,

$z_{ji} \sim Exp(30), \quad j = 1,2.$

Les variables z_3 et z_4 ont été générées indépendamment de y telles que $z_{ji} \sim \mathrm{Exp}(30)$, $\mathrm{j} = 3,4$. Les paramètres du modèle $ß_0, ß_1$ et $ß_2$ ont respectivement été fixés à 20, 0.5 et 0.5. Finalement, nous avons généré les données de telle sorte que le R^2 du modèle (7.6) soit approximativement égal à 0.8. 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 catégoriques seulement (voir section 7.2.2). La moyenne de la population ainsi créée est égale à $\overline{Y} = 49.87$. De cette population, $R = 1000$ échantillons de taille $n = N = 2000$ (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.

7.3.2. Méthode par croisement

Les classes sont formées à partir de combinaisons des 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.
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 réalisée. Certaines classes ne les respecteront pas. Dans ce

cas, on ôte la variable la moins significative et on croise les $q-2$ variables restantes. On répète le processus jusqu' à que chaque receveur ait trouvé un donneur.

7) Un estimateur imputé de la moyenne $\overline{Y}$ est alors donné par (2.2) .

Pour cette méthode, les points 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 (3.8) de l'estimateur imputé (2.2).

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 points :

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 score 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 score servant à former les classes : $\hat{y}$ et $\hat{p}$.
- ✓ 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 :
  - $y = ß_0 + ß_1 z_1 + ß_2 z_2$ (bon modèle)
  - $y = ß_0 + ß_2 z_2$ (mauvais modèle avec une variable en moins)

Les graphiques 3.1-3.5, présentés dans l'Annexe 3, exhibent le biais relatif (3.8) de l'estimateur « ajusté » (7.3).

7.3.4. Discussion des résultats

Méthode par croisement : Annexe 2

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

- 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_2 est vraisemblablement éliminée de la liste $(z_{1c}\, z_{2c}\, z_{3c}\, z_{4c})$, ce qui explique le biais de l'estimateur.

- Pour le mécanisme ignorable 2, nous constatons que le biais reste petit même quand le nombre minimal de donneurs augmente. Ce résultat s'explique par le fait que la variable z_1 n'est jamais éliminée de la liste. Elle est donc présente dans le modèle d'imputation même lorsque les contraintes deviennent plus sévères.

- Pour le mécanisme non-ignorable 4, l'estimateur est biaisé dans tous les cas de figure comme il fallait s'y attendre. Le biais augmente que le nombre minimal de donneurs augmente.

- Lorsque les variables classées dans le mauvais ordre $(z_{4c}\, z_{3c}\, z_{2c}\, z_{1c})$, les estimateurs sont d'emblée biaisés. Ceci s'explique facilement que les variables les plus significatives sont les premières à être éliminées lorsque les contraintes ne sont pas satisfaites.

- Lorsque la contrainte "plus de donneurs que de receveurs" est "activée," 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 le biais.

Méthode des scores : Annexe 3

- Les graphiques 3.1-3.4 montrent que la méthode des quantiles égaux et l'algorithme de classification mènent à des résultats très similaires.

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

- 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 compte tenu de l'expression du biais (7.2).

- 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 augemente (voir Graphique 3.1- 3.4).

- Pour le mécanisme non-ignorable 4 et le bon modèle d'imputation, bien que le biais diminue lorsque le nombre de classes augmente, ce dernier ne devient pas négligeable (voir Graphique 3.1et 3.2). Ce résultat n'est pas surprenant comme l'indique l'expression du biais (7.5).

- 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 superflue.

- 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 :

- L'utilisation d'un mauvais modèle d'imputation mènera presque toujours à des estimateurs biaisés.

- La méthode par croisement est sensible aux contraintes et au mauvais classement des variables dans la liste.

- L'estimation de la variance sera vraisemblablement ardue dans le cas de la méthode par croisement alors qu'elle sera relativement aisée dans le cas de la méthode des scores, puisque dans ce cas, les classes sont disjointes.

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

- Pour la méthode des scores, il est à prévoir que l'utilisation du score $\hat{p}$ mènera à des estimateurs avec une plus grande erreur quadratique moyenne que ceux obtenus lorsque le score $\hat{y}$ est utilisé.

Bibliographie

[1] Chen, H., Rao, J. N. K., Sitter, R. R., "Efficient random imputation for missing data in complex surveys", *Statistica Sinica*, vol 10, pp 1153-1169, 2000.

[2] Deville, J. C., Särndal, C. E., "Variance estimation for the regression imputed Horvitz-Thompson estimator", *Journal of Official Statistics*, vol 10, pp 381-394, 1994.

