Techniques d'enquête, décembre 1980
Vol. 6, n° 2
Statistique Canada, n° 12-001-X au catalogue

Causes des données incomplètes, ajustements et effets

Richard Platek[1]

Résumé

L'article donne un aperçu général des concepts de données incomplètes et de la non-réponse. Il est reconnu que la non-réponse est un indice important de la qualité des données puisqu'elle affecte les estimateurs en y introduisant un biais et une augmentation de la variance à cause d'une réduction de la taille effective de l'échantillon. La relation entre le biais et le taux de non-réponse est moins évidente puisqu'elle dépend de l'ampleur de la non-réponse et aussi de la différence des diverses caractéristiques entre les répondants et les non-répondants.

Le moyen le plus efficace de traiter les effets de la non-réponse est d'en minimiser l'ampleur. Cependant, toute tentative de contrôler l'ampleur de la non-réponse doit être fondée sur une bonne compréhension de ses origines. Les causes de la non-réponse et son ampleur sont fondamentalement liées i) au type d'enquête, ii) aux méthodes de saisie des données et iii) au plan d'échantillonnage. Toutefois, étant donné un plan d'échantillonnage, l'ampleur de la non-réponse sera influencée par des facteurs tels le type de région et le type de non-réponse.

Il y a plusieurs façons de traiter les données incomplètes. Chacune d'elles, en fin de compte, attribue une valeur aux données manquantes ou incorrectes; à moins qu'il ne soit décidé de publier des données « brutes ». La procédure d'attribution de valeurs s'appelle imputation et une telle valeur imputée décrit, présumément, la caractéristique du non-répondant.

L'article donne une brève discussion philosophique sur le sujet de la validation et de l'imputation et leurs applications à la méthodologie des diverses procédures d'imputation. Parmi celles-ci, mentionnons la pondération, réplication, « hot deck » substitution par des données antérieures et remplacement par la valeur zéro. L'application de l'imputation par rapport aux méthodes employées par l'enquête sur la population active au Canada y est aussi discutée. Une table de décision est fournie indiquant les diverses étapes à suivre pour un cas particulier d'un questionnaire de l'EPA partiellement complet.

Mots-clés : Données incomplètes; non-réponse; validation; imputation.

1 Introduction

Au cours des dernières années, le nombre d'enquêtes réalisées en vue de fournir des estimations sur divers sujets a augmenté considérablement dans la plupart des pays, y compris le Canada. La fiabilité des estimations fondées sur les enquêtes dépend de nombreux facteurs, dont l'effet de la non-réponse et des données incohérentes ou incomplètes. Toute enquête, indépendamment de son type et de la méthode de collecte des données, souffrira de non-réponse parce que a) les unités de la

1. Richard Platek, directeur, Division du recensement et des méthodes d'enquête auprès des ménages, Statistique Canada.

population n'ont pas toutes été incluses dans la base de sondage, b) les unités sélectionnées et classées comme étant admissibles n'ont pas pu être trouvées ou c) ces unités ont refusé de participer à l'enquête. Outre la non-réponse, certains enregistrements ne sont que partiellement remplis ou contiennent des réponses invalides.

Une question fréquemment soulevée est celle de savoir si, en ignorant la non-réponse et les données incomplètes, les estimations basées sur l'information fournie par les unités répondantes satisferaient les besoins pour lesquels une enquête est conçue. Par exemple, si l'on estimait un item considéré comme représentant, disons, 15 % de la population, quel serait l'effet possible sur l'estimation si le taux de non-réponse était de 15 % ou de 20 % ? Dans quelle mesure le biais potentiel dû à 15 % ou 20 % de non-réponse surpasserait-il l'erreur due à l'échantillonnage ?

La plupart des statisticiens praticiens ou des analystes des données reconnaissent que les mesures de la non-réponse et de l'incomplétude des données sont un indice important de la qualité de ces dernières, car la non-réponse et l'incomplétude des données ont une incidence sur les estimations en introduisant éventuellement un biais ainsi qu'un accroissement de la variance d'échantillonnage en raison d'une réduction de la taille effective de l'échantillon. La relation entre le biais et l'ampleur de la non-réponse est moins évidente, puisqu'elle dépend tant de l'ampleur de la non-réponse que de la différence entre les caractéristiques des répondants et des non-répondants.

Dans le cas de nombreuses enquêtes à grande échelle, on estime généralement que, si elles étaient mesurées, les erreurs dues à la non-réponse et aux données incomplètes dépasseraient grandement celles dues à l'échantillonnage, du moins pour des désagrégations dont le niveau de détail n'est pas trop fin. Toutefois, dans la plupart des enquêtes, seule l'erreur d'échantillonnage est identifiée et les autres composantes sont consignées et analysées inadéquatement. Pourtant, il s'agit de sources possibles de biais et ne pas tenir compte de leur effet sur les estimations pourrait mener à des résultats d'enquête d'une qualité inacceptable. Par conséquent, la réduction des effets de la non-réponse et des réponses incomplètes ou invalides est très importante et devrait être entreprise à diverses étapes, dont la conception de l'enquête, la collecte des données, le traitement des données et l'estimation.

Un moyen de traiter la non-réponse, après la collecte des données, consiste à appliquer des méthodes d’imputation ou d’ajustement des pondérations à l’étape du traitement et de l’estimation. Tandis que les ajustements pour tenir compte de la non-réponse peuvent réduire le biais plus ou moins efficacement, des opérations de collecte des données bien conçues maintiendront la non-réponse à un niveau acceptable et à un coût raisonnable, réduisant ainsi la nécessité d’appliquer des ajustements.

Dans le présent article, la question de la non-réponse et des données incomplètes, y compris les réponses invalides, dans les enquêtes auprès des ménages sera abordée sous l’angle de leur origine, des diverses méthodes permettant de les réduire, ainsi que de l’ajustement pour en tenir compte dans les estimations finales.

2 Non-réponse

La non-réponse peut être définie comme ne pas réussir à obtenir une réponse utilisable auprès d’une unité déclarante qui est incluse légitimement dans l’échantillon dans une enquête particulière, et elle peut être de deux types :

a) non-réponse totale, quand un questionnaire d’enquête n’est pas rempli pour une unité désignée;

b) non-réponse partielle, quand un questionnaire d’enquête est rempli pour une unité, mais que les réponses à une ou à plusieurs questions ne sont pas obtenues.

La non-réponse se produit à cause de difficultés opérationnelles, de contraintes de temps et de coût, d’un manque de coopération de la part du répondant, de l’incapacité ou de la réticence de l’intervieweur à rechercher les répondants manquants, ou pour certaines autres raisons. La gravité des problèmes de non-réponse est mesurée par les taux de non-réponse qui sont calculés comme étant le pourcentage de ménages non répondants parmi l’ensemble des ménages échantillonnés.

Les taux de non-réponse varient considérablement d’une enquête à l’autre. Dans certaines enquêtes, ils atteignent 40 % et plus, tandis que dans d’autres ils peuvent n’être que de l’ordre de 4 %. La question de savoir si les taux de non-réponse sont trop élevés ou trop bas dépend de l’objet de l’enquête. Si l’objectif est d’estimer un item qui représente 15 % de la population, selon la fiabilité avec laquelle il faut estimer les 15 %,

même un taux de non-réponse de 5 % peut avoir une grande incidence sur l'estimation, particulièrement si la caractéristique des non-répondants est corrélée avec les variables importantes. Par ailleurs, cela ne signifie pas nécessairement qu'une enquête dont le taux de non-réponse est élevé ne peut pas fournir des renseignements utiles. Par exemple, supposons que l'objectif d'une enquête soit de déterminer si 15 % d'une population achèteraient un produit particulier. Si 20 % des enquêtés répondaient qu'ils achèteraient le produit et 30 %, qui ne l'achèteraient pas, et que 50 % étaient des non-répondants, l'objectif de l'enquête serait atteint si tous les non-répondants rentraient dans la catégorie des personnes qui n'achèteraient pas le produit. En général, cependant, le biais possible dans l'estimation sera d'autant plus grand et les objectifs d'une enquête seront d'autant moins susceptibles d'être satisfaits que le taux de non-réponse sera élevé.

La question de l'ampleur de la non-réponse ne peut pas être résolue simplement en débutant avec un échantillon excédentaire pour tenir compte de la non-réponse éventuelle, puisqu'en présence de non-réponse, l'échantillon n'est plus probabiliste. La difficulté tient au fait que les non-répondants diffèrent de certaines façons et à des degrés divers des personnes qui répondent. Nous pouvons supposer que chaque personne (si sélectionnée) est un répondant potentiel, c'est-à-dire que l'individu i, quand il est dans l'échantillon, répondra avec la probabilité δ_i et sera un non-répondant avec la probabilité $1 - \delta_i$. Si la probabilité de réponse δ_i est la même pour tout i, la situation est corrigée facilement en ajustant (gonflant) les poids des répondants. Mais la probabilité de réponse peut dépendre de la caractéristique d'intérêt et l'ajustement des poids des répondants pour tenir compte de la non-réponse donnera lieu à un biais de non-réponse. La grandeur du biais de non-réponse dépendra de la relation entre la caractéristique d'intérêt et la probabilité de réponse. Le problème peut être illustré au moyen des exemples qui suivent sous différents plans d'échantillonnage :

a) Échantillonnage aléatoire simple à un degré : Dans une population de N individus, n sont sélectionnés par EASSR et le sujet de la recherche est une variable quantitative Y (par exemple, revenu, nombre d'accidents, etc.). Supposons que la probabilité d'obtenir une réponse diminue quand la valeur de y augmente. En l'absence de non-réponse, le total $Y = \sum_{t=1}^{N} y_t$ serait estimé sans biais par

