Techniques d'enquête, juin 1986
Vol. 12, n° 1, pp. 1–17
Statistique Canada

Le traitement des données d'enquête manquantes

GRAHAM KALTON et DANIEL KASPRZYK[1]

RÉSUMÉ

La non-réponse intégrale et la non-réponse à une question sont les causes des données d'enquête manquantes. On corrige habituellement la non-réponse intégrale par une certaine correction des pondérations, tandis que la non-réponse à une question est corrigée par une sorte d'imputation. Les auteurs examinent les méthodes de correction par pondération et imputation ainsi que leurs propriétés.

MOTS CLÉS: Non-réponse; non-réponse à une question; corrections de pondération; imputation.

1. INTRODUCTION

En règle générale, les enquêtes recueillent les réponses à un grand nombre de questions pour chaque élément échantilloné. Le problème des données manquantes se pose lorsque une partie ou la totalité des réponses ne sont pas recueillies pour un élément échantillonné, ou lorsque certaines réponses sont supprimées parce qu'elles ne répondent pas aux critères de vérification. On distingue habituellement entre la non-réponse intégrale (ou à un questionnaire), quand aucune réponse à l'enquête n'existe pour un élément échantillonné, et la non-réponse à une question, quand quelques-unes des réponses, mais pas toutes, sont données. La non-réponse intégrale résulte de refus, de l'incapacité à participer à l'enquête, d'absences et d'éléments non relevés. La non-réponse à une question s'explique par des refus, l'ignorance, des omissions et des réponses supprimées lors de la vérification.

On examine dans cette communication des méthodes générales qui existent pour le traitement des données d'enquête manquantes. La distinction entre non-réponse intégrale et non-réponse à une question est utile ici, puisqu'on utilise des méthodes de correction différentes dans les deux cas. En général, la seule information existante à propos de l'ensemble des non-répondants est celle concernant la base de sondage d'où l'échantillon a été tiré (les strates et les UPÉ dans lesquelles ils se trouvent, par exemple). Il est habituellement possible d'incorporer rapidement les aspects importants de ces renseignements dans des corrections de pondérations afin d'essayer de compenser les données manquantes. On se sert habituellement des corrections de pondérations en règle générale pour la non-réponse intégrale. On examine à la section 2 les méthodes pour faire de telles corrections.

Par contre, dans le cas de la non-réponse à une question, on dispose d'un grand nombre d'informations supplémentaires pour les éléments en cause, c'est-à-dire non seulement les renseignements provenant de la base de sondage, mais également les réponses aux autres questions de l'enquête. Afin de conserver toutes les réponses pour des éléments contenant des non-réponses à des questions, la méthode habituelle de correction donne des enregistrements d'analyse qui incorporent des réponses réelles aux questions pour lesquelles les réponses ont été acceptables, et des réponses imputées aux autres questions. Les méthodes d'imputation pour l'affectation des réponses aux questions manquantes sont examinées à la section 3.

En général, le choix entre les corrections de pondérations et l'imputation pour le traitement des données d'enquête manquantes est assez aisé, mais le choix n'est pas toujours aussi clair. On peut citer des cas de ce que l'on pourrait appeler une non-réponse partielle, lorsque certaines données sont obtenues pour un élément échantillonné, mais une quantité appréciable de données est manquante. La non-réponse partielle peut se produire par exemple lorsque un

[1] Graham Kalton, Survey Research Center, University of Michigan, Ann Arbor, Michigan, 48106-1248 et Daniel Kasprzyk, Population Division, U.S. Bureau of the Census, Washington, D.C., 20233. Les auteurs aimeraient remercier les critiques pour leurs commentaires utiles.

répondant met fin prématurément à une interview, lorsque des données ne sont pas recueillies pour un ou plusieurs membres d'un ménage autrement coopératif (pour l'analyse au niveau des ménages) ou lorsqu'une personne échantillonnée fournit des données pour une partie seulement des questions d'un panel. Cox et Cohen (1985) et Kalton (1986) examinent le choix entre la pondération et l'imputation pour compenser la non-réponse lors d'une enquête par panel.

Bien que les corrections de pondérations et d'imputation soient traitées comme des approches distinctes dans la discussion qui suit, elles sont en faites étroitement reliées. La relation et les différences entre ces deux approches sont examinées brièvement à la section 4, qui contient également quelques autres solutions au problème des données d'enquête manquantes.

2. CORRECTIONS DES PONDÉRATIONS

Ces corrections sont destinées avant tout à compenser la non-réponse intégrale. L'essence de toute procédure de correction de pondérations est d'accroître les poids de répondants précis de façon à ce que ces derniers représentent des non-répondants. Les procédures nécessitent des renseignements auxiliaires sur les non-répondants ou la population totale. Les quatres types suivants de correction des pondérations sont examinés brièvement ci-dessous: corrections des pondérations de la population, corrections des pondérations de l'échantillon, méthode itérative du quotient et poids basés sur les probabilités de réponse. Kalton (1983) contient davantage de détails.

2.1 Corrections des pondérations de la population

L'information auxiliaire utilisée dans la correction des pondérations de la population est la distribution de la population sur une ou plusieurs variables telles que la répartition de la population selon l'âge, le sexe et la race obtenues à partir d'estimations démographiques types. L'échantillon des répondants est partagé en un ensemble de classes, appelées ici classes de pondération, définies par l'information auxiliaire disponible (hommes blancs âgés de 15-24 ans, femmes non blanches âgées de 25-34 ans, etc.). Les poids de tous les répondants à l'intérieur d'une classe de pondération sont ensuite corrigés par le même facteur multiplicatif avec différents facteurs dans différentes classes. La correction est faite de façon à ce que la distribution des répondants pondérée pour l'ensemble des classes de pondération soit conforme à la distribution de la population.

On appelle souvent ce type de correction post-stratification. Nous évitons ce terme ici, cependant, parce que même si la pondération de population ressemble à la post-stratification, il existe une importante différence entre les deux. Tout comme la pondération de population, la post-stratification pondère l'échantillon de façon à rendre la distribution d'échantillonnage conforme à la distribution de la population pour un ensemble de classes (ou strates). Cependant, la théorie habituelle de la post-stratification ne porte que sur les fluctuations d'échantillonnage qui pourraient pousser la distribution d'échantillonnage à s'écarter de la distribution de la population, et non pas sur les écarts plus importants qui peuvent résulter de taux de réponse différents pour l'ensemble des classes. Les corrections de post-stratification ressemblent davantage à un affinement de l'échantillon, se traduisant en général par de faibles variations seulement des poids à travers les strates. Par conséquent, et à condition que les strates ne soient pas petites, la post-stratification se traduit par des erreurs types moins élevées pour les estimations d'enquête. Par contre, les corrections des pondérations de la population peuvent entraîner davantage de corrections majeures et causer des erreurs types plus élevées.

Les corrections des pondérations de la population tentent de réduire le biais entraîné par la non-réponse et les erreurs d'observation. Considérons l'estimation d'une moyenne de population $\bar{Y}$ à partir d'un échantillon dans lequel les éléments sont sélectionnés avec une

probabilité égale. Supposons que la population est divisée en un ensemble de classes de pondération, avec une proportion W_h d'éléments dans la classe h. Supposons de plus que les répondants répondent à chaque fois, et que les non-répondants n'existent pas. Soit R_h et M_h les proportions des répondants et non-répondants respectivement de la classe h, et soit $\bar{R} = \Sigma W_h R_h$ le taux de réponse global. Alors, d'après Thomsen (1973), le biais de la moyenne des répondants non corrigée ($\bar{y}$) peut s'écrire sous la forme

$$B(\bar{y}) = \bar{R}^{-1}\sum W_h(\bar{Y}_{rh} - \bar{Y}_r)(R_h - \bar{R}) + \sum W_h M_h(\bar{Y}_{rh} - \bar{Y}_{mh}) = A + B \qquad (1)$$

où $\bar{Y}_{rh}$ et $\bar{Y}_{mh}$ sont les moyennes des répondants et des non-répondants de la classe h respectivement, et $\bar{Y}_r$ est la moyenne de la population des répondants. L'emploi de la correction des pondérations de la population donne la moyenne de l'échantillon pondérée, $\bar{y}_p = \Sigma W_h \bar{y}_{rh}$, où $\bar{y}_{rh}$ est la moyenne de l'échantillon de répondants de la classe h. Le biais de $\bar{y}_p$ est simplement le second terme de $B(\bar{y})$, c'est-à-dire $B(\bar{y}_p) = B$.