[3] Efron, B.,"Bootstrap methods: another look at the jackknife", *Annals of Statistics*, vol 7, pp 1-26, 1979.

[4] Eltinge, J. L., Yansaneh, I. S.,"Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U. S. Consumer Expenditure Survey", *Survey Methodology*, vol 23, pp 33-40, 1997.

[5] Fay, R. E., "A design-based perspective on missing data variance", *Proceedings of the 1991 Annual Research Conference, U.S. Bureau of the Census*, pp 420-440, 1991.

[6] Hansen, M. H., Hurvitz, W. N., Madow, W. G. "Sample Survey Methods and theory", vol I et II, New York Wiley, 1953.

[7] Haziza, D., Beaumont, J. F., "Construction des classes d'imputation dans les enquêtes", *manuscrit*, 2002.

[8] Haziza, D., Rao, J. N. K., "Model-assisted approach to inference for totals under imputation for missing data in two-stage cluster sampling", *Proceedings of the Survey Methods Section, American Statistical Association*, 2001a.

[9] Haziza, D., Rao, J. N. K., "Inference for regression coefficient under imputation for missing data", *Proceedings of the Survey Methods Section, Statistical Society of Canada*, pp 420-440, 2001b.

[10] Haziza, D., Rao, J. N. K., "Inference for bivariate statistics under imputation for missing survey data in stratified multistage sampling", *manuscrit*, 2002.

[11] Haziza, D., Rao, J. N. K., "Inference for population means under unweighted imputation for missing survey data", *Survey Methodology*, à paraître, 2003.

[12] Kalton, G., Kasprzyk, D., "The treatment of missing survey data", *Survey Methodology*, vol 12, pp 1-16, 1986.

[13] Kovar, J. G., Whitridge, P. J., "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, pp 403-423, 1995.

[14] Little, R. J. A., "Survey nonresponse adjustments", *International Statistical Review*, vol 54, pp 139-157, 1986.

[15] Oh, H. L., Scheuren, F. J., "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, pp 143-184, 1983.

[16] Qin, J., Leung, D., Shao, J.,"Estimation with survey data under nonignorable nonresponse or informative sampling", *Journal of the American Statistical Association*, vol 97, pp 193-200, 2002.

[17] Quenouille, M., "Approximation tests of correlation in time series", *Journal of the Royal Society*, pp 18-84, 1949.

[18] Rancourt, E., "Edit and imputation: from suspicious to scientific techniques", *Actes, l'Association internationale des statisticiens d'enquête*, pp 605-633, 2001.

[19] Rao, J. N. K.,"Variance estimation under imputation for missing data", *Technical report, Statistics Canada, Ottawa*, 1990.

[20] Rao, J. N. K., Shao, J.,"Jackknife variance estimation with survey data under hot-deck imputation", *Biometrika*, vol 79, pp 811-822, 1992.

[21] Rao, J. N. K., Sitter, R. R.,"Variance estimation under two-phse sampling with application to imputation for missing data", *Biometrika*, vol 82, pp 453-460, 1995.

[22] Rubin, D. B., "Inference and missing data", *Biometrika*, vol 63, pp 581-590, 1976.

[23] Rubin, D. B.,"Multiple imputation in sample surveys- a phenomenological Bayesian approach to nonresponse", *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp 20-34, 1978.

[24] Santos, R., "Effect of imputation on regression coefficients", *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp 140-145, 1981.

[25] Särndal, C. E.,"Methods for estimating the precision of survey estimates when imputation has been used", *Proceedings of Symposium 1990*, Measurement and improvement of data quality, pp 337-347, 1990.

[26] Särndal, C. E., "Methods for estimating the precision of survey estimates when imputation has been used", *Survey Methodology*, vol 18, pp 241-252, 1992.

[27] Shao, J., Sitter, R. R.,"Bootstrap for imputed survey data", *Journal of the American Statistical Association*, vol 91, pp 1278-1288, 1996.

[28] Shao, J., Steel, P., "Variance estimation for survey data with composite imputation and nonnegligible sampling fractions", *Journal of the American Statistical Association*, vol 94, pp 254-265, 1999.

[29] Shao, J., Wang, H., "Sample correlation coefficients based on survey data under regression imputation", *Journal of the American Statistical Association*, vol 97, pp 544-552, 2002.

[30] Skinner, C.J., Rao, J. N. K., "Jackknife variance estimation for multivariate statistics under hot-deck imputation from common donors", *Journal of Statistical Planning and Inference*, pp 149-167, 2002.