$\hat{Y} = \frac{N}{n}\sum_{i=1}^{n} y_i$, mais Y est observée pour $n^*(<n)$ et $\tilde{Y} = \frac{N}{n^*}\sum_{i=1}^{n^*} y_i$ est utilisé pour estimer le total Y. Donc $\tilde{Y}$ aurait été sans biais si la probabilité de réponse ne dépendait pas des valeurs de y, mais dans cette situation, puisque les probabilités de réponse sont faibles pour les grandes valeurs de y, Y sous-estime le total Y; c'est-à-dire que nous avons une corrélation négative entre la probabilité de réponse et la caractéristique, ce qui donne lieu à un biais de non-réponse négatif.

b) Échantillonnage avec probabilité proportionnelle à la taille (PPT) à un degré : Dans une population de N unités (entreprises), n sont sélectionnées avec PPT pour estimer $Y = \sum_{t=1}^{N} y_t$, où y est un attribut quantitatif (par exemple, production, etc.) et p_t est la mesure de taille pour l'unité $t, t = 1, 2 \ldots, N$; $\sum_{t=1}^{N} p_t = 1$, où $\pi_i = np_i$ = probabilité que l'unité i soit dans l'échantillon.

y_i = valeur observée pour l'unité i si elle est sélectionnée et qu'elle répond. En l'absence de non-réponse, $\hat{Y} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$ est une estimation sans biais de Y. Supposons que les valeurs y sont observées pour $n^*(<n)$ unités sur n sélectionnées, alors $\tilde{Y} = \left(\frac{n}{n^*}\right)\sum_{i=1}^{n^*} \frac{y_i}{\pi_i}$ est utilisé pour estimer Y, et sera sans biais si les probabilités de réponse sont corrélées avec le ratio Y_i/P_i. Le biais sera vraisemblablement petit comparativement au total Y si la variabilité des ratios Y_i/P_i est faible. Habituellement, un plan d'échantillonnage PPT est employé lorsqu'il existe une forte corrélation positive entre les valeurs Y_i et les tailles P_i des unités. Souvent, cette forte corrélation résulte en une faible variation des ratios Y_i/P_i quand la courbe de régression de Y_i sur P_i passe par l'origine.

c) Échantillonnage PPT stratifié à plusieurs degrés : Dans les enquêtes auprès des ménages, les ménages sont sélectionnés selon un plan d'échantillonnage PPT stratifié à plusieurs degrés et le plan est habituellement autopondéré. Dans certaines enquêtes, on peut obtenir auprès d'un membre du ménage la réponse au sujet de tous les membres admissibles dans le ménage. Il est concevable que la non-réponse ait tendance à être plus élevée dans un ménage d'une seule personne que dans les ménages comptant deux membres ou plus. Dans de telles conditions, si l'ajustement du poids pour tenir compte de la non-réponse ne tient pas compte de la taille

du ménage, l'estimation résultante de la population présentera un biais à la hausse. Dans ce contexte, considérons un autre exemple. Supposons que le sujet de la recherche est une variable qualitative, c'est-à-dire la présence ou l'absence d'une caractéristique particulière. Supposons en outre que la probabilité de réponse est élevée si la caractéristique est absente et faible si elle est présente. Si une estimation du nombre de personnes ayant la caractéristique est obtenue par ajustement des pondérations des répondants, le résultat sera une sous-estimation, c'est-à-dire que nous aurons une corrélation négative entre les probabilités de réponse et la caractéristique.

Idéalement, pour éliminer le biais de non-réponse, on pondérerait simplement chaque réponse dans l'échantillon par l'inverse du produit des probabilités de sélection et de réponse réelles pour chaque unité répondante. Il s'agit toutefois d'une tâche impossible, puisque la vraie probabilité de réponse est inconnue pour chaque unité. En pratique, nous employons les probabilités de réponse moyennes estimées par les taux de réponse pour ajuster les pondérations dans les estimations. L'effet de cette procédure sur la réduction du biais de non-réponse dépendra du degré de relation entre les probabilités de réponse et les taux de réponse.

Le moyen le plus favorable de traiter l'effet de la non-réponse consiste à minimiser la taille de cette dernière. Cependant, tout effort systématique en vue de contrôler l'ampleur de la non-réponse doit être fondé sur une bonne compréhension de ses origines. Fondamentalement, les causes et l'ampleur de la non-réponse sont liées i) au type d'enquête, ii) aux méthodes de collecte des données, et iii) au plan d'échantillonnage. Mais, même pour un plan d'enquête donné, l'ampleur de la non-réponse sera influencée par des facteurs tels que le type de région, le type de non-réponse, etc. En guise d'illustration, nous procéderons à un bref examen de la non-réponse pour l'Enquête sur la population active au Canada.

L'Enquête sur la population active au Canada est réalisée auprès d'un échantillon probabiliste mensuel de logements. Les ménages dans les logements sélectionnés sont interviewés une fois par mois pendant six mois consécutifs. Chaque mois, durant une semaine particulière (appelée semaine d'enquête), 56 000 logements à travers le Canada sont contactés

par environ 1 100 intervieweurs. Ces derniers recueillent l'information sur les caractéristiques démographiques et l'activité sur le marché du travail des membres de la population civile de 15 ans et plus, sauf les personnes vivant en établissement institutionnel, qui sont membres des ménages résidant dans ces logements.

Dans le cas de l'Enquête sur la population active au Canada, des renseignements détaillés sont consignés sur la non-réponse totale, qui peut être ventilée en plusieurs composantes ayant chacune une cause différente et nécessitant un traitement différent. Par exemple, dans l'EPA ou toute autre enquête auprès des ménages, on distingue les composantes suivantes : a) absence temporaire du ménage, b) personne à la maison, c) refus, d) pas d'intervieweur disponible, e) mauvais temps, etc. L'ampleur de la non-réponse due aux deux dernières causes est de peu d'importance générale.

Le tableau 1 donne un exemple de taux de non-réponse à l'EPA, et est suivi d'une brève discussion des diverses composantes des taux de non-réponse totale.

Tableau 1
Taux de non-réponse (%) en fonction du mode d'occupation des ménages dans l'EPA (juillet 1977 à juin 1978)

Nombre de mois dans l'enquête	Taux de non-réponse (%)			
	Non-réponse totale	Refus	Personne à la maison	Absence temporaire
1	8,04	1,43	2,96	2,94
2	5,09	1,21	1,44	1,99
3	4,71	1,32	1,10	1,90
4	4,65	1,46	1,09	1,79
5	4,62	1,51	0,99	1,77
6	4,45	1,52	0,78	1,73

D'après les résultats présentés au tableau 1, les commentaires qui suivent peuvent être formulés :

a) Le taux de non-réponse totale le plus élevé a eu lieu durant le premier mois, probablement parce que les intervieweurs avaient plus de difficulté à trouver les gens à domicile parce qu'ils n'avaient pas encore déterminé le meilleur moment pour les appeler, comme en témoigne le taux plus élevé de cas « Personne à la maison », par exemple. Ensuite, le taux a diminué fortement le deuxième mois et a continué de baisser au cours des troisième et quatrième mois.

b) Le taux de refus a diminué au deuxième mois, augmenté progressivement au cours du troisième, quatrième et cinquième mois, et s'est stabilisé au sixième mois.

c) Le taux de cas « Personne à la maison » a diminué, d'environ 50 %, du premier mois au deuxième. Il a continué de baisser du deuxième au troisième mois, mais a diminué très progressivement au cours des quatrième et cinquième mois. Une baisse importante a ensuite eu lieu au sixième mois. Le comportement du taux de cas « Personne à la maison » au cours des six mois de participation des ménages à l'enquête était le plus vraisemblablement dû au fait qu'à mesure que la participation d'un ménage à l'enquête s'allonge, l'intervieweur sait de mieux en mieux quand le répondant sera le plus probablement à la maison.

d) Le taux d'« Absence temporaire » a diminué tout au long des six mois, particulièrement du premier au deuxième. Il est difficile d'expliquer ce phénomène, puisque le taux d'« Absence temporaire » ne devrait pas, en principe, dépendre de la durée de participation du ménage à l'enquête. On peut supposer que les intervieweurs pourraient avoir confondu les types de non-réponse « Personne à la maison » et « « Absence temporaire ».

Tableau 2
Taux de non-réponse (%) selon le type de région (moyenne mensuelle, 1978)

Type de région	Proportion approximative de l'échantillon	Non-réponse totale	Refus	Personne à la maison	Absence temporaire
UNAR[1]	0,48	5,0	1,2	1,3	2,0
-urbaine[2]	0,18	5,3	1,0	1,4	2,4
-rurale[2]	0,30	4,9	1,3	1,3	1,8
UAR[3]	0,51	5,7	1,7	1,6	2,0
-zones bâties[4]	0,37	5,5	1,6	1,5	2,0
-banlieues[4]	0,10	4,8	1,6	1,2	1,8
-appartement[5]	0,04	9,8	2,7	3,6	3,0

[1] Les unités non-autoreprésentatives sont les régions en dehors des UAR qui englobent les régions rurales et les petits centres urbains.

[2] Chaque unité primaire d'échantillonnage dans une UNAR est divisée en une partie urbaine et une partie rurale.

[3] Les unités non-autoreprésentatives sont les villes dont la population dépasse 15 000 habitants ou dont les caractéristiques uniques nécessitent qu'elles soient considérées comme des UAR. Chaque UAR est sélectionnée avec certitude.

[4] Les UAR sont stratifiées en sous-unités, et les sous-unités sont classées comme « zones bâties » ou « banlieues » en fonction de leur possibilité de croissance future. Généralement parlant, les ménages des banlieues des UAR appartiennent aux banlieues ou aux régions suburbaines.

[5] Pour 17 grandes villes du Canada, il existe une base de sondage distincte des appartements dans les immeubles comptant au moins cinq étages et au moins 30 unités.