Si A et B sont de même signe, la correction des pondérations de la population réduit de A le biais absolu dans l'estimation de $\bar{Y}$. Si $\bar{Y}_{rh} = \bar{Y}_{mh}$, ce qui se produit lorsqu'on s'attend à ce que les non-répondants soient absents au hasard à l'intérieur des classes de pondération, alors $B = 0$. Dans ce cas, la correction des pondérations de la population élimine le biais. Le terme A est un terme de genre covariance entre les taux de réponses de classe et les moyennes des répondants de classe. Il sera égal à 0 si le taux de réponse ou la moyenne des répondants ne varie pas entre classes. Dans l'un ou l'autre de ces cas, la correction des pondérations de la population n'a aucun effet sur le biais de l'estimateur. On peut observer que les corrections des pondérations de la population peuvent accroître le biais absolue de l'estimation de $\bar{Y}$. Ceci sera le cas lorsque A et B seront de signes opposés, et si $|A| < 2|B|$.

Les corrections des pondérations de la population nécessitent des données extérieures sur les distributions de la population pour les variables à utiliser. Il faut prendre soin de s'assurer que les données sur lesquelles reposent les distributions de population soient exactement comparables aux données d'enquête; sinon, on obtiendra des poids qui ne conviennent pas. Comme la procédure touche les distributions des populations, elle fait plus que simplement de corriger la non-réponse. Elle compense les erreurs de champ d'observation et apporte une correction de post-stratification.

2.2 Correction des pondérations de l'échantillon

Comme pour les corrections des pondérations de la population, les corrections des pondérations de l'échantillon consistent à diviser l'échantillon en classes de pondération; des poids différents sont alors affectés à ces classes afin d'essayer de réduire le biais de non-réponse. La différence essentielle entre les deux procédures réside dans les informations auxiliaires utilisées. Comme l'a décrit plus haut, les corrections des pondérations de la population sont basées sur des distributions de population d'origine externe. Aucune donnée n'est nécessaire pour les non-répondants d'échantillon. Par contre, les corrections des pondérations de l'échantillon n'utilisent que des données internes à l'échantillon et nécessitent des informations sur les non-répondants.

Dans le cas des corrections des pondérations de l'échantillon, les poids des corrections de non-réponse des classes de pondération sont établis de façon proportionnelle aux inverses des taux de réponse des classes. Afin de calculer ces taux de réponses, il faut déterminer le nombre de répondants et de non-répondants dans les classes. Il faut par conséquent connaître à quelle classe chaque répondant et non-répondant appartiennent. Comme en règle générale on ne dispose que de très peu d'informations sur les non-répondants, le choix de la classe de pondération est habituellement très sérieusement limité. Il se trouve souvent limité à des variables de plans de sondage généraux (UPÉ et strates), aux caractéristiques de ces

variables (urbaine/rurale, région géographique) et parfois, à quelques variables supplémentaires se trouvant dans la base de sondage. À l'occasion, il peut également être possible de recueillir des informations sur une ou deux variables pour les non-répondants, comme par exemple par observation de l'interviewer.

Tout comme les corrections des pondérations de la population ressemblent à la post-stratification, de même les corrections des pondérations de l'échantillon ressemblent à l'échantillonnage à deux degrés. L'échantillon du premier degré est l'échantillon total des répondants et des non-répondants, l'échantillon du deuxième degré est le sous-échantillon des répondants, sélectionné avec différentes fractions de sondage (taux de réponses) dans différentes strates (classes de pondération). La moyenne pondérée de l'échantillon peut être représentée par la formule $\bar{y}_s = \Sigma w_h \bar{y}_{rh}$, où w_h est la proportion de l'échantillon total dans la classe de pondération h. En supposant qu'il n'y a pas d'erreurs d'observation, $E(w_h) \doteq W_h$, la proportion de la population dans la classe h, telle qu'elle est utilisée dans l'estimateur pondéré de la population $\bar{y}_p = \Sigma W_h \bar{y}_{rh}$. Le biais de $\bar{y}_s$ est le même que celui de $\bar{y}_p$, à savoir $B(\bar{y}_s) = B$ d'après l'équation (1). Par conséquent, l'effet de la correction des pondérations de l'échantillon sur le biais de l'estimation d'enquête est le même que celui de la correction des pondérations de la population. Comme les corrections de pondérations de l'échantillon n'utilisent que des données pour l'échantillon, elles ne corrigent pas les erreurs d'observation (contrairement aux corrections des pondérations de la population).

Les deux types de correction ont des besoins en données différents, et rectifient ainsi des causes possibles différentes de biais. Dans la pratique, les deux types de correction sont utilisés en combinaison. En règle générale, les corrections des pondérations de l'échantillon sont appliquées d'abord, et les corrections des pondérations de la population, ensuite. Une façon habituelle de procéder est d'abord de déterminer les poids d'échantillon nécessaires pour corriger les probabilités inégales de sélection, ensuite de réviser ces poids afin de corriger les taux de réponse inégaux dans différentes classes de pondération d'échantillon (classes urbaines/rurales à l'intérieur de régions géographiques, par exemple), et enfin, de réviser les poids une nouvelle fois afin d'aligner la distribution d'échantillon pondérée pour certaines caractéristiques (âge, sexe, par exemple) sur la distribution d'une population connue pour ces caractéristiques. L'utilisation de cette méthode dans la U.S. Current Population Survey est décrite par Bailar *et al.* (1978).

Tout comme pour les corrections des pondérations de la population, les corrections des pondérations de l'échantillon visent à réduire le biais que la non-réponse pourrait causer dans les estimations d'enquête. Un des effets des corrections des pondérations de l'échantillon est d'augmenter les variances des estimations d'enquête. Il faut par conséquent trouver un compromis entre la réduction du biais et l'augmentation de la variance.

On peut avoir une idée de l'accroissement de la variance imputable à la pondération en examinant le cas où les variances des éléments à l'intérieur des classes de pondération sont toutes les mêmes et les variances entre les moyennes de classe sont négligeables lorsqu'on les compare aux variances intra-classes. Alors, la perte de précision résultant de la pondération est à peu près la même que celle provenant de l'utilisation d'un échantillon stratifié disproportionné, alors qu'un échantillon stratifié proportionné est optimal; Kish (1965, Section 11.7C; 1976) examine ce dernier cas.

Dans ces conditions, la pondération accroît la variance de la moyenne d'un échantillon d'approximativement $L = (\Sigma W_h k_h)(\Sigma W_h / k_h)$, où W_h est la proportion de la population, et k_h est le poids de la classe h. Une autre forme de L est $(\Sigma n_h)(\Sigma n_h k_h^2) / (\Sigma n_h k_h)^2$, où n_h est la taille de l'échantillon dans la classe h. Le facteur L devient grand lorsque la variance des poids est importante.

Une variance importante dans les poids peut résulter de la segmentation de l'échantillon en un grand nombre de classes de pondération avec peu d'éléments échantillonnés dans chacune. Lorsque les classes de pondération sont peu nombreuses, leurs taux de réponse sont instables, ce qui se traduit par une forte variation des poids. Afin d'éviter cette

situation, on limite habituellement le degré de segmentation de l'échantillon. Mais dans ce dernier cas, il peut y avoir encore des classes de pondération nécessitant des poids élevés. Parfois, ce problème est traité par fusionnement avec des classes voisines, et parfois leurs poids sont réduits à une valeur maximum acceptable quelconque (voir Bailar *et al.* 1978, et Chapman *et al.* 1986, pour des exemples). Ces procédés évitent l'augmentation de la variance associée à l'utilisation de poids extrêmes, mais ils peuvent se traduire par une augmentation du biais; leur effet sur ce dernier est cependant inconnu.

Dans certains cas, il semble souhaitable d'utiliser plusieurs variables auxiliaires dans la constitution des classes de pondération pour les corrections des pondérations de la population ou de l'échantillon. Cependant, si les classes sont constituées en prenant la classification recoupée intégrale des variables, on se trouvera en présence d'un grand nombre de classes de pondération. À moins que l'échantillon ne soit très grand, les tailles des échantillons dans les classes de pondération obtenues seront petites, et l'instabilité des taux de réponse se traduira par une importante variance des poids et par une perte de précision des estimations d'enquête. Une façon de traiter ce problème est de réduire le nombre de classes en fusionnant des cellules, en laissant de côté par exemple quelques-unes des variables auxiliaires, ou encore, en utilisant des classifications plus grossières. Une autre solution consisterait à baser les poids sur un modèle, comme on le fait dans le cas de la méthode itérative du quotient, examinée ci-dessous.