Annexe 1

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

Mécanisme 1

mécanisme uniforme, c'est-à-dire $p_i = 0.7 \; \forall i$.

Mécanisme 2

Mécanisme ignorable qui dépend de z_1 tel que

$$\log\left(\frac{p_i}{1-p_i}\right) = \exp\left(l_0 + l_1 z_1\right)$$

où l_0 et l_1 sont choisis de manière à obtenir un taux global de réponse de 70%.

Mécanisme 3

Mécanisme ignorable : idem mécanisme 2 mais remplacer z_1 par z_2.

Mécanisme 4

Mécanisme non-ignorable : idem mécanisme 2 mais remplacer z_1 par *y*.

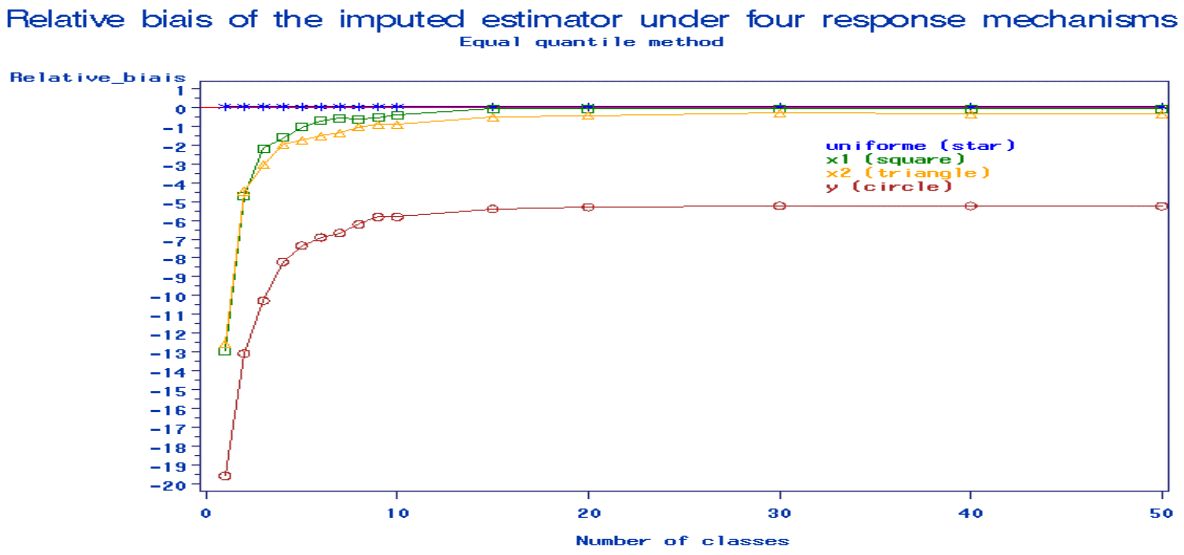
Annexe 2

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
z1c z2c z3c z4c	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
z4c z3c z2c z1c	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