Le taux de non-réponse totale était plus élevé dans les zones bâties que dans les banlieues des UAR, en raison de taux « Personne à la maison » et « Absence temporaire » plus élevés. Donc, il semble que les personnes qui vivent dans les centres des villes ont tendance à être plus difficiles à joindre que celles qui vivent dans les banlieues; néanmoins, les différences n'étaient pas importantes.

Le taux de non-réponse totale était plus élevé pour les appartements situés dans les UAR que pour n'importe quelle autre zone présentée dans le tableau 2. En fait, le taux de non-réponse totale dans l'échantillon d'appartements des UAR était presque égal à deux fois le taux dans échantillon hors appartement des UAR (constitué des zones bâties et des banlieues). Les composantes « Refus », « Personne à la maison » et « Absence temporaire » étaient aussi plus importantes pour les appartements. Pour l'instant, cette catégorie est petite, mais elle grandit assez rapidement. Le taux de cas « Personne à la maison » était presque trois fois plus élevé dans l'échantillon d'appartements que dans l'échantillon d'autres logements que les appartements. Cette différence importante pourrait être due à des modes de vie différents des personnes vivant en appartement et de celles ne vivant pas dans ce genre de logement. Les ménages qui vivent en appartement sont habituellement constitués de personnes seules ou de petites familles qui ont tendance à être plus mobiles et difficiles à trouver à la maison, tandis que les ménages qui vivent dans d'autres logements que les appartements sont plus susceptibles de comprendre des familles plus grandes avec enfants.

Dans les UNAR, le taux de non-réponse totale était plus élevé dans la partie urbaine, à cause des taux plus élevés d'« Absence temporaire »[1] parmi les ménages urbains des UNAR. Les taux de cas « Personne à la maison »[2] étaient à peu près les mêmes dans les parties urbaine et rurale, mais le taux de refus de 1,3 % dans les zones rurales des UNAR était en moyenne 30 % plus élevé que le taux de 1,0 % observé dans les zones urbaines des UNAR.

À la présente section, j'ai traité principalement de la non-réponse totale ou non-réponse d'une unité. Les problèmes associés à la non-réponse partielle, définie au début de la section, sont examinés

[1] Le ménage était absent pendant toute la semaine de l'enquête.
[2] Aucune prise de contact n'a pu avoir lieu avec l'occupant après plusieurs tentatives.

brièvement à la section sur la collecte des données et à la section suivante. Les réponses incomplètes ou invalides sont examinées à la section sur l'imputation.

3 Traitement de la non-réponse

Le plan d'enquête comprend fondamentalement trois étapes, à savoir a) le plan de sondage, b) la collecte des données et c) l'estimation.

Aucune de ces étapes ne peut être entreprise purement sur des bases techniques ni purement sur des bases pratiques. Le plan d'enquête est établi à la lumière de ce qui est faisable en pratique et souhaitable en théorie, afin de répondre aux besoins des utilisateurs. L'importance de la non-réponse, par la voie de ses effets sur les estimations de l'enquête, fait partie intégrante du plan d'enquête et peut difficilement être laissée au hasard. Des dispositions doivent être prises à toutes les étapes possibles d'une enquête afin de contrôler l'ampleur de la non-réponse et de réduire au minimum l'effet de la non-réponse finale.

Même si la non-réponse se produit effectivement durant l'étape de collecte des données, il est possible d'influer considérablement sur son ampleur à l'étape de la planification en examinant l'effet possible que divers facteurs du plan d'enquête peuvent avoir sur la non-réponse.

En outre, une préparation minutieuse et appropriée à la collecte des données axée sur les méthodes d'interview et de motivation des répondants et des intervieweurs aura un effet important sur l'ampleur de la non-réponse.

À l'étape du traitement et de l'estimation, un effort est fait en vue de réduire au minimum l'effet de la non-réponse sur les estimations finales grâce à l'imputation des valeurs manquantes.

Considérons maintenant comment l'effet de la non-réponse peut être influencé à chacune de ces étapes.

Plan de sondage (planification et élaboration)

À l'étape de la planification, être conscient de l'effet de la non-réponse sur l'erreur quadratique moyenne (EQM) des données d'enquête peut mener à un plan d'enquête qui aura une incidence sur l'ampleur de la non-réponse. L'un des facteurs importants de la planification d'une

enquête est le choix d’un niveau de tolérance de la non-réponse, et un concepteur d’enquête chevronné peut estimer assez exactement, pour une enquête particulière, le niveau de réponse auquel on peut s’attendre sous diverses conditions. Par exemple, pour des estimations nationales sur un échantillon de grande taille, l’effet de la non-réponse sur la variance d’échantillonnage et de réponse sera vraisemblablement peu important, et le biais sera probablement la composante prédominante de l’EQM. Toutefois, pour des estimations infranationales, les variances seront vraisemblablement grandes, de sorte que le biais pourrait être relativement moins important.

Le coût d’enquête est un autre élément important qui influera sur de nombreux facteurs de l’élaboration de l’enquête, y compris la non-réponse. Il importe de trouver un équilibre entre les autres facteurs et le coût, afin de réaliser un taux de non-réponse suffisamment faible pour desservir les objectifs de l’enquête. Il faut aussi se rendre compte que, dans des limites raisonnables, il est parfois préférable d’accepter un échantillon un peu plus petit qu’il n’était prévu au départ et de transférer les ressources à des procédures appropriées de collecte des données, de suivi et d’estimation. Cela serait particulièrement avantageux si le concepteur de l’enquête soupçonnait de grandes différences entre les caractéristiques des répondants et des non-répondants.

À l’étape de la planification et de l’élaboration de l’enquête, un certain nombre de facteurs doivent être pris en considération pour arriver au plan final. Ces facteurs peuvent être classés en trois groupes :

Groupe I
a) taille de l’échantillon
b) stratification
c) degré de groupement
d) répartition de l’échantillon
e) méthode de sélection

Groupe II
a) base de sondage
b) méthode d’interview
c) sélection, formation et supervision du personnel
d) longueur et énoncé du questionnaire

e) caractère sensible des questions

f) type de région dans laquelle l'enquête est réalisée

g) faisabilité et coût des rappels

h) publicité

Groupe III a) vérification et imputation

b) estimation

c) estimation de la variance

Toutes ces opérations influent sur l'erreur quadratique moyenne qui donne une mesure de la fiabilité des données.

Supposons que l'erreur quadratique moyenne peut être décomposée en les composantes suivantes :

$$EQM = V_S + V_R + V_{CR} + (B_S + B_R)^2$$

où

V_S = variance d'échantillonnage

V_R = variance de réponse simple

V_{CR} = variance de réponse corrélée

B_S = biais d'échantillonnage

B_R = biais de réponse (y compris biais de non-réponse)

La variance d'échantillonnage (V_S) et le biais d'échantillonnage (B_S) sont affectés par tous les facteurs du groupe I et du groupe III, ainsi que par l'ampleur de la non-réponse. Il importe de souligner que chaque programme d'enquête établit ses propres exigences en ce qui concerne le plan, le questionnaire et les méthodes d'interview. Par exemple, un plan de sondage sans mise en grappes peut produire un taux de non-réponse plus élevé qu'un plan de sondage en grappes. Cela pourrait tenir à la nécessité d'effectuer de nombreux déplacements dans le cas d'interviews sur place où, pour des raisons de coût, les visites de rappel répétées doivent être limitées. Plus la non-réponse est grande, plus son effet sur la variance d'échantillonnage et le biais de non-réponse sera grand. Les composantes non dues à l'échantillonnage de l'erreur quadratique

moyenne (V_R,V_{CR},B_R) sont influencées par tous les facteurs du groupe II. En outre, il est assez évident que tous les facteurs du groupe II sont une source et une cause potentielles de non-réponse. Ainsi, nous avons vu que les taux de non-réponse dépendent du type de région dans laquelle est réalisée l'enquête. La longueur du questionnaire et son caractère sensible auront sans aucun doute un effet sur l'importance de la non-réponse. Donc, l'examen minutieux des facteurs du groupe II à l'étape de l'élaboration du plan permet d'éviter des problèmes sérieux liés à la non-réponse.

Collecte des données

Dans la discussion des approches en vue de réduire au minimum la non-réponse, on peut faire la distinction entre deux catégories. L'une, telle que « Personne à la maison » ou « Absence temporaire », est en fait un problème de type « Pas de contact » et est principalement axée sur les opérations. L'autre catégorie est le problème de non-réponse réelle, où un contact a été établi avec le répondant, mais une réponse acceptable n'a pas été obtenue.

Habituellement, les solutions au problème de type « Pas de contact » sont évidemment de nature opérationnelle. Dans le cas d'une interview par téléphone ou sur place, le moment et les profils des appels téléphoniques ou des visites au répondant sont importants. La taille de la tâche et le temps accordé pour la collecte des données doivent être adéquats. Dans le cas d'une enquête par la poste, il est essentiel de s'assurer que les adresses sur la liste d'envoi sont correctes, que les procédures de suivi sont efficaces et que le matériel est approprié. L'ampleur de la non-réponse due aux cas « Personne à la maison » ou « Absence temporaire » donne une idée importante des problèmes opérationnels.