2.3 Corrections selon la méthode itérative du quotient

Lorsque les classes de pondération sont choisies comme étant des cellules de la classification recoupée des variables auxiliaires, les corrections des pondérations de la population rendent la distribution conjointe des variables auxiliaires de l'échantillon conforme à celle de la population. De même, les corrections des pondérations de l'échantillon rendent la distribution conjointe des variables auxiliaires de l'échantillon de répondants conforme à celle de l'échantillon total. Comme on l'a dit plus haut, cependant, cette méthode peut avoir l'effet indésirable de créer un grand nombre de petites classes de pondération, qui sont par conséquent instables. Par ailleurs, il n'est pas toujours possible d'utiliser cette méthode dans les corrections des pondérations de la population; souvent, les distributions marginales de la population, et peut-être, quelques distributions bivariées, des variables auxiliaires existent, mais la distribution conjointe intégrale est inconnue.

Une autre solution serait de définir des poids qui rendent les distributions marginales des variables auxiliaires de l'échantillon conformes aux distributions marginales de la population (pondérations de la population) ou aux distributions d'échantillon totales marginales (pondérations de l'échantillon), sans garantir pour cela la conformité de la distribution conjointe intégrale. La méthode itérative du quotient, ou plus simplement, méthode du quotient, peut être utilisée pour obtenir des poids qui répondent à ces conditions. La méthode du quotient correspond à l'ajustement proportionnel itératif dans l'analyse des tableaux de contingence (voir, par exemple, Bishop *et al.* 1975).

Considérons l'utilisation de la méthode du quotient dans le cas simple de deux variables auxiliaires. Soit W_{hk} la proportion de la population dans la cellule (h, k) de la classification recoupée, et soit $\tilde{w}_{hk}$ la proportion attribuée à cette cellule par l'algorithme de la méthode du quotient. Suivant les tailles de l'échantillon total et de celui des répondants dans les cellules, et en supposant que toutes les cellules ont au moins un répondant, le biais de la moyenne de l'échantillon corrigé de la méthode du quotient $\bar{y}_q = \Sigma\Sigma\tilde{w}_{hk}\bar{y}_{hk}$ est

$$B(\bar{y}_q) = \sum\sum W_{hk}M_{hk}(\bar{Y}_{rhk} - \bar{Y}_{mhk}) + \sum\sum(\tilde{W}_{hk} - W_{hk})(\bar{Y}_{rhk} - \bar{Y}_{rh.} - \bar{Y}_{r.k} + \bar{Y}_r)$$

où $\tilde{W}_{hk} = E(\tilde{w}_{hk})$. Le premier terme de ce biais correspond au terme du biais B dans l'équation (1) pour les corrections des pondérations de la population et de l'échantillon. L'espérance mathématique de ce terme est zéro si les non-répondants de la cellule sont des

sous-ensembles aléatoires des populations de la cellule. Le deuxième terme est zéro si $\tilde{W}_{hk} = W_{hk}$, ou s'il n'y a aucune interaction dans $\bar{Y}_{rhk}$ pour la classification.

À la base de la méthode du quotient se trouve un model logit pour les taux de réponse de cellules. Avec le model $\ln[R_{hk}/(1 - R_{hk})] = \alpha_h + \beta_k$ pour les taux de réponse dans une classification bidimensionnelle, $\tilde{W}_{hk} = W_{hk}$. Par conséquent, dans ce modèle, le deuxième terme dans $B(\bar{y}_q)$ est zéro.

La pondération par la méthode du quotient est examinée plus en détail par Oh et Scheuren (1978a, 1978b, 1983). Oh et Scheuren (1987a) donnent également une bibliographie sur la méthode itérative du quotient.

2.4 Pondération avec probabilités de réponse

Bien qu'un certain nombre de méthodes pour la pondération avec des probabilités de réponse aient été proposées, cette approche n'a pas encore été adoptée de façon générale comme procédure de correction. À la base de cette méthode, on suppose que tous les éléments de la population ont des probabilités (habituellement devant être différentes de zéro) de répondre à l'enquête. On utilise une méthode pour estimer les probabilités de réponse des éléments répondants. Ces derniers se voient à leur tour attribuer des poids de correction de la non-réponse inversement proportionnels à leurs probabilités de réponse estimatives.

Une des premières applications de cette méthode est le procédé bien connu de Politz et Simmons (1949, 1950). Un contact simple (le soir) est pris pour chaque ménage sélectionné, et au cours de l'interview, on demande aux répondants pendant combien des cinq soirs précédents ils se trouvaient à la maison à peu près à la même heure. Leurs probabilités de réponse sont ensuite calculées comme étant une fraction des six soirs (y compris celui de l'interview) pendant lesquels ils étaient à la maison, et les inverses de ces probabilités servent à l'analyse. Il est à noter que cette méthode ne prend pas en compte ceux qui étaient absents les six soirs, et ceux qui ont refusé de répondre.

Une autre façon d'estimer les probabilités de réponse est de procéder à la régression du statut de réponse (1 pour les répondants, 0 pour les non-répondants) sur un ensemble de variables existant pour les répondants et les non-répondants, en utilisant pour cela une régression logistique ou probit. Les valeurs prédites à partir de la régression pour les répondants sont ensuite retenues comme étant leurs probabilités de réponse, et les poids inversement proportionnels à ces valeurs prédites servent à l'analyse. Un cas spécial se pose lorsque les variables explicatives sont des variables dichotomiques qui identifient un ensemble de classes. Les probabilités de réponse prédites sont alors les taux de réponse de classe, et la méthode se ramène à une correction des pondérations de l'échantillon. La méthode convient le mieux aux situations lorsque l'on dispose de beaucoup d'informations pour les non-répondants, comme par exemple lorsque les non-répondants sont des pertes après la première vague d'une enquête par panel. Little et David (1983) ont examiné la possibilité d'appliquer cette méthode à la non-réponse par panel. Il convient de noter que si la régression prédit très bien le statut de réponse, les poids en résultant pourront varier de façon appréciable, ce qui se traduira par une perte sensible de la précision des estimations de l'enquête.

Drew et Fuller (1980, 1981) présentent une approche pour l'estimation des probabilités de réponse à partir du nombre de répondants certains à des contacts successifs. Dans leur modèle, la population est divisée en classes. À l'intérieur de chacune d'elle, chaque élément par hypothèse reçoit la même probabilité de réponse, qui demeure la même pour chaque contact. Le modèle prévoit également une proportion de non-répondants endurcis que l'on suppose constante pour les différentes classes. Dans ces hypothèses, les probabilités de réponse de chaque classe et la proportion de non-répondants endurcis peuvent être estimées, et par conséquent il est possible d'apporter des corrections des pondérations. Thomsen et Siring (1983) ont adopté une méthode semblable, utilisant un modèle plus complexe.

Enfin, il faut mentionner une approche voisine qui compense la non-réponse par une hausse de la pondération des répondants difficiles à interviewer. Bartholomew (1961), par exemple, a proposé de ne faire que deux contacts dans une enquête, et de pondérer à la hausse les répondants lors du deuxième appel afin de représenter les non-répondants. L'hypothèse à

la base de cette façon d'agir est que les non-répondants sont comme les répondants tardifs. Cette hypothèse semble douteuse, cependant, et les preuves empiriques provenant d'une étude de suivi intensive des non-répondants dans l'U.S. Current Population Survey ne la confirment pas (Palmer et Jones 1966; Palmer 1967).

3. IMPUTATION

Un vaste éventail de méthodes d'imputation ont été mises au point pour attribuer les valeurs pour les réponses aux questions manquantes. On se contentera ici de présenter une brève vue d'ensemble des méthodes, les différences fondamentales entre elles et quelques-unes des questions soulevées par l'imputation. Kalton et Kasprzyk (1982) donnent un traitement plus poussé.

Les méthodes d'imputation peuvent aller de simples procédures *ad hoc* utilisées pour obtenir des enregistrements complets pour l'introduction des données aux techniques perfectionnées hot-deck et de régression. Voici quelques procédures d'imputation habituelles:

(a) *Imputation déductive.* Il est parfois possible de déduire la réponse manquante à une question avec certitude à partir des réponses aux autres questions. Des vérifications doivent examiner la cohérence entre les réponses à des questions connexes. Lorsque les vérifications limitent une réponse manquante à une valeur possible seulement, on peut utiliser l'imputation déductive. L'imputation déductive est la forme idéale de l'imputation.