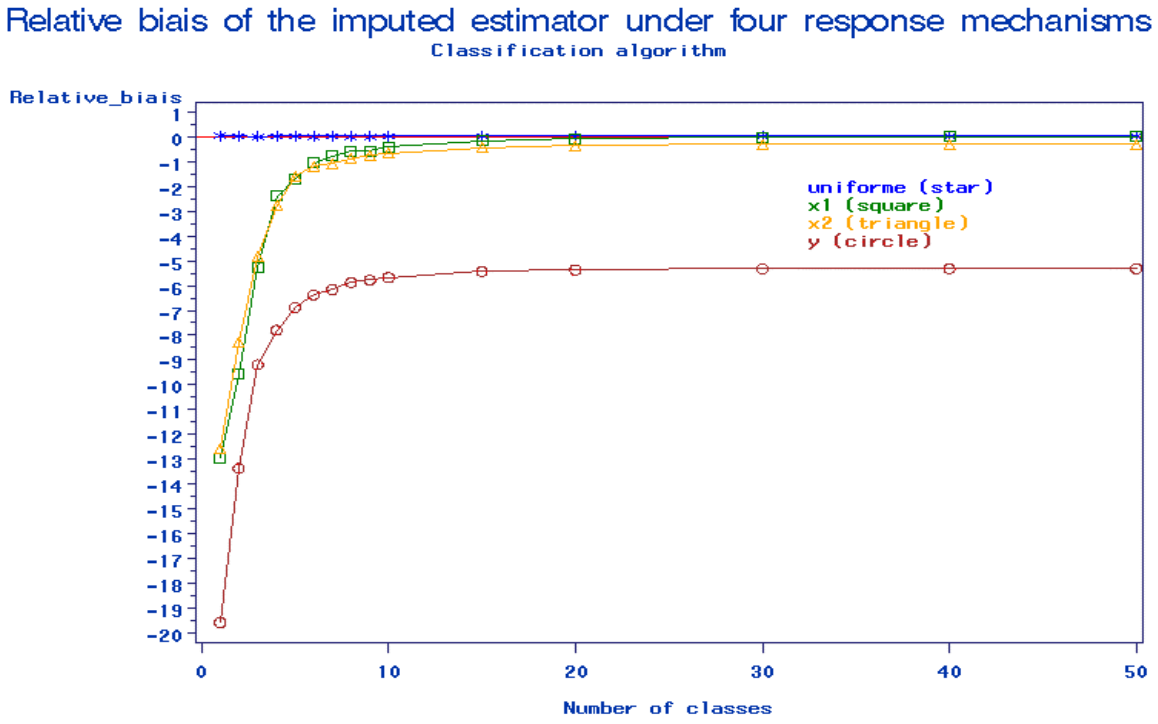
Annexe 3

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



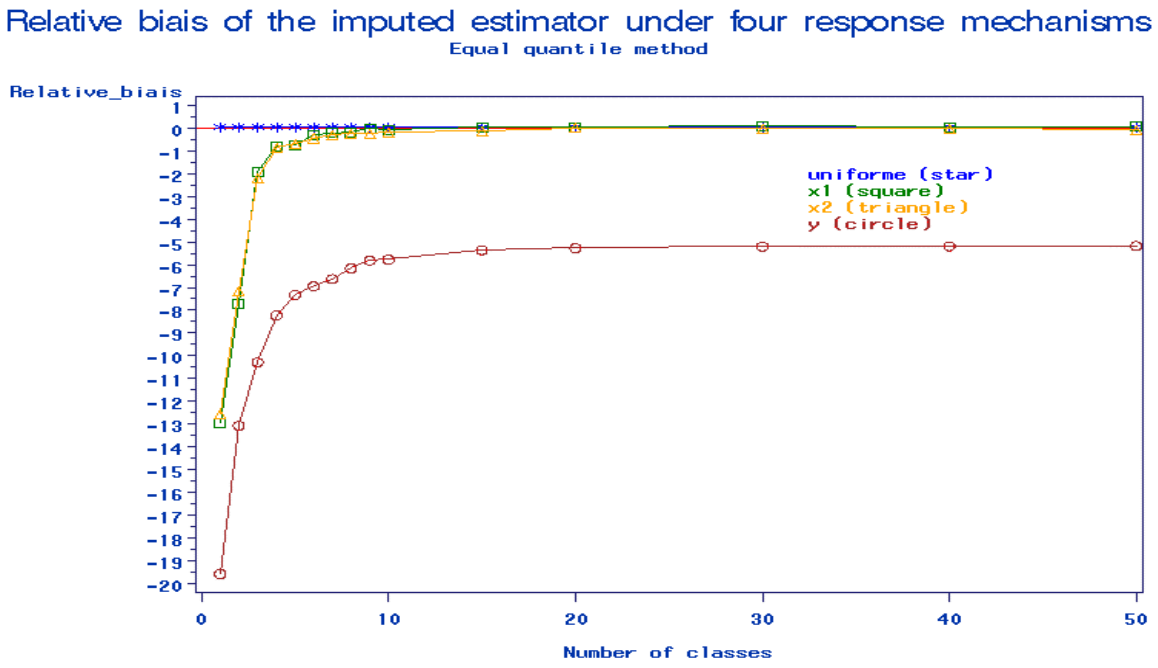
note1: 2000 repetitions of 1000 units
note2: Response rate : 70 %
note3: population mean 49.870572279