L'existence de refus pose un ensemble différent de problèmes. Il faut reconnaître d'emblée que les taux de refus ne sont pas toujours aussi simples à expliquer qu'on pourrait l'imaginer. Un intervieweur pourrait préférer consigner un refus dans la catégorie « Personne à la maison » ou un répondant pourrait simplement ne pas ouvrir la porte comme moyen de refus et être donc enregistré comme un cas « Personne à la maison ». Dans une enquête par la poste, il n'est pas toujours certain que le répondant a reçu le questionnaire et, quand il l'a reçu, s'il a ou non

simplement négligé de le renvoyer par la poste. Donc, la distinction entre la non-réponse réelle et d'autres causes n'est pas faite facilement. La réponse invalide présente encore un autre ensemble de problèmes, puisqu'un intervieweur inexpérimenté pourrait ne pas se rendre compte que les données sont invalides ou illogiques jusqu'à ce qu'une routine de vérification en fasse la découverte. En outre, l'intervieweur peut maladroitement coder la réponse dans un emplacement incorrect sur le questionnaire, produisant ainsi une donnée invalide qui doit être écartée. Quoi qu'il en soit, indépendamment de la façon dont la non-réponse est consignée, le problème semble consister à motiver le répondant approprié à produire une réponse valide.

En ce qui concerne la motivation, considérons que le répondant est neutre à l'égard de l'enquête et examinons les éléments qui pourraient le motiver à répondre ou à non répondre. Des facteurs tels que la difficulté à comprendre les questions, l'utilisation du temps de réponse, la protection de la vie privée, l'indifférence, des difficultés à se rappeler l'information, des questions embarrassantes ou personnelles sont tous des exemples de motivation à ne pas répondre. Par ailleurs, l'intérêt pour l'enquête, la volonté d'aider, le devoir, la compréhension de l'importance des résultats de l'enquête sont des exemples de motivation à répondre.

Le problème se résume donc à trouver les moyens d'accentuer la motivation positive et de réduire la motivation négative jusqu'à ce que la balance penche en faveur de la réponse. L'élément essentiel est le répondant et tout facteur qui affecte sa capacité et sa motivation à répondre doit être pris en considération par un concepteur d'enquête.

Les lettres de présentation de l'enquête, les exemples d'utilisations des données, et les brochures décrivant les objectifs de l'enquête et les lois autorisant cette dernière sont souvent d'excellents moyens d'éviter l'hostilité et la méfiance des répondants.

L'atteinte à la vie privée est associée au contenu du questionnaire, bien que la réaction des répondants soit relativement variable. De nombreuses procédures existent pour minimiser l'effet sur le répondant, et les procédures choisies doivent être adaptées à la situation particulière. Dans certains cas, il peut être préférable de permettre à l'enquêté de répondre de manière entièrement anonyme. Cela peut s'accomplir en faisant

appel à l'autodéclaration sans aucune identification sur le questionnaire. Néanmoins, assez souvent, il est essentiel d'avoir un code postal ou une désignation d'échantillon pour les besoins de la pondération et de l'estimation du suivi et, dans ce cas, il faut veiller à ce que le répondant ne perçoive pas cela comme un moyen d'identifier ses réponses.

En plus de garantir la protection de la vie privée, certaines formes de rémunération du répondant en dédommagement de son temps et de son effort ont été appliquées par certains organismes d'enquête.

a) Substitution sur le terrain - Une méthode en vue de traiter la non-réponse à l'étape de la collecte des données consiste à substituer sur le terrain d'autres unités non sélectionnées antérieurement. Il convient de souligner qu'il s'agit néanmoins encore d'une non-réponse et que la substitution est un mode d'imputation. Deux catégories fondamentales de substitution sont utilisées :

a. sélection d'une unité de remplacement aléatoire;

b. sélection d'une unité de remplacement désignée spécialement.

Dans le cas d'une méthode de substitution aléatoire, une unité supplémentaire de la population est sélectionnée selon une méthode probabiliste pour remplacer un non-répondant. Pour de nombreuses procédures de substitution aléatoire, les unités de remplacement possibles sont sélectionnées avant la collecte des données afin d'éviter tout délai ou problème qui pourrait exister si les unités de remplacement étaient sélectionnées durant ou après la collecte des données.

Dans une procédure qui utilise des unités de remplacement désignées spécialement (par exemple, un voisin immédiat), l'intention est de trouver une unité de remplacement dont les caractéristiques sont similaires à celles du non-répondant. Malheureusement, cela peut entraîner un biais d'échantillonnage, surtout si le voisin vit en dehors de la base de sondage. Alors que toute unité originale peut être sélectionnée avec une probabilité connue conformément au plan de sondage, la substitution d'autres répondants non sélectionnés antérieurement pour remplacer les répondants non coopératifs, d'une façon non contrôlée ou même d'une façon contrôlée, modifiera les probabilités d'inclusion. Un biais d'échantillonnage de grandeur inconnue pourrait être introduit

(puisque les probabilités de sélection sont inconnues). Bien que la variance d'échantillonnage puisse être réduite en raison d'un accroissement de la taille effective de l'échantillon, il n'y aurait probablement pas de réduction du biais de réponse ni du biais de non-réponse. Même si les probabilités d'inclusion pouvaient être calculées, un certain biais de non-réponse persisterait, puisque les unités non coopératives n'ont essentiellement aucune chance d'être incluses. La question fondamentale concernant le mérite des procédures de substitution est celle de savoir si le recours à la substitution fournit de meilleures valeurs substitutives pour les non-répondants que celles fournies par d'autres procédures d'imputation. Sans aucun doute, le recours à des procédures de substitution présente des avantages et des inconvénients. Le premier avantage est qu'il s'agit d'un moyen commode d'équilibrer l'échantillon en ce qui concerne la taille de ce dernier. L'autre est la réduction de la variance d'échantillonnage due à l'augmentation de la taille effective de l'échantillon.

L'un des principaux inconvénients de l'utilisation de la substitution est une tendance à y recourir au lieu de faire tous les efforts possibles pour obtenir les réponses auprès des unités originales. Donc, l'utilisation de procédures de substitution requiert que des mesures de contrôle appropriées soient prises pour s'assurer qu'un effort maximal est fait en vue d'obtenir les réponses auprès des unités originales de l'échantillon. Un autre inconvénient est la tendance à ignorer le niveau et la fréquence de la substitution dans le calcul des taux de réponse à l'enquête.

b. Rappels – Dans de nombreuses enquêtes, les rappels sont utilisés abondamment afin de réduire la non-réponse et les biais résultants. Les rappels peuvent prendre diverses formes selon le type d'enquête. Dans le cas des enquêtes par la poste ou par interview, il peut s'agir d'une lettre, d'un appel téléphonique ou d'une interview sur place. Dans les enquêtes téléphoniques et les enquêtes par interview sur place, les appels répétés sont la forme normale de rappel. Il est donc nécessaire d'étudier divers types de rappels en ce qui a trait à la qualité des données, du coût et des réactions des répondants.

Les rappels peuvent être utilisés uniquement pour réduire la non-réponse, ou pour fournir de l'information à un système d'imputation, ou pour étudier l'effet sur les quantités.

c. Règle de sélection des répondants – Afin d'éviter toute ambiguïté quant à l'admissibilité des répondants à une enquête donnée, une procédure appelée « règle de sélection des répondants » doit être définie et suivie sur le terrain.

Pour les enquêtes qui comportent un répondant désigné, deux règles sont utilisées le plus souvent :

1. le répondant désigné doit être interviewé et s'il est capable de répondre, mais non disponible, des rappels répétés sont effectués jusqu'à ce qu'un contact puisse être établi;
2. si le répondant désigné est absent ou incapable de répondre (sourd, malade, etc.), un répondant de substitution est choisi. Les variantes de cette règle comprennent différentes définitions des répondants de substitution permis.

Pour les enquêtes dans lesquelles les réponses pour chaque membre admissible du ménage sont requises, il existe un choix entre plusieurs règles de sélection des répondants possibles :

1. chaque membre du ménage doit répondre personnellement (autoréponse);
2. un membre du ménage peut répondre pour chaque membre du ménage (réponse par personne interposée);
3. un mélange entre l'autoréponse et la réponse par personne interposée, c'est-à-dire que certaines réponses sont des autoréponses, tandis que d'autres sont des réponses par personne interposée selon que le répondant est disponible au moment de l'interview ou qu'une règle particulière de sélection du répondant est appliquée, par exemple, les personnes non apparentées au chef du ménage doivent répondre elles-mêmes.

Les études méthodologiques des effets de l'utilisation de diverses règles de sélection des répondants ont porté principalement sur deux aspects fondamentaux. Le premier comporte les différences entre les nombres de

rappels nécessaires pour prendre contact avec le répondant souhaité, approche proposée par Deming et Cochran, 1977; le second comprend les différences de qualité des données obtenues qui peuvent être évaluées grâce à un programme de ré-interview des répondants originaux.

L'utilisation d'un répondant substitut diminue le nombre de rappels, donc réduit le coût de l'enquête et le temps nécessaire pour obtenir les données. Par ailleurs, il peut être désavantageux d'utiliser un répondant substitut du fait que les données fournies par ce dernier pourraient être moins exactes que celles obtenues par autoréponse. L'application d'une règle particulière de sélection des répondants doit être examinée très soigneusement en regard du type et de la qualité des données requises, du coût de l'obtention des données et de l'actualité de ces dernières pour la publication. Ces considérations varieront d'une enquête à l'autre selon le sujet de l'enquête, le budget et l'organisation du travail sur le terrain.