(b) *Imputation par la moyenne globale.* Cette méthode attribue la moyenne globale des répondants à toutes les réponses manquantes.

(c) *Imputation par la moyenne de classe.* L'échantillon total est divisé en classes selon les valeurs des variables auxiliaires servant à l'imputation (comparables aux classes de pondération). À l'intérieur de chaque classe d'imputation, la moyenne de classe des répondants est attribuée à toutes les réponses manquantes.

(d) *Imputation globale aléatoire.* On choisi au hasard un répondant dans l'échantillon de répondants total, et la valeur du répondant choisi est attribuée au non-répondant. Cette méthode est la forme la plus simple de l'imputation hot-deck, c'est-à-dire une procédure d'imputation dans laquelle la valeur attribuée à une réponse manquante est prise chez un répondant à l'enquête courante.

(e) *Imputation aléatoire intra-classe.* Dans cette méthode hot-deck, on choisit un répondant au hasard dans une classe d'imputation, et la valeur ainsi obtenue est attribuée au non-répondant.

(f) *Imputation hot-deck itérative.* Ce terme sert ici à décrire la procédure utilisée pour les questions sur la population active de l'U.S. Current Population Survey (Brooks et Bailar 1978). La procédure commence par un ensemble de classes d'imputation. Une valeur unique pour la question faisant l'objet de l'imputation est attribuée à chaque classe (peut-être tirée d'une enquête antérieure). Les enregistrements du fichier de données de l'enquête sont alors examinés à leur tour. Si un enregistrement a une réponse à la question, la réponse obtenue remplace la valeur réservée pour la classe d'imputation à laquelle il appartient. Si l'enregistrement a une réponse manquante, il se voit attribuer la valeur réservée pour sa classe d'imputation.

La méthode hot-deck est semblable à l'imputation aléatoire intra-classe. Si l'ordre des enregistrements dans le fichier de données était aléatoire, les deux méthodes seraient équivalentes, à l'exception du début. L'ordre non aléatoire de la liste favorise en général la méthode hot-deck, puisqu'elle donne un appariement plus étroit des donneurs et des receveurs, à condition que l'ordre du fichier donne une autocorrélation positive. Les avantages sont, cependant, probablement peu appréciables.

La méthode hot-deck itérative présente l'inconvénient de pouvoir facilement trouver des utilisations multiples pour les donneurs, caractéristique qui se traduit par une perte de précision dans les estimations d'enquête. Une telle utilisation multiple des donneurs se produit lorsque, à l'intérieur d'une classe d'imputation, un enregistrement avec une

réponse manquante est suivi d'un ou de plusieurs enregistrements avec des réponses manquantes. Le nombre de classes d'imputation qui peut être utilisé avec cette méthode doit également être limité afin de garantir que des donneurs existent à l'intérieur de chaque classe.

On trouve des discussions utiles de la méthode hot-deck itérative dans Bailar *et al* (1978), Bailar et Bailar (1978, 1983), Ford (1983), Oh et Scheuren (1980), Oh *et al.* (1980), et Sande (1983).

(g) *Imputation hot-deck hiérarchique.* Cette méthode n'a pas les inconvénients mentionnés plus haut de l'imputation hot-deck itérative. Il s'agit d'une forme d'imputation hot-deck mise au point pour les questions de l'enquête supplémentaire sur le revenu de mars de la Current Population Survey. Cette méthode consiste à répartir les répondants et les non-répondants dans un grand nombre de classes d'imputation à partir d'une catégorisation détaillée d'un ensemble important de variables auxiliaires. On procède ensuite à l'appariement des non-répondants avec les répondants sur une base hiérarchique, en ce sens que si un appariement ne peut être fait dans la classe d'imputation initiale, les classes sont fusionnées et on fait l'appariement à un niveau de détail moins élevé. Coder (1978) et Welniak et Coder (1980) donnent d'autres renseignements sur cette procédure.

(h) *Imputation par régression.* Cette méthode utilise des données de répondants pour régresser la variable pour laquelle des imputations sont nécessaires sur un ensemble de variables auxiliaires. L'équation de régression sert ensuite à prédire les valeurs des réponses manquantes. La valeur imputée peut être soit la valeur prédite, soit la valeur prédite plus un résidu. Il y a plusieurs façons d'obtenir ce dernier, comme on le verra plus loin.

(i) *Appariement par fonction de distance.* Cette méthode attribue à un non-répondant la valeur du répondant "le plus proche", "le plus proche" étant défini en termes d'une fonction de distance pour les variables auxiliaires. Diverses formes de fonctions de distance ont été proposées (Sande 1979; Vacek et Ashikago 1980; etc.), et la fonction peut être construite de façon à réduire l'utilisation multiple d'enregistrements donneurs en incorporant une pénalité pour chaque utilisation (Colledge *et al.* 1978).

Bien qu'à première vue ces procédés peuvent sembler assez disparates, ils peuvent être presque tous réunis à l'intérieur d'un cadre unificateur unique. Les méthodes peuvent toutes être décrites, au moins de façon approximative, comme des cas particuliers du modèle général de régression

$$\hat{y}_{mi} = b_{ro} + \sum b_{rj} z_{mij} + \hat{e}_{mi} \tag{2}$$

où $\hat{y}_{mi}$ est la valeur imputée de l'enregistrement i avec une valeur manquante y, z_{mij} sont les valeurs qui reflètent les variables auxiliaires pour l'enregistrement, b_{ro} et b_{rj} sont les coefficients de régression pour la régression de y sur x pour les répondants, et $\hat{e}_{mi}$ est un résidu choisi selon un schéma précisé pour la méthode d'imputation utilisée.

L'équation (2) représente la méthode d'imputation par régression de façon claire. Si les $\hat{e}_{mi}$ sont posés comme étant égaux à zéro, alors la valeur imputée est la valeur prédite à partir de la régression; sinon, on peut ajouter un résidu d'une forme quelconque. L'équation représente également l'imputation moyenne de classe en définissant les z_j comme des variables dichotomiques qui représentent des classes, et en posant $\hat{e}_{mi} = 0$. L'équation de régression se réduit alors à $\hat{y}_{mi} = \bar{y}_{rh}$, la moyenne de classe. L'imputation aléatoire intra-classe se fait en ajoutant un résidu à la moyenne de classe, résidu qui est l'écart par rapport à la moyenne de classe d'un des répondants. Dans ce cas, $\hat{y}_{mi} = \bar{y}_{rh} + e_{rhk}$, où e_{rhk} est l'écart par rapport au répondant k dans la classe h; ceci se ramène à $\hat{y}_{mi} = y_{rhk}$, qui est la valeur pour le répondant en question. Les méthodes hot-deck itérative et hiérarchique ressemblent à la méthode aléatoire intra-classe. La méthode de la moyenne globale et la méthode d'imputation globale aléatoire sont des cas dégénérés de la méthode de la moyenne de classe et de la méthode aléatoire intra-classe qui n'utilisent pas de données auxiliaires.

Une considération importante dans le choix des méthodes d'imputation est le type de variables que l'on impute. Toutes les méthodes ci-dessus peuvent être appliquées de façon courante en présence de variables continues, mais certaines d'entre elles ne conviennent pas

dans le cas de variables catégoriques ou discrètes (telles que être un membre de la population active (1) ou non (0), et le nombre d'années d'études). Les méthodes de la moyenne globale, de la moyenne de classe et d'imputation par régression imputent des valeurs telles que 0.7 pour un actif (c'est-à-dire une probabilité de 70%) et de 10.7 pour le nombre d'années d'études achevées. Ces valeurs ne conviennent pas pour des répondants individuels, et leur arrondissement à des nombres entiers se traduit par un biais. Pour cette raison, ces méthodes d'imputation ne donnent pas de bons résultats pour les variables catégoriques et discrètes. Un avantage appréciable de toutes les méthodes hot-deck est qu'elles donnent toujours des valeurs plausibles, puisque les valeurs proviennent des répondants.

Les méthodes d'imputation ci-dessus ont deux caractéristiques importantes sur lesquelles il faut s'arrêter un peu: lorsque l'on ajoute un résidu, et dans ce cas, sa forme, et si les données auxiliaires sont utilisées sous forme de variables dichotomiques pour représenter des classes, ou si elles sont utilisées directement dans la régression. Ces deux caractéristiques sont examinées dans les deux sous-sections suivantes. Les sous-sections suivantes examinent d'autres problèmes soulevés par l'utilisation de l'imputation.