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



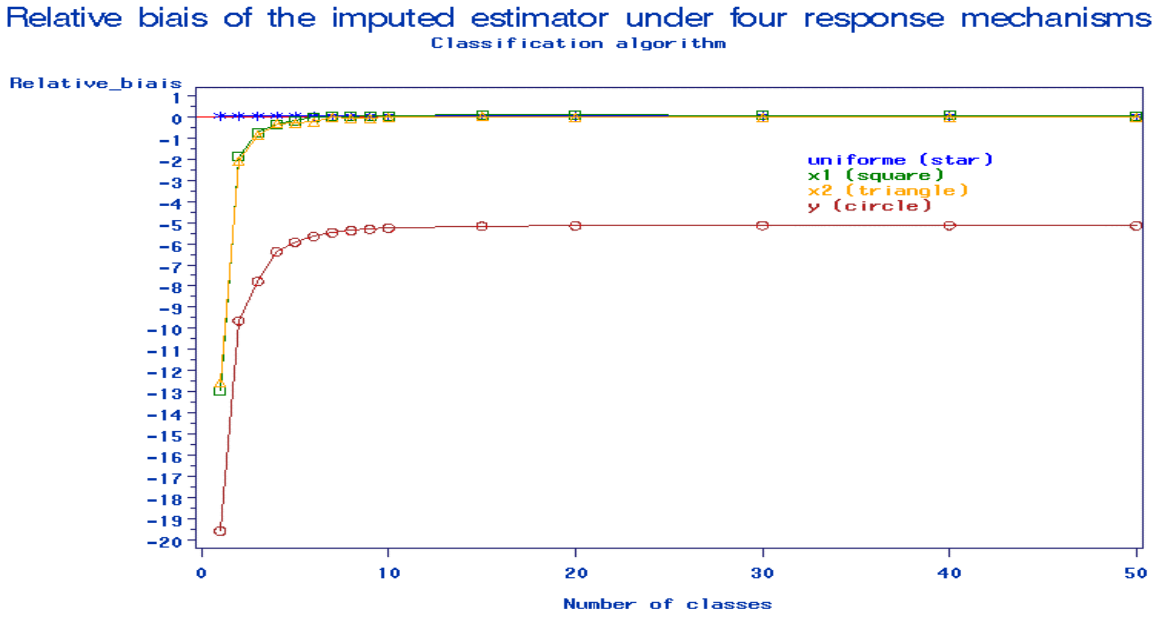
note1: 1000 repetitions of 2000 units
note2: Response rate : 70 %
note3: population mean 49.870572279

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



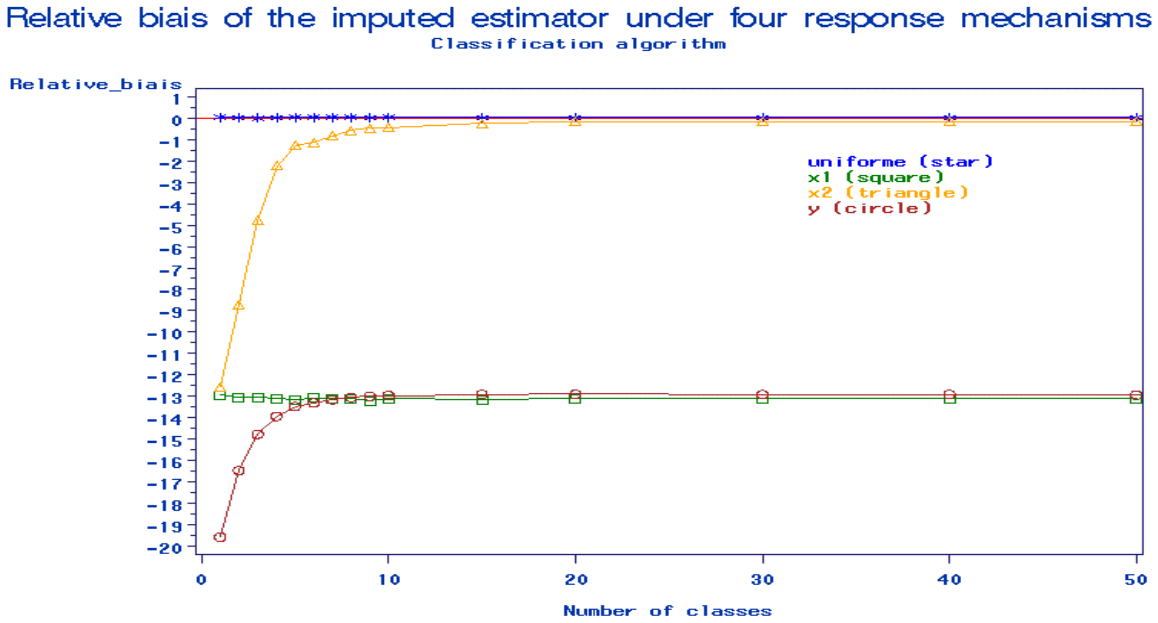
note1: 2000 repetitions of 1000 units
note2: Response rate : 70 %
note3: population mean 49.870572279

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



note1: 1000 repetitions of 2000 units
note2: Response rate : 70 %
note3: population mean 49.870572279

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



note1: 1000 repetitions of 2000 units
note2: Response rate : 70 %
note3: population mean 49.870572279