Imputation

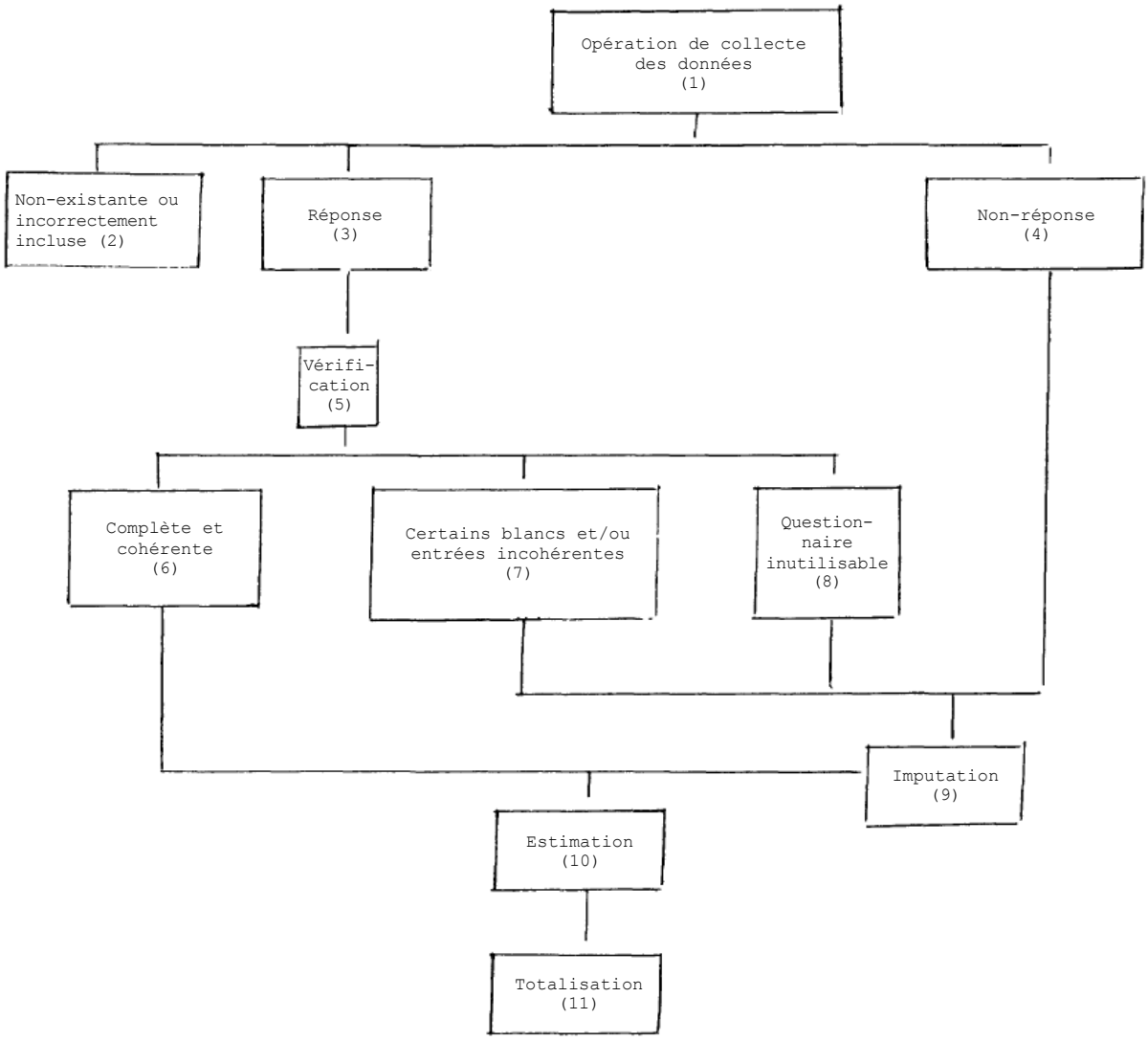
À l'étape du traitement et de l'estimation, les données d'enquête sont habituellement classées en fonction de la non-réponse totale, de la non-réponse partielle et des réponses invalides.

Il est très important d'intégrer un système de contrôle efficace dans le plan d'enquête, c'est-à-dire s'assurer que les unités sélectionnées (et non les autres) soient interviewées, que les unités non déclarantes soient classées convenablement (par exemple, non-réponse, non existantes), que les lacunes dans la base de sondage soient décelées, que la saisie des données soit complète, etc.

Il existe divers moyens de traiter les réponses incomplètes ou invalides. Chacun d'eux aboutit à l'attribution d'une valeur pour les données manquantes ou invalides, à moins qu'il soit décidé de publier des données « brutes ». La procédure d'attribution de la valeur est appelée imputation, et une valeur imputée est supposée être en rapport avec la caractéristique du non-répondant. Donc, on produit un ensemble de données « épuré », c'est-à-dire qu'une valeur est donnée à chaque unité dans l'enquête. Avant de discuter des diverses méthodes d'imputation, examinons les problèmes conceptuels d'imputation.

À mesure que l’information chemine de la collecte des données à la totalisation, les divers types de réponses peuvent être identifiés et ces types de réponses sont présentés au graphique 1.

Graphique 1 : Graphique de cheminement pour chaque unité échantillonnée



Il s’agit, naturellement, d’un diagramme très simplifié du processus, présenté uniquement pour les besoins de la discussion du présent article.

Le graphique 1 montre que deux des trois groupes de questionnaires obtenus après l'étape de la vérification requièrent une mesure supplémentaire avant l'estimation. Il s'agit des groupes (8) questionnaire inutilisable, (7) questionnaire contenant certains blancs et/ou entrées incohérentes et (4) questionnaire contenant une non-réponse. Un questionnaire inutilisable pourrait être classé comme une non-réponse totale ou être associé à un ménage répondant avec certains blancs ou entrées incohérentes. Dans chaque cas, cependant, une mesure supplémentaire désignée (8) imputation serait nécessaire. Les cas de non-réponse totale sont habituellement pondérés d'une certaine manière, sauf pour le recensement. Par ailleurs, les questionnaires déficients rentrent dans deux catégories, à savoir (7) entrées incohérentes ou réponses en blanc illégitimes.

Les entrées incohérentes peuvent être des impossibilités logiques ou des entrées plausibles, mais hautement improbables. Il semble naturel, si les entrées sont des impossibilités logiques et qu'elles peuvent être décelées en tant que telles, de devoir les ajuster, même s'il se peut qu'elles n'influencent pas fortement les données. L'ajustement évitera bien de l'embarras aux analystes de la division spécialisée associés aux rapports publiés.

Dans le cas d'entrées plausibles, mais très improbables, un choix difficile se présente entre laisser ce qui pourrait sembler être une distribution non naturelle ou éliminer de la distribution les valeurs extrêmes qui pourraient en fait représenter la situation réelle. Idéalement, le choix entre les deux options doit se faire en se basant sur l'expérience concernant les mécanismes d'erreur et sur la nature de la distribution véritable fondée sur le savoir des spécialistes du domaine. En tout cas, il faut pouvoir cerner les cas posant problème, c'est-à-dire disposer de règles de vérification appropriées quand on est confronté à des événements impossibles ou fortement improbables, ainsi que d'une méthode pour les traiter (c'est-à-dire imputation).

Il existe une différence fondamentale entre la vérification et l'imputation. Considérons l'ensemble de toutes les combinaisons possibles de codes sur un questionnaire. La vérification peut être définie comme la division de cet ensemble en deux sous-ensembles s'excluant mutuellement, à savoir les combinaisons qui sont jugées acceptables et celles qui sont jugées inacceptables, les dernières

incluant les questionnaires contenant des champs en blanc invalides et des entrées incohérentes. Donc, la vérification est essentiellement un diagnostic et, du point de vue opérationnel, elle doit être définie par un ensemble de règles. Par contre, l'imputation a davantage la nature d'un traitement des données.

L'imputation peut être définie comme l'attribution de données à des champs vides (y compris la non-réponse totale) ou au replacement de données invalides dans certains champs en suivant un ensemble de règles. Il n'existe aucune méthode d'imputation sans biais connue, mais certaines méthodes peuvent être plus appropriées que d'autres.

Il se peut qu'au lieu d'effectuer une imputation pour la non-réponse au moment où sont préparées les totalisations des données de l'enquête, les totalisations soient présentées accompagnées du taux de non-réponse. Dans ce cas, les utilisateurs pourraient choisir entre diverses méthodes d'imputation à partir de données totalisées. À première vue, cette approche semble avoir des avantages, puisqu'elle donne aux utilisateurs la possibilité de choisir leur propre méthode d'imputation. Cependant, elle présente de sérieux inconvénients. En raison des méthodes d'imputation différentes employées, des estimations contradictoires seraient produites par divers utilisateurs et des problèmes de cohérence et d'intégration des données auraient lieu. De surcroît, l'organisme chargé de la collecte des données est habituellement mieux placé, en raison de sa proximité par rapport aux sources des données, pour prendre des décisions d'imputation. Pour ces raisons, l'imputation est ordinairement effectuée par les organismes de collecte des données plutôt que par les utilisateurs individuels. Le principe général de l'imputation repose sur l'espérance qu'une procédure appropriée, qu'il s'agisse de la non-réponse ou de champs en blanc résultant du rejet à la vérification, fournira une relation plus logique entre les données croisées et réduira également l'erreur quadratique moyenne des estimations.

La situation la plus simple a lieu quand une seule valeur possible peut être imputée pour un champ de façon telle qu'après l'imputation, l'enregistrement soit cohérent. On parle alors d'imputation déterministe. Par exemple, si le champ « épouse » est codé « masculin », il n'existe qu'une seule valeur possible à imputer pour le champ « sexe » pour assurer la cohérence avec cette information. Parfois, plus d'une valeur rendrait l'enregistrement cohérent. Le cas échéant, on choisirait

une valeur particulière qui prédomine davantage en ce qui concerne la proportion de la fréquence totale ou qui est plus plausible. Un bon exemple de cette situation s'observe dans le cas de l'Enquête sur la population active, pour laquelle, durant les mois allant de l'automne au printemps, pour les personnes de 15 et 16 ans, si aucune caractéristique de population active n'est entrée, on impute qu'elles « fréquentent l'école », même s'il n'est pas du tout impossible qu'elles ne le fassent pas. À condition que la proportion de cas de ce genre soit suffisamment faible, cette imputation aura pour effet une légère augmentation du biais, mais il y aura une certaine réduction de la variance.