3.1 Choix des résidus

On peut classer les méthodes d'imputation en déterministes ou stochastiques selon que les $\hat{e}_{mi}$ sont posés comme étant égaux à zéro ou non. Pour chaque méthode d'imputation déterministes, il y a une contrepartie stochastique. Soit $\hat{y}_{mid}$ la valeur imputée par la méthode déterministe, et $\hat{y}_{mis} = \hat{y}_{mid} + \hat{e}_{mi}$ celle imputée par la méthode stochastique correspondante. On obtient alors $E_2(\hat{y}_{mis}) = \hat{y}_{mid}$, où E_2 indique l'espérance mathématique pour l'échantillonnage des résidus compte tenu de l'échantillon initial, à condition que $E_2(\hat{e}_{mi}) = 0$ (ce qui est en général le cas).

Le choix entre une méthode déterministe et une méthode stochastique correspondante dépend de la nature de l'analyse d'enquête effectuée. Examinons d'abord l'estimation de la moyenne de la population de la variable y en utilisant pour cela la moyenne d'échantillon des valeurs des répondants et des valeurs imputées des non-répondants. Comme Kalton et Kasprzyk (1982) le montrent, avec $E_2(\hat{y}_{mis}) = \hat{y}_{mid}$, il s'en suit que l'espérance mathématique de la moyenne de l'échantillon est la même, quelle que soit la méthode d'imputation utilisée (détermiste ou stochastique). Les méthodes ont donc par conséquent le même effet sur le biais de l'estimation. Cependant, l'addition de résidus aléatoires dans la méthode stochastique entraîne une perte de précision dans la moyenne de l'échantillon. Bien qu'il soit possible de maîtriser cette perte en choisissant une méthode appropriée d'échantillonnage des résidus (Kalton et Kish 1984), il n'y en reste pas moins qu'il y a une certaine perte de précision. Pour cette raison, un schéma déterministe est préférable dans le cas de l'estimation de la moyenne de la population.

Considérons maitenant l'estimation de l'écart type des éléments et la distribution de la variable y. Les méthodes d'imputation déterministes ne donnent pas de bons résultats, puisque qu'elles entraînent une réduction de l'écart type et une déformation de la forme de la distribution. On peut illuster simplement ceci en termes de la méthode d'imputation de la moyenne de classe. En attribuant la moyenne de classe à toutes les valeurs manquantes dans une classe, la forme de la distribution est indiscutablement déformée, avec une série de pointes comme moyennes de classe. L'écart type de la distribution se trouve atténué parce que les valeurs imputées ne prennent en compte que la variance inter-classes et non pas la variance intra-classes. L'intérêt des méthodes d'imputation stochastique est que le terme résiduel saisit la variance inter-classes (ou résiduelle), et évite ainsi la réduction de l'écart type de l'élément et la distortion de la distribution.

Comme certaines analyses d'enquête vont probablement faire intervenir les distributions des variables, on préfère en général des méthodes d'imputation stochastiques telles que les méthodes hot-deck. Une fois que l'on a décidé d'utiliser une méthode stochastique, la question du choix des résidus se pose. Si l'on accepte les hypothèses de régression habituelles, il est possible de choisir les résidus à partir d'une distribution normale avec une moyenne de zéro et une variance égale à la variance des résidus provenant de la régression des répondants.

Cependant, cette façon d'agir comporte l'utilisation du modèle uniquement. Une autre solution qui éviterait l'hypothèse de normalité consisterait à chosir les résidus au hasard à partir de la distribution empirique des résidus des répondants. Une autre solution serait de choisir un résidu à partir d'un répondant qui correspond "de près" au non-répondant, avec une mesure "proche" en termes de valeurs semblables des variables auxiliaires. Cette solution intéressante évite l'hypothèse d'homoscédasticité et empêche une mauvaise spécification de distribution du terme résiduel. À la limite, le répondant le plus proche est celui qui a les mêmes valeurs de toutes les variables auxiliaires que le non-répondant. Dans ce cas, le non-répondant reçoit une des valeurs des répondants appariés. C'est ce qui se produit avec les méthodes hot-deck, lorsque les non-répondants et les répondants sont appariés en termes des variables auxiliaires et que les non-répondants se voient attribuer des valeurs provenant des répondants appariés.

Une autre considération dans le choix des résidus est de rendre les valeurs imputées plausibles. Comme on l'a noté plus haut, les méthodes déterministes peuvent imputer des valeurs pour les variables catégoriques et discrètes qui ne sont pas plausibles. Certaines méthodes stochastiques résolvent ce problème, par la répartition des résidus. En particulier, l'utilisation des résidus des répondants avec les méthodes aléatoire intra-classes et les méthodes hot-deck itérative et hiérarchique font en sorte que les valeurs imputées sont plausibles.

3.2 Classe d'imputation ou imputation par régression

Comme on l'a noté plus haut, les deux méthodes de classes d'imputation ou d'imputation par régression relèvent du modèle d'imputation donné par l'équation (2). La différence entre ces deux méthodes réside dans les façons dont elles emploient des variables auxiliaires.

Les méthodes de classes d'imputation divisent l'échantillon en un ensemble de classe. À cette fin, il faut catégoriser les variables auxiliaires continues. La constitution des classes est parfaitement flexible, et l'utilisation symétrique des variables auxiliaires dans différentes parties de l'échantillon n'est pas nécessaire. Ainsi, par exemple, lorsqu'on impute le taux de rémunération horaire dans un échantillon d'employés, on peut d'abord commencer par diviser l'échantillon en deux parties, soit les syndiqués et les non-syndiqués, ensuite on peut constituer les classes d'imputation pour les membres en termes d'âge et de profession, tandis que les classes pour les non-syndiqués peuvent être constituées en termes de sexe et de branche d'activité. En règle générale, on vise à construire des classes de taille appropriée qui expliquent le plus possible la variance de la variable à imputer. Lorsque les classes sont constituées par une classification recoupée intégrale des variables auxiliaires, le modèle qu'on utilise, effectivement, contient tous les effets principaux et toutes les interactions pour la classification recoupée. Les méthodes de classe d'imputation sont limitées par le fait que le nombre de classes constituées doit être calculé de façon à garantir qu'il y a un nombre minimum de répondants dans chaque classe. La méthode hot-deck hiérarchique tente d'accroître la quantité de données auxiliaires utilisées, mais même avec cette méthode, il est souvent impossible de procéder à l'appariement des répondants et des non-répondants à des niveaux de détail plus fins. Combinées à l'utilisation des résidus des répondants aléatoires à l'intérieur d'une classe, les méthodes de classe d'imputation ont la propriété précieuse selon laquelle les valeurs imputées sont des valeurs acceptables, c'est-à-dire que les valeurs imputées sont les valeurs des répondants.

Les méthodes d'imputation par régression présentent un avantage par rapport à celles de classe d'imputation pour ce qui est du nombre et du niveau de détail des variables auxiliaires qu'elles peuvent utiliser. On peut prendre l'âge, par exemple, comme variable continue au lieu d'être catégorisée en quelques classes seulement. Le modèle de régression permet d'inclure davantage d'effets principaux dans le modèle, mais au prix d'une baisse du nombre d'interactions. Naturellement, les modèles de régression peuvent inclure certaines interactions, mais celles-ci doivent être précisées. Les modèles peuvent également inclure des termes polynomiaux et utiliser des transformations, mais là encore, il faut les préciser. Le modèle de régression présente le potentiel de donner de meilleures prédictions pour les valeurs imputées, mais pour cela il faut procéder à une modélisation minutieuse. Une modélisation d'imputation minutieuse n'est pas réaliste pour toutes les variables d'une enquête, mais elle peut être réalisable pour une ou deux variables importantes, et en particulier dans le cas d'enquêtes périodiques. Sans

une modélisation minutieuse, on court le risque sérieux d'avoir de mauvaises imputations, bien que, comme on l'a mentionné plus tôt, ce risque peut être réduit par l'affectation de résidus aléatoires de répondants "proches".

Si l'imputation par régression attribue au résidu d'un répondant exactement les mêmes valeurs des variables auxiliaires, la valeur imputée sera nécessairement une valeur acceptable. Cependant, s'il existe une différence, même minime, entre les valeurs du répondant et du non-répondant pour les variables auxiliaires, la valeur imputée peut ne pas être acceptable. Une variante de l'imputation par régression qui évite ce problème, appelée appariement de moyennes prédictives est décrite par Little (1986b) (Little attribue cette méthode à Rubin). Selon cette méthode, le non-répondant est apparié au répondant qui possède la valeur prédite la plus proche. Alors, au lieu d'ajouter le résidu du répondant à la valeur prédite du non-répondant, on attribue à ce dernier la valeur du répondant. La méthode devient une méthode hot-deck semblable à l'appariement de la fonction de distance.

Le choix entre les méthodes de classe d'imputation et d'imputation par régression devrait dépendre en partie des efforts consacrés à la mise au point des modèles de régression. À moins que des ressources appropriées ne soient consacrées à la mise au point d'un modèle de régression, la méthode de classe d'imputation pourrait être plus sûre. Le choix doit également en partie dépendre de la taille de l'échantillon. Lorsque l'échantillon est important, les méthodes hot-deck vont probablement utiliser suffisamment de classes pour bénéficier de toutes les principales variables indépendantes; cependant, dans le cas des petits échantillons, ceci ne pourrait pas être le cas, et les méthodes de régression pourraient présenter un plus grand potentiel. David *et al.* (1986) décrivent une intéressante étude qui compare les modèles de régression pour l'imputation des salaires et traitements dans la U.S. Current Population Survey aux imputations hot-deck hiérarchiques. En dépit des efforts importants consacrés à la mise au point de modèles de régression, les imputations hot-deck ne devaient pas se révéler inférieures dans cet important échantillon.

3.3 Effet de l'imputation sur les relations

Bien que la plus grande partie de la bibliographie consacrée à l'imputation traite de son effet sur les statistiques univariées telles que les moyennes et les distributions, une importante partie de l'analyse d'enquêtes se rapporte aux relations bivariées et multivariées. Dans ce cas, l'analyse des relations peut être considérée en termes généraux de façon à regrouper les totalisations recoupées, la corrélation ou l'analyse de régression, la comparaison des moyennes ou des proportions intra-classe et toute autre analyse faisant intervenir deux ou plusieurs variables. Comme on le montrera ci-dessous, l'imputation peut avoir des effets nocifs sur toutes les analyses de relations, atténuant souvent des associations entre variables. On trouve dans Santos (1981), Kalton et Kaspryzk (1982) et Little (1986a) des discussions des effets des imputations sur les relations.

La nature générale de l'effet de l'imputation sur les relations est mise en évidence si l'on examine son effet sur l'estimation de la covariance de l'échantillon dans la situation simple où la variable y a des réponses manquantes qui sont manquantes au hasard pour l'ensemble de la population et où la variable x n'a pas de données manquantes. La covariance de l'échantillon s_{xy} est calculée comme d'habitude, à partir des valeurs réelles pour les répondants et des valeurs imputées pour les non-répondants, comme une estimation de la covariance de la population S_{xy}. Si l'on utilise le fait que $E_2(\hat{y}_{mis}) = \hat{y}_{mid}$ comme ci-essus, on peut montrer facilement que la valeur attendue de s_{xy} dans une méthode d'imputation déterministe est la même que celle en vertu de la méthode stochastique correspondante.

Comme le montre Santos (1980), le biais relatif de s_{xy} lorsqu'on utilise la méthode de la moyenne globale ou celle de l'imputation globale aléatoire est approximativement $-\bar{M}$, où $\bar{M}$ est le taux de non-réponse. Cette situation s'explique par le fait que les valeurs y imputées ne sont pas reliées à leurs valeurs x, et que les cas avec des valeurs imputées atténuent la covariance vers zéro. Cette atténuation se trouve réduite en importance par les méthodes d'imputation qui utilisent les variables auxiliaires. Avec l'imputation de la moyenne de classe ou de l'imputation aléatoire intra-classe, le biais relatif est d'environ $-\bar{M}(S_{xy.z}/S_{xy})$,

où $S_{xy.z} = \Sigma W_h S_{xyh}$ est la covariance intra-classe moyenne pour les classes formées par les variables auxiliaires z, S_{xyh} est la covariance à l'intérieur de la classe h, et W_h est la proportion de la population de la classe h. Avec une imputation par régression prédite ou une imputation par régression avec un résidu aléatoire, toutes deux avec une variable auxiliaire unique z, le biais relatif est approximativement $-\bar{M}[1 - (\rho_{xz}\rho_{yz}/\rho_{xy})]$, où ρ_{uv} est la corrélation entre u et v.

La caractéristique inquiétante de ce résultat est que, à moins que $\bar{M}$ ne soit petit, s_{xy} calculé avec des valeurs imputées en vertu de l'une de toutes ces méthodes d'imputation peut avoir un biais appréciable, même dans le cas du modèle avec données manquantes au hasard. Les estimations s_{xy} calculées avec des valeurs imputées obtenues grâce aux méthodes d'imputation de classe et par régression sont sans biais seulement si la covariance partielle $S_{xy.z}$ est zéro. En général, il n'y a pas de raison de supposer les yeux fermés que $S_{xy.z}$ est 0. Toutefois, il y a un important cas lorsque $S_{xy.z} = 0$. Ceci se produit lorsque $x = z$, c'est-à-dire lorsque x est utilisé comme variable auxiliaire dans la procédure d'imputation. Dans ce cas, la covariance de l'échantillon est sans biais selon le modèle de données manquantes au hasard. Ce résultat permet de croire que si la relation entre x et y doit former une partie importante de l'analyse d'enquête, x devrait être utilisé comme variable auxiliaire dans l'imputation des données manquantes y.

La théorie ci-dessus suppose qu'il a des données manquantes seulement pour la variable y. Dans la pratique, la variable x sera également souvent incomplète. Dans ce cas, la covariance d'échantillon peut être réduite en raison des imputations pour les deux variables. Une situation particulière se présente lorsque x et y manquent dans un enregistrement. Si les deux valeurs font l'objet de deux imputations distinctes, la covariance est réduite, mais si elles sont imputées conjointement, en utilisant le même répondant comme le donneur des deux valeurs, la structure de covariance subsiste. Ceci permet de croire que lorsqu'un enregistrement a plusieurs valeurs connexes manquantes, elles doivent être prises du même donneur. Coder (1978) décrit l'utilisation de l'imputation conjointe du même donneur dans l'enquête supplémentaire sur le revenu de mars de la Current Population Survey.

À titre d'illustration de la façon dont les arguments ci-dessus à propos de l'atténuation des covariances s'appliquent à d'autres formes de relation, nous allons donner un exemple numérique simple de l'effet de l'imputation sur la différence entre deux proportions. Supposons que la variable qui nous intéresse soit le fait qu'une personne ait un caractère particulier ou non, et supposons que la moitié des répondants ne répondent pas à la question. Les réponses manquantes sont imputées par une méthode aléatoire d'imputation intra-classe qui utilise deux classes, A et B. L'objectif est maintenant de comparer les pourcentages des hommes et des femmes possédant ce caractère. Les données sont présentées au tableau 1. Comme 60% du total des répondants de la classe A ont le caractère, 60 des 100 hommes et 60 des 100 femmes non-répondants de cette classe seront imputés comme ayant le caractère. De même, dans la classe B, 40% du total des répondants on le caractère, et 40 hommes et 40 femmes non-répondants seront imputés comme l'ayant. La proportion des hommes réels

Tableau 1

Nombre de répondants possédant le caractère et nombre de personnes échantillonnées selon la classe, le sexe et le type de réponse.

	Classe A			Classe B		
	H	F	Total	H	F	Total
Répondants possédant le caractère	80	40	120	60	20	80
Total, répondants	100	100	200	100	100	200
Non-répondants	100	100	200	100	100	200
Total, échantillon	200	200	400	200	200	400

et imputés ayant le caractère est par conséquent $(80+60+60+40)/400 = 0.6$, ou 60%. Pour les femmes, la proportion correspondante est $(40+60+20+40)/400 = 0.4$, ou 40%. La différence entre ces deux pourcentages est de 20%.

Si l'on avait également pris le sexe en compte dans la formation des classes d'imputation, les pourcentages des hommes et des femmes possédant le caractère auraient été de 70% et de 30%, soit une différence de 40%. La non-inclusion du sexe comme variable auxiliaire dans l'imputation s'est traduite ainsi par une réduction appréciable de la mesure de la relation entre le sexe et la présence du caractère.