Dans d'autres situations où on pourrait raisonnablement imputer toute une gamme de valeurs, d'autres critères sont nécessaires. Un critère possible consisterait à minimiser l'erreur quadratique moyenne des estimations résultantes. Toutefois, la question est l'erreur quadratique moyenne de quelle estimation ? Étant donné la demande sans cesse croissante de microdonnées totalisées de plusieurs façons différentes et imprévues, on ne sait pas vraiment quelle erreur quadratique moyenne on doit minimiser. En outre, on ne connaîtrait pas tous les types d'agrégats auxquels un enregistrement particulier pourrait contribuer dans différents types de totalisations. Par conséquent, on pourrait préférer utiliser un autre critère qui produirait l'entrée la plus appropriée pour un champ dans un enregistrement particulier par rapport au reste de l'information figurant dans l'enregistrement. Autrement dit, quel est le meilleur moyen de prédire la valeur d'un champ en se basant sur ce que l'on sait des autres champs de l'enregistrement ? Un bon exemple de ce genre d'imputation est l'utilisation des données du mois précédent dans l'Enquête sur la population active; pour une personne particulière, il serait difficile de trouver une meilleure valeur imputée, particulièrement dans les cas où les caractéristiques démographiques évoluent lentement. Si l'on ne dispose pas d'information fondée sur le passé, on devra recourir à des méthodes telles que l'imputation par la régression ou l'imputation hot deck.

4 Procédures d'imputation

Depuis plusieurs années, diverses procédures d'imputation des données manquantes dues à la non-réponse ont été appliquées dans les enquêtes auprès des ménages et les recensements. Le choix d'une procédure

particulière était, autant que je sache, justifié principalement en se basant sur la convenance, l'intuition et l'expérience. Il était souvent supposé que les probabilités que les unités répondent étaient uniformes, et le biais de non-réponse était en grande partie ignoré.

Bien que des variations des taux de réponse aient été décelées entre les unités en fonction de leurs caractéristiques, l'effet d'unités individuelles répondant ou ne répondant pas sur le biais et la variance des estimations n'a habituellement pas été examiné suffisamment.

Pour faciliter un examen détaillé de l'effet, Platek et Gray (1980) ont élaboré une méthodologie concernant le biais et la variance découlant de plusieurs procédures d'imputation. L'élaboration de l'expression du biais et de la variance des estimations s'appuie sur le concept fondamental qu'une unité, si elle est sélectionnée, répond ou ne répond pas avec une certaine probabilité de réponse associée à chaque unité. Il s'agit d'une extension de l'approche adoptée par Platek, Singh et Tremblay en 1978 en ce qui concerne les recensements.

La définition de diverses procédures d'imputation comprend les éléments suivants :

i) l'utilisation de cellules pour l'imputation; les cellules peuvent être des régions d'équilibrage ou des classes de pondération; ou

ii) l'ajustement des pondérations en utilisant les probabilités de réponse estimées à l'intérieur des cellules.

Les régions d'équilibrage sont souvent appelées « régions d'équilibrage dépendantes du plan de sondage » pour les besoins de l'imputation. Une région d'équilibrage est une région géographique dans laquelle un échantillon déficient en raison de données manquantes est agrandi jusqu'au niveau prescrit par imputation des données manquantes. Habituellement, une région d'équilibrage est une strate, mais il pourrait s'agir d'autres régions dépendantes du plan de sondage, telles qu'une unité primaire d'échantillonnage, une grappe, un groupe de strates ou même l'échantillon entier. Les régions d'équilibrage peuvent être délimitées avant ou après la réalisation de l'enquête.

Les classes de pondération sont définies par les poststrates (strates définies après l'échantillonnage) formées en se fondant sur l'information

au sujet des répondants et des non-répondants dans l'échantillon. L'information peut être obtenue auprès des non-répondants partiels pour lesquels certaines caractéristiques sont connues, même si la caractéristique particulière que l'on veut estimer est inconnue pour ces unités, comme dans le cas de la non-réponse partielle, par exemple. Les caractéristiques utilisées dans les poststrates pourraient également être obtenues auprès de sources externes. Du point de vue opérationnel, une classe de pondération est très similaire à une région d'équilibrage, sauf que les unités ayant des caractéristiques similaires sont groupées en classes ou poststrates sans tenir compte du plan de sondage original. Le choix des caractéristiques et la taille des régions d'équilibrage et des classes de pondération sont importants, car la variance et le biais d'une estimation calculée d'après l'échantillon dépendront de l'homogénéité des caractéristiques des répondants et des non-répondants dans les régions d'équilibrage et les classes de pondération.

Les régions d'équilibrage et les classes de pondération sont également appelées cellules ou unités d'équilibrage.

a) « Pondération » dans les cellules d'ajustement

La « méthode d'imputation par pondération », applicable en pratique pour éliminer la non-réponse, consiste à gonfler les poids de sondage ou les probabilités d'inclusion inversées en les multipliant par l'inverse du taux de réponse dans une cellule. Implicitement, la valeur imputée pour la valeur manquante de chaque non-répondant est égale à la moyenne des valeurs de tous les répondants dans la cellule, avec certains ajustements pour tenir compte des probabilités de sélection.

Si une cellule « b » contient n_b unités dans l'échantillon et qu'un nombre m_{by} d'entre elles ont répondu à une question particulière ou à plusieurs questions qui détermineraient la caractéristique y, alors l'estimation de Horvitz-Thompson[1] du total de la caractéristique y dans cette cellule sera donnée par (en laissant tomber y dans m_{by}),

$$\hat{Y}_{b1} = \frac{n_b}{m_b} \sum_{i=1}^{n_b} \delta_{iy} Y_i/\Pi_i = \sum_{i=1}^{n_b} \delta_{iy} Y_i / (\Pi_i \hat{\alpha}_{iy}) ; \quad \text{où} \qquad (1)$$

[1] En se limitant à cet estimateur uniquement.

δ_{iy} = 1 ou 0 selon que l'unité i répond ou ne répond pas pour la caractéristique « y ».

$\hat{\alpha}_{iy} = m_b/n_b$ est une estimation de la probabilité de réponse, qui peut ou non être égale à la probabilité de réponse réelle, donnée par $E\delta_{iy}$ qui est définie par α_{iy}.

Y_i = réponse pour l'unité I comme il est défini plus haut.

Une région dans laquelle les ajustements sont effectués serait constituée de cellules d'ajustement mutuellement exclusives et exhaustives, de sorte que les estimations globales du total pour la caractéristique y seraient données par la somme des estimations sur l'ensemble des cellules; c'est-à-dire

$$\hat{Y}_1 = \sum_b \hat{Y}_{b1} . \qquad (1)$$

L'expression (1) montre que $\hat{Y}_{b1}$ peut être considérée comme une estimation pondérée à la hausse où le poids de l'unité i comprend la probabilité de sélection inverse π_i^{-1} et l'estimation de la probabilité de réponse inverse, données par $\hat{\alpha}_{iy}^{-1}$ ou $(m_b / n_b)^{-1}$. Puisque nous connaissons rarement les probabilités de réponse individuelles au moment du traitement des données, nous devons employer la meilleure estimation disponible des probabilités de réponse et, en choisissant convenablement les régions d'équilibrage ou les classes de pondération, cette estimation est habituellement le taux de réponse dans la cellule ou la moyenne estimée des probabilités de réponse des unités.

Il est facile de montrer que $\hat{Y}_1$ peut présenter un biais de réponse et un biais de non-réponse ou d'imputation donnés par :

$$\sum_b \bar{\alpha}_{by}^{-1} \sum_{i=1} \alpha_{iy} B_{Riy}$$

et

$$\sum_b \bar{\alpha}_{by}^{-1} \sum_{i=1} (\alpha_{iy} - \bar{\alpha}_{by}) Y_i \qquad (2)$$

respectivement.

Ici, $\bar{\alpha}_{by}$, est le taux de réponse prévu ou la probabilité de réponse moyenne prévue dans la cellule « b », en ce qui concerne la caractéristique y. $\bar{\alpha}_{by} = \sum_i \pi_i \alpha_{iy} / \sum_i \alpha_i$, la sommation étant effectuée sur l'ensemble des unités dans la cellule « b ».

B_{Riy} est le biais de réponse pour l'unité i en ce qui concerne la caractéristique y, c'est-à-dire $Ey_i | (\delta_{iy} = 1) = Y_i + B_{Riy}$.

Le biais d'imputation existera seulement si les probabilités de réponse α_{iy} varient à l'intérieur d'une cellule d'ajustement et si une corrélation existe entre les probabilités de réponse, α_{iy}, et la caractéristique, Y_i.

b) Méthode de répétition

La « méthode d'imputation par répétition » consiste à combler la déficience de l'échantillon dans une cellule en raison de la non-réponse par répétition de l'ensemble des répondants ou d'un sous-échantillon de ceux-ci.

Dans la méthode de répétition, si nous considérons une classe de pondération b contenant n_b unités sélectionnées et m_b répondants en ce qui concerne la caractéristique y, l'estimation dans la cellule « b » sera donnée par :

$$\hat{Y}_{b2} = \sum_{i=1}^{n_b} \delta_{iy} W_{iy} Y_i / \Pi_i \qquad (3)$$

où W_{iy} est le nombre de fois que l'unité i est répétée pour tenir compte $(n_b - m_b)$ non-répondants dans la cellule « b » de façon que $\sum_i \delta_{iy} W_{iy} = n_b$. Notons que W_{iy} est défini uniquement pour les répondants, c'est-à-dire quand $\delta_{iy} = 1$.