3.4 Imputations multiples

Idéalement, l'analyste qui utilise un ensemble de données avec des valeurs imputées devrait être en mesure d'obtenir des résultats valables pour toute analyse en utilisant pour cela des techniques normalisées pour les données complètes. Cependant, comme on l'a remarqué dans la section précédente, l'imputation peut déformer les mesures des relations entre variables. Elle peut également déformer l'estimation de l'erreur type.

Toutes les méthodes d'imputation à l'exception de l'imputation par déduction fabriquent dans une certaine mesure les données. L'ampleur de cette fabrication dépend de la qualité de la prédiction des valeurs manquantes par le modèle d'imputation. Si ce dernier n'explique qu'une petite partie de la variance dans la variable entre les répondants, la quantité de fabrication dans chaque valeur imputée sera probablement appréciable. Si le modèle d'imputation explique une proportion élevée de la variance des répondants, cette quantité sera probablement moins importante. Cependant, il ne faut pas perdre de vue que l'ajustement du modèle d'imputation pour les répondants n'est pas nécessairement une bonne mesure de l'ajustement pour les non-répondants.

Les erreurs types calculées de la manière habituelle à partir d'un ensemble de données contenant des valeurs imputées seront généralement sous-estimées en raison du degré de fabrication contenu dans les valeurs imputées. Rubin (1978, 1979) a suggéré l'emploi de la méthode des imputations multiples pour obtenir des inférences valides à partir d'ensembles de données contenant des valeurs imputées (voir également Herzog et Rubin 1983, et Rubin et Schenker 1986). Lorsqu'on utilise les imputations multiples aux fins de l'estimation de l'erreur type, la construction de l'ensemble complet des données par imputation des réponses manquantes s'effectue en plusieurs fois (m par exemple) en utilisant la même procédure de l'imputation. Les estimations de l'échantillon z_i $(i = 1, 2, \ldots, m)$ du paramètre de la population étudiée Z sont calculées à partir de chacun des ensembles de données répétés, et on calcule leur moyenne $\bar{z}$. Un estimateur de la variance pour $\bar{z}$ est ensuite obtenu par la formule $\hat{V} = \hat{W} + [(m+1)/m]\hat{B}$, où $\hat{W}$ est la moyenne de la variance intra-répétition de $\bar{z}$ et $\hat{B} = \Sigma(z_i - \bar{z})^2/(m-1)$ est la variance entre répétitions. Même si l'on inclut cette dernière composante, cependant, les étendues des intervalles de confiance pour Z basées sur $\hat{V}$ sont toujours surestimées, le degré de surestimation augmentant avec le niveau de non-réponse.

Cette surestimation des niveaux de confiance peut être corrigée en modifiant la procédure d'imputation, comme le décrivent Rubin et Schenker (1986). Ils considèrent la méthode d'imputation globale aléatoire, et l'une de leurs modifications prend en compte l'incertitude entourant la moyenne et la variance de la population de la façon suivante. Avec la méthode d'imputation globale aléatoire normale, les moyenne et variance attendues conditionnelles des valeurs imputées sont la moyenne et la variance des répondants de l'échantillon. Avec la modification, la moyenne et la variance attendues des valeurs imputées d'une répétition sont tirées au hasard des distributions appropriées. Les valeurs imputées sont donc une sélection aléatoire de valeurs de répondants, modifiées pour la moyenne et la variance choisies au hasard. Lorsque l'on estime la moyenne de la population, l'effet du changement de la moyenne et de la variance attendues entre répétitions est d'accroître la composante de la variance intra-répétitions dans $\hat{V}$. Ceci se traduit par une meilleure délimitation des intervalles de confiance obtenus.

Un problème important posé par l'utilisation d'imputation multiples est la quantité d'analyses informatiques supplémentaires nécessaires, qui augmente avec le nombre de

répétitions, m. Pour cette raison, on peut préférer une valeur de m peu élevée, comme par exemple $m = 2$. Une telle valeur de m peu élevée peut cependant se traduire par un niveau de précision bas pour l'estimateur de la variance. Même avec un m peu élevé, on peut se demander si la méthode de l'imputation multiple est plausible pour les analyses de routine. Elle serait mieux utilisée dans des études spéciales comme celles décrites par Herzog et Rubin (1983).

En plus de fournir des erreurs types appropriées, un autre avantage des imputations multiples résultant de la même procédure d'imputation est qu'elle réduit la perte de précision dans les estimations d'enquête résultant de la sélection aléatoire de répondants comme donneurs de valeurs imputées (voir section 3.1). Cette perte se trouve réduite dans le cas des imputations multiples par une mise en moyenne des répétitions. Un petit nombre de répétitions convient bien à cette fin. Comme on l'a dit plus tôt, Kalton et Kish (1984) décrivent d'autres moyens de choisir l'échantillon de répondants pour y arriver.

Une autre application potentielle importante des imputations multiples est la production d'imputations pour les différentes répétitions grâce à des procédures d'imputation différentes, retenant des hypothèses différentes à propos des non-répondants. Supposons par exemple que les taux horaires de rémunération doivent être imputés pour certains salariés de l'échantillon. Une procédure qui pourrait être utilisée est la méthode de l'imputation aléatoire intra-classe, qui repose sur l'hypothèse que les non-répondants manquent au hasard à l'intérieur des classes. Si l'on estime que les non-répondants pourraient dans la réalité provenir davantage des salariés touchant des taux plus élevés de rémunération dans chaque classe, une simple modification de la méthode aléatoire intra-classe consisterait à imputer des valeurs qui dépasseraient de 50 cents les valeurs des donneurs, par exemple. D'autres procédures d'imputation, utilisant par exemple des classes d'imputation différentes, pourraient être également essayées. La comparaison des estimations d'enquête obtenues des ensembles de données auxquels on a appliqué les différentes procédures d'imputation fournit alors une indication importante de la sensibilité des estimations aux valeurs imputées. Si les estimations se révèlent être très semblables, on peut les accepter avec plus de confiance; si elles diffèrent sensiblement, il faut les traiter avec beaucoup de prudence.

4. CONCLUSION

On a présenté la pondération et l'imputation comme deux méthodes distinctes pour le traitement des données d'enquête manquantes, mais en fait il existe une relation étroite entre elles. À titre d'illustration, on peut considérer une méthode d'imputation qui attribue les valeurs des répondants aux non-répondants. Dans les analyses univariées, ce procédé équivaut à l'abandon des enregistrements des non-répondants et à l'addition des poids des non-répondants à ceux des répondants donneurs (Kalton 1986).

Les différences entre la pondération et l'imputation ressortent lorsqu'on examine la nature multivariée des données d'enquête. Il est possible d'imputer les réponses d'un non-répondant intégral en prenant toutes les réponses d'un donneur unique; cependant, la pondération est généralement plus simple dans ce cas et elle permet d'éviter la perte de précision résultant de l'échantillonnage de répondants devant servir comme donneurs. Il n'est pas pratique d'utiliser la pondération pour régler la non-réponse à une question, puisque cela se traduirait par des ensembles différents de poids pour chaque question; cette façon d'agir entraînerait de graves difficultés dans le cas des totalisations recoupées et d'autres analyses des relations entre variables.

La pondération est un ajustement global unique qui essaie de compenser les réponses manquantes pour toutes les questions simultanément. L'imputation, par contre, traite chaque question séparément. Cette différence se répercute sur la façon dont les données auxiliaires sont utilisées. Lors de la constitution des classes de pondération, l'objectif et de déterminer les classes dont les taux de réponse diffèrent. Le choix des variables auxiliaires utilisées dans l'imputation, cependant, est déterminé d'abord en fonction de leurs possibilités de prédire les réponses manquantes.

Une hypothèse se trouvant à la base de toutes les procédures examinées ici est celle selon laquelle une fois que les variables auxiliaires ont été prises en compte, les valeurs manquantes le sont au hasard. Ainsi, par exemple, on suppose que les non-répondants sont comme les répondants à l'intérieur des classes de pondération et d'imputation. Cette hypothèse peut être évitée en utilisant des modèles de censure stochastiques, comme l'ont fait Greenlees *et al.*(1982) en imputant les salaires et traitements de la Current Population Survey. Cependant, comme Little (1986b) le fait remarquer, ces modèles sont très sensibles aux hypothèses de distribution retenues.

Une autre méthode pour le traitement des données d'enquête manquantes consisterait à laisser les valeurs manquantes dans l'ensemble de données et laisser l'analyste incorporer des modèles de données manquantes appropriés dans l'analyse (Little 1982). Cette méthode a beaucoup de caractéristiques intéressantes, mais les ressources en personnel et calculs nécessaires à sa mise en oeuvre l'empêchent de l'utiliser comme stratégie générale. Cette approche semble plutôt mieux convenir à un petit éventail d'analyses spéciales. Afin de permettre à l'analyste d'adopter cette approche, il est essentiel que toutes les valeurs imputées soient indiquées afin de signaler lesquelles ne sont pas des réponses réelles, de sorte qu'il soit possible de les laisser de côté dans l'analyse.

Enfin, nous devons noter que toutes les méthodes de traitement des données manquantes d'enquête doivent dépendre d'hypothèses non testables. Si les hypothèses sont gravement erronées, les analystes peuvent donner des conclusions erronées. La seule façon sûre d'éviter d'importants biais de non-réponse dans les estimations d'enquête est de limiter la quantité de données manquantes.

BIBLIOGRAPHIE

BAILAR III, J.C., et BAILAR, B.A. (1978). Comparison of two procedures for imputing missing survey values. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 462-467.

BAILAR, B.A., et BAILAR III, J.C. (1983). Comparison of the biases of the hot-deck imputation procedure with an "equal-weights" imputation procedure. Dans *Incomplete Data in Sample Surveys. Volume 3, Proceedings of the Symposium*. (éd. W.G. Madow et I. Olkin), New York: Academic Press, 299-311.

BAILAR, B.A., BAILEY, L., et CORBY, C.A. (1978). A comparison of some adjustment and weighting procedures for survey data. Dans *Survey Sampling et Measurement*, (éd. N.K. Namboodiri), New York: Academic Press, 175-198.

BARTHOLOMEW, D.J. (1961). A method of allowing for 'not at home' bias in sample surveys. *Applied Statistics*, 10, 52-59.

BISHOP, Y.M.M., FIENBERG, S.E., et HOLLAND, P.W. (1975). *Discrete Multivariate Analyses*. Cambridge, Mass: The MIT Press.

BROOKS, C.A., et BAILAR, B.A. (1978). *An Error Profile: Employment as Measured by the Current Population Survey*. Statistical Policy Working Paper 3. U.S. Department of Commerce. Washington, D.C.: U.S. Government Printing Office.

CHAPMAN, D.W., BAILEY, L., et KASPRZYK, D. (1986). Nonresponse adjustment procedures at the U.S. Census Bureau. *Survey Methodology*, en voie de rédaction.

CODER, J. (1978). Income data collection and processing from the March Income Supplement to the Current Population Survey. *The Survey of Income and Program Participation Proceedings of the Workshop on Data Processing*, February 23-24, 1978, (éd. D. Kasprzyk), Chapter II. Washington, D.C.: U.S. Department of Health, Education and Welfare.

COLLEDGE, M.J., JOHNSON, J.H., PARE, R., et SANDE, I.G. (1978). Large scale imputation of survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 431-436.

COX, B.G., et COHEN, S.B. (1985). *Methodological Issues for Health Care Surveys*. New York: Marcel Dekker.

DAVID, M., LITTLE, R.J.A., SAMUHEL, M.E., et TRIEST, R.K. (1986). Alternative methods for CPS income imputation. *Journal of the American Statistical Association*, 81, 29-41.

DREW, J.H., et FULLER, W.A. (1980). Modelling nonresponse in surveys with callbacks. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 639-642.

DREW, J.H., et FULLER, W.A. (1981). Nonresponse in complex multiphase surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 623-628.

FORD, B.L. (1983). An overview of hot-deck procedures. Dans *Incomplete data in Sample Surveys, Volume 2, Theory and Bibliographies*, (éd. W.G. Madow, I. Olkin et D.B. Rubin), New York: Academic Press, 185-207.

GREENLEES, W.S., REECE, J.S., et ZIESCHANG, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251-261.

HERZOG, T.N., et RUBIN, D.B. (1983). Using multiple imputation to handle nonresponse in sample surveys. Dans *Incomplete data in Sample Surveys, Volume 2, Theory and Bibliographies*, (éd. W.G. Madow, I. Olkin et D.B. Rubin), New York: Academic Press, 209-245.

KALTON, G. (1983). *Compensating for Missing Survey Data*. Ann Arbor: Survey Research Center, University of Michigan.

KALTON, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, 2, en voie de rédaction.

KALTON, G., et KASPRZYK, D. (1982). Imputing for missing survey responses. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 22-31.

KALTON, G., et KISH, L. (1984). Some efficient random imputation methods. *Communications in Statistics - Theory and Methods*, 13(16), 1919-1939.

KISH, L. (1965). *Survey Sampling*. New York: Wiley.

KISH, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society*, Ser. A, 139, 80-95.

LITTLE, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

LITTLE, R.J.A. (1986a). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.

LITTLE, R.J.A. (1986b). Missing data in Census Bureau surveys. *Proceedings of the Second Annual Census Bureau Research Conference*, 442-454.

LITTLE, R.J.A., et DAVID, M.H. (1983). Weighting adjustments for non-response in panel surveys. Document de travail. Washington, D.C.: U.S. Bureau of the Census.

OH, H.L., et SCHEUREN, F. (1978a). Multivariate raking ratio estimation in the 1973 Exact Match Study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 716-722.

OH, H.L., et SCHEUREN, F. (1978b). Some unresolved application issues in raking ratio estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 723-728.

OH, H.L., et SCHEUREN, F. (1980). Estimating the variance impact of missing CPS income data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 408-415.

OH, H.L., et SCHEUREN, F. (1983). Weighting adjustment for unit nonresponse. Dans *Incomplete data in Sample Surveys, Volume 2, Theory and Bibliographies*. (éd. W.G. Madow, I. Olkin et D.B. Rubin), New York: Academic Press, 143-184.

OH, H.L., SCHEUREN, F., et NISSELSON, H. (1980). Differential bias impacts of alternative Census Bureau hot deck procedures for imputing missing CPS income data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 416-420.

PALMER, S. (1967). On the character and influence of nonresponse in the Current Population Survey. *Proceedings of the Social Statistics Section, American Statistical Association*, 73-80.

PALMER, S., et JONES, C. (1966). A look at alternate imputation procedures for CPS noninterviews. Washington, D.C.: U.S. Bureau of the Census document interne.

POLITZ, A., et SIMMONS, W. (1949). I. An attempt to get the 'not at homes' into the sample without callbacks. II. Further theoretical considerations regarding the plan for eliminating callbacks. *Journal of the American Statistical Association*, 44, 9-31.

POLITZ, A., et SIMMONS, W. (1950). Note on an attempt to get the 'not at homes' into the sample without callbacks. *Journal of the American Statistical Association*, 45, 136-137.

RUBIN, D.B. (1978). Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 20-34.

RUBIN, D.B. (1979). Illustrating the use of multiple imputations to handle nonresponse in sample surveys. *Bulletin of the International Statistical Institute*. 48(2), 517-532.

RUBIN, D.B., et SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.

SANDE, G. (1979). Numerical edit and imputation. Article présenté à l'International Association for Statistical Computing, 42nd Session of the International Statistical Institute.

SANDE, I.G. (1983). Hot-deck imputation procedures. Dans *Incomplete Data in Sample Surveys, Volume 3, Proceedings of the Symposium*, (éd. W.G. Madow et I. Olkin), New York: Academic Press, 339-349.

SANTOS, R.L. (1981). Effects of imputation on regression coefficients. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 140-145.

THOMSEN, I. (1973). A note on the efficiency of weighting subclass means to reduce the effects of nonresponse when analyzing survey data. *Statistisk Tidskrift*, 4, 278-283.

THOMSEN, I., et SIRING, E. (1983). On the causes and effects of nonresponse: Norwegian experiences. Dans *Incomplete Data in Sample Surveys, Volume 3, Proceedings of the Symposium*, (éd. W.G. Madow et I. Olkin), New York: Academic Press, 25-29.

VACEK, P.M., et ASHIKAGA, T. (1980). An examination of the nearest neighbor rule for imputing missing values. *Proceedings of the Statistical Computing Section, American Statistical Association*, 326-331.

WELNIAK, E.J., et CODER, J.F. (1980). A measure of the bias in the March CPS earnings imputation system. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 421-425.