Il existe plusieurs moyens de répéter les unités. L'un de ceux-ci, et le plus simple à traiter du point de vue du développement méthodologique, est la sélection aléatoire d'unités sans remise pour la répétition.

c) Hot deck

Les procédures hot deck sont des méthodes courantes d'ajustement des ensembles de données pour tenir compte des valeurs manquantes. En général, une procédure hot deck est un processus de répétition - une valeur déclarée dans l'échantillon est répétée pour représenter une

valeur manquante. Donc, les termes « procédure d'imputation » et « procédure hot deck » ne sont pas interchangeables. Une procédure qui impute la moyenne de toutes les valeurs déclarées pour chaque valeur manquante est une imputation, mais non une procédure hot deck.

La raison principale de l'utilisation d'une procédure hot deck ou cold deck est d'essayer de réduire le biais de non-réponse. La différence essentielle entre les deux procédures tient à la façon dont l'information pour les données manquantes est spécifiée. Une procédure cold deck est déterministe, c'est-à-dire que dans chaque condition spécifiée, la même valeur est substituée pour une non-réponse partielle. Dans les procédures hot deck, pour chaque condition spécifiée, la valeur substituée pour la non-réponse partielle est la valeur de l'item en question qui a été trouvée dans le dernier enregistrement « acceptable ». Donc, dans la procédure hot deck, la substitution est probabiliste, et reflète les fréquences des items trouvés dans les enregistrements acceptables satisfaisant les mêmes conditions.

En tant que méthode d'imputation, les procédures hot deck présentent des caractéristiques intéressantes, dont les suivantes : a) les procédures aboutissent à un moyen relativement facile de construire des poststrates [voir I.P. Fellegi et Holt], b) l'appariement des enregistrements ne pose pas de problèmes particuliers et c) aucune hypothèse de modélisation forte n'est nécessaire afin d'estimer les valeurs individuelles pour les items manquants.

En ce qui concerne l'évaluation des procédures hot deck, on voudrait savoir quels sont les effets de la taille des groupes de classification (souvent appelés classes de pondération), la fréquence des données manquantes, le choix des items d'appariement, etc., sur le biais et la fiabilité des principales estimations. Certains travaux théoriques concernant les procédures hot deck ont été effectués par (I.P. Fellegi et Holt), (Bailar et Bailar 1978) et (Cox B et Folsom R.E. 1978).

Les travaux théoriques à venir consisteront à essayer de généraliser les procédures hot deck qui ont été élaborées et à établir des expressions pour le biais et la variance des estimations fondées sur les procédures. Le biais résulte de la probabilité estimée de réponse à un item concernant une unité par rapport à la probabilité de réponse réelle (qui est évidemment inconnue) et de la corrélation entre la valeur de l'item et

la probabilité de réponse. La variance comprend des composantes supplémentaires au-delà de celles qui entrent en jeu quand la pondération est appliquée dans une classe de pondération. En fonction des méthodes et des contraintes imposées à la procédure hot deck, les termes de variance supplémentaires peuvent devenir assez compliqués.

Quoi qu'il en soit, une extension de la théorie concernant la variance élaborée jusqu'à présent par Bailar et Bailar semble être la voie à suivre la plus prometteuse.

d) Méthode de substitution de données historiques

La « méthode de substitution de données historiques » consiste à substituer des données provenant de sources historiques ou externes, comme le Recensement, une enquête antérieure ou des données administratives, pour une unité en vue de remplacer les données manquantes causées par la non-réponse. Les deux cas d'application de la « méthode de substitution de données historiques » qui suivent peuvent être envisagés :

Cas (i) où des données de source historique ou externe sont disponibles pour toutes les unités qui ont omis de répondre;

Cas (ii) où les données de source externe sont disponibles pour certaines unités, mais pas toutes, et une imputation par une autre méthode, par exemple, « par pondération », doit également être appliquée.

Quand des données de source externe sont disponibles pour chaque unité, la valeur imputée des données manquantes est donnée par z^1_{iy}, la valeur observée de la caractéristique y^1 dans l'enquête précédente, le Recensement ou le fichier de données administratives. L'estimation est alors donnée par :

$$\hat{Y}_3 = \sum_{i=1}^{n} [\delta_{iy}\, y_i/\Pi_i + (1 - \delta_i)\, z^1_{iy}\ /\Pi_i] \qquad (4)$$

Il convient de souligner que y_i ainsi que z^1_{iy} sont sujettes à une erreur de réponse (pouvant toutes deux contenir un biais de réponse et une variance de réponse) en ce qui concerne la valeur vraie pour l'unité i. Si des données de source historique ou externe sont disponibles pour chaque unité qui omet de répondre pour la caractéristique étudiée, des

cellules d'ajustement telles que celles utilisées dans les méthodes de pondération et de répétition deviennent redondantes.

On peut montrer que le biais de l'estimation $\hat{Y}_3$ est la somme de deux composantes, à savoir

$\sum_{i=1}^{N} \alpha_{iy} B_{Riy}$ ··· le biais de réponse, dû aux erreurs de réponse pour les réponses courantes,

et $\sum_{i=1}^{N} (1-\alpha_{iy}) B^1_{Riy}$ ··· le biais d'imputation, dû à la substitution de données historiques et l'introduction du biais de réponse des enregistrements historiques concernant la caractéristique « y » dans l'enquête courante. Ici B^1_{Riy} est le biais d'imputation se rapportant à l'unité i.

Quand des données de source historique ou externe ne sont disponibles que pour certains non-répondants, une autre procédure d'imputation, telle que la pondération, doit être appliquée en même temps que la substitution de données historiques. Quand la pondération est conjuguée à la substitution de données historiques, des cellules d'ajustement sont nécessaires et l'estimation est donnée par :

$$\hat{Y}_4 = \sum_b \hat{Y}_{b4}$$

$$\hat{Y}_{b4} = \frac{n_b}{m_b + m_b^1} \sum_{i=1}^{n_b} [\delta_{iy}\, y_i/\pi_i + (1-\delta_{iy})\, \delta^1_{iy}\, Z^1_{iy}/\pi_i] \qquad (5)$$

Dans la formule (5), $\delta^1_{iy} = 1$ ou 0 selon que des données de source historique ou externe représentant la caractéristique y de l'unité i sont disponibles ou non. Le taux de réponse estimé $(m_b + m_b^1)/n_b$ contient deux composantes; m_b/n_b, le taux de réponse habituel pour l'enquête courante tel qu'il est défini en (1) et m_b^1/n_b, le taux de réponse pour des données historiques acceptables parmi les $(n_b + m_b)$ non-répondants de l'enquête courante.

La quantité $(m_b + m_b^1)$ égale $\sum_{i=1}^{n_b} [\delta_{iy} + (1+\delta_{iy})\delta^1_{iy}]$, le nombre de répondants et de non-répondants pour lesquels des données historiques acceptables sont disponibles.

e) Méthode de substitution de la valeur zéro

La « méthode d'imputation par substitution de la valeur zéro » consiste à ignorer les données manquantes dues aux non-répondants, ce qui implique de substituer zéro pour leurs valeurs manquantes.

Dans certains cas, les cellules manquantes sont simplement étiquetées « non déclaré » et il est laissé à l'analyste le choix de traiter les données de la façon qui lui convient le mieux.

5 Probabilités de réponse estimées

Le biais de l'estimation sous chaque procédure d'imputation, à part le biais de réponse, résulte du fait que les probabilités de réponse estimées diffèrent des probabilités de réponse réelles, ainsi que d'une corrélation entre les valeurs de la caractéristique et les probabilités de réponse réelles. En délimitant des cellules pour les besoins de l'imputation, par exemple des strates ou des grappes, on s'efforce d'employer des probabilités de réponse estimées aussi proches que possible des valeurs réelles. La procédure la plus courante consiste à utiliser des taux de réponse qui sont équivalents à des estimations des probabilités de réponse moyennes dans les cellules d'ajustement. Les probabilités de réponse estimées sont présentées au tableau 3 selon la procédure d'imputation.

Tableau 3
Probabilité de réponse estimée $\hat{\alpha}_{iy}$ selon la procédure d'imputation

Procédure	Formule estimée (numéro entre parenthèses)	Probabilité de réponse estimée
Pondération	$\hat{Y}_{b1}$ (1)	m_b / n_b
Répétition	$\hat{Y}_{b2}$ (3)	$[n_b / m_b]^{-1}$ ou $[n_b / m_b + 1]^{-1}$
Substitution de données historiques		
Cas (i)	$\hat{Y}_3$ (4)	1 (incluant la probabilité de données historiques disponibles)
Cas (ii)	$\hat{Y}_{b4}$ (5)	$(m_b + m_b^1) / n_b$ (incluant la probabilité de données historiques disponibles parmi les non-répondants)

Les probabilités de réponse estimées sont, dans tous les cas, sauf pour $\hat{Y}_3$, obtenues par le taux de réponse dans la cellule « b ».

6 Variances des estimations

Sous le concept de probabilités de réponse plutôt que le concept de strates de répondants et de non-répondants, le développement des variances des estimations sous diverses procédures d'imputation aboutit à des expressions très complexes tant pour la variance d'échantillonnage que pour la variance non due à l'échantillonnage. Les composantes de ces expressions sont les suivantes :

i) variance d'échantillonnage;

ii) variance de réponse simple;

iii) variance de réponse corrélée;

iv) composante de variance due à la variation en cas de réponse/non-réponse pour chaque unité;

v) composante de covariance due à la covariance entre les événements de réponse/non-réponse pour chaque paire d'unités.

Les trois premières composantes ont été traitées en profondeur dans le cas d'une réponse complète par Fellegi (1964) sous échantillonnage aléatoire simple sans remise et par Koch (1973) sous échantillonnage avec probabilité proportionnelle à la taille sans remise. Les deux dernières composantes ont été développées par R. Platek et G. Gray (1978).

On peut montrer que la méthode de répétition produira une variance légèrement plus grande que la simple augmentation des poids dans les classes de pondération ou les régions d'équilibrage (voir Hansen et coll. 1953), et que l'accroissement s'applique aux cinq composantes. La substitution de données historiques produira presque certainement des variances plus faibles si les données historiques sont fortement corrélées aux valeurs courantes des caractéristiques, puisque la forte corrélation résultera effectivement en un plus grand échantillon. De nouveau, les cinq composantes seraient vraisemblablement réduites. La méthode de substitution de la valeur zéro présente une plus faible variance d'échantillonnage, principalement en raison de la sous-estimation du total; toutefois, les variances non dues à l'échantillonnage ne sont pas nécessairement plus faibles et l'erreur quadratique moyenne globale de l'estimation par la méthode de substitution de la valeur zéro est vraisemblablement plus grande que celle des autres estimations en raison de la sous-estimation importante.

Les résultats susmentionnés ont été corroborés par un exemple hypothétique présenté par Platek et Gray [1978].

7 Application de l'imputation aux enquêtes auprès des ménages à Statistique Canada

Les procédures d'imputation fréquemment utilisées dans les enquêtes par sondage et les recensements au Canada comprennent l'ajustement des pondérations, la répétition, la substitution de données historiques ou externes et les procédures hot deck.

Enquête sur la population active

L'une des principales enquêtes continues réalisées par Statistique Canada est l'Enquête sur la population active d'après laquelle sont obtenues des estimations mensuelles du chômage, de l'emploi et de nombreuses autres caractéristiques. Les données sont publiées quelques semaines après la réalisation de l'enquête auprès d'environ 56 000 ménages durant approximativement la troisième semaine de chaque mois. Il est impossible de prendre contact avec chacun des ménages avec lesquels il aurait fallu le faire à cause de l'échéancier rigoureux de collecte et de traitement des données. Certains ménages sont absents pendant la semaine complète, absents chaque fois que l'intervieweur appelle, ou bien ils refusent d'être interviewés. La non-réponse peut bien sûr avoir d'autres causes, mais celles-ci représentent seulement une faible part de la non-réponse totale, qui est maintenue autour de 5 % la plupart du temps à part des augmentations à 7 % ou 8 % durant les mois d'été en raison d'« absences temporaires » dues aux vacances.

L'imputation pour la non-réponse complète des ménages est effectuée en fonction des critères qui suivent, i) pour environ le tiers des non-répondants, substitution des valeurs du mois précédent s'il y a lieu (avec des transformations appropriées dans certains champs pour mettre à jour les données du mois précédent) ou ii) gonflement par l'inverse du taux de « réponse » dans les unités d'équilibrage (la « réponse » incluant dans ce cas les valeurs substituées pour les non-répondants qui ont effectivement répondu à l'enquête précédente et fourni des données applicables). Dans le cas de (ii), l'imputation pour les deux tiers restants des non-répondants est implicitement la moyenne de l'ensemble des répondants, dans l'unité d'équilibrage (unité primaire

d'échantillonnage, petites parties urbaines et rurales des UPE, ou sous-unités dans les grandes villes).

L'imputation pour la non-réponse partielle ou les rejets à la vérification est effectuée de l'une de trois façons, en fonction du ou des items pour lesquels des données manquent ou sont incorrectes et en fonction de la situation de réponse et des caractéristiques de l'unité lors de l'enquête précédente.

i) la bonne réponse à l'item qui a été omis peut être déduite sans ambiguïté d'après la suite du questionnaire (une table de décision rendrait certaine une réponse unique et cohérente);

ii) la substitution de la réponse à l'item de l'enquête précédente si elle est disponible et si elle est appropriée selon une table de décision;

iii) l'application d'une procédure hot deck en vertu de laquelle un enregistrement similaire est obtenu dans la même UPE, en suivant le même cheminement (un de six possibles) dans la série de questions, et pour le même groupe âge-sexe. Ici, un regroupement des classes de pondération peut être nécessaire pour trouver un enregistrement de type similaire; habituellement les catégories âge-sexe plutôt que les UPE ou les cheminements suivis dans le questionnaire sont groupées pour les besoins de l'imputation s'il y a lieu.

Afin d'illustrer la procédure d'imputation pour des questionnaires partiellement remplis, considérons les exemples qui suivent, tirés de l'Enquête sur la population active (EPA) au Canada.

Exemple 1 (Enquête sur la population active au Canada)

Chaque personne de 15 ans et plus faisant partie des ménages sélectionnés pour l'EPA se voit poser la question 80 (Q80) « Fréquente-t-il/elle l'école ? » Dans l'affirmative, la réponse est codée « 1 », et dans la négative, elle est codée « 2 ». Pour les personnes ne fréquentant pas l'école, aucune question connexe n'est posée. Cependant, pour celles fréquentant l'école, les questions de suivi sont : question 81 (Q81) « Est-il/elle inscrit(e) comme étudiant(e) à temps plein ou à temps partiel ? ». Dans l'affirmative, la question 82 (Q82) est posée : « De quel genre d'école s'agit-il ? ». Donc, la situation suivante peut se produire.

Q80 ≠ 1 ou 2, ce qui est une erreur.

La relation requise entre les questions doit être du type suivant :

a. si Q80 = 1, c'est-à-dire fréquente une école,
   alors Q81 = 1 ou 2, c'est-à-dire à temps plein ou à temps partiel
   et Q82 = 1 ou 2 ou 3 ou 4, c'est-à-dire le type d'école;

b. si Q80 = 2, c'est-à-dire ne fréquente pas une école,
   alors Q81 = blanc (sans objet)
   et Q82 = blanc (sans objet).

Les réponses à d'autres questions peuvent aussi être reliées à Q80 et Q82, à savoir : Q14 et Q36 : Fréquentation d'une école comme raison de travailler < 30 heures, Q58 « Va à l'école » comme réponse à la question de savoir ce que le répondant faisait immédiatement avant de chercher du travail et « Va à l'école », comme raison, à la question Q64, de ne pas être disponible pour travailler la semaine précédente. Dans chaque cas, le code est « 3 ».

En guise d'exemple d'un ensemble particulier de conditions**, supposons que la réponse à Q80 n'est pas disponible ou qu'elle ne concorde pas avec la logique des questions subséquentes. La table de décision qui suit résume les étapes suivies pour imputer la valeur appropriée pour Q80.

	Règle d'imputation						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Q81 = 1 ou 2	O	N	N	N	N	N	N
Q82 = 1, 2, 3 ou 4	O	N	N	N	N	N	N
Q14, 36, 58 ou 64 = 3	|	O	N	N	N	N	N
Âge = 15 ou 16 ans	|	|	O	O	N	N	N
Mois d'enquête = juillet ou août*	|	|	O	N	N	N	N
Q80 le mois précédent = 1	|	|	|	|	O	N	N
= 2	|	|	|	|	|	O	N
≠ 1 ou 2	|	|	|	|	|		O
Alors Q80 le mois courant	1	1	2	1	1	2	hot deck

Ici, hot deck signifie une recherche dans la même UPE, le même cheminement de questions, et le même groupe âge-sexe comme il est mentionné plus haut.

* Pour les personnes non âgées de 15 ou de 16 ans, cela n'a pas d'importance que le mois d'enquête soit juillet ou août pour les règles d'imputation 5, 6 et 7.

** Cela exclut d'autres conditions, par exemple si Q81 n'est pas 1 ou 2, mais que Q82 est 1, 2, 3 ou 4, et inversement.

Chaque colonne représente une série distincte d'étapes qui doivent être suivies afin d'arriver à la valeur imputée indiquant si une personne fréquente ou non l'école (1 ou 2). Les O = Oui et N = Non dans le corps de la table de décision correspondent aux énoncés de condition.

Voici la discussion détaillée de chaque règle d'imputation :

1. La valeur imputée est fondée sur la logique interne des questions connexes. Donc, si les réponses à Q81 et Q82 sont valides, alors seule la valeur de Q80 concordant avec l'autre information est attribuée, c'est-à-dire si Q81 = 1 ou 2 et Q82 = 1, 2, 3 ou 4, alors imputer « 1 » pour Q80, c'est-à-dire va à l'école.

2. Si l'information pour Q81 et Q82 n'est pas disponible, alors l'étape suivante consiste à chercher l'information pertinente ailleurs dans le questionnaire. Les autres questions sont posées en fonction du « cheminement » qu'une personne peut avoir suivi. Si elle a indiqué plus tôt qu'elle allait à l'école, alors Q80 est également codée comme cela, c'est-à-dire si Q81 ≠ 1, 2 et Q82 ≠ 1, 2, 3, 4, mais Q14, 36, 58 ou 64 = 3, alors Q80 = 1, c'est-à-dire va à l'école.

3-4. Si aucune information directement reliée n'est disponible, alors utiliser l'information qui est indirectement reliée à la question en conflit à la vérification. Ici, il s'agit de l'âge et du mois de l'enquête. Si une personne est âgée de 15 ou de 16 ans, alors durant les mois de juillet et août, elle ne va fort probablement pas à l'école et donc Q80 est codée 2, c'est-à-dire ne va pas à l'école. Par contre, pour tous les autres mois, la réponse est codée comme fréquentant l'école, c'est-à-dire Q81 = 1, c'est-à-dire va à l'école.

   Ces imputations sont des exemples de cas rares, de sorte que l'attribution de la « valeur la plus probable » est justifiée, même si l'imputation hot deck pourrait théoriquement être meilleure, mais éventuellement plus coûteuse.

5-6. Si aucune information n'est disponible dans les données du mois courant et que la personne n'a ni 15 ni 16 ans, l'option suivante est d'utiliser les données du mois précédent. Quelle

qu'elle soit, si la réponse du mois précédent est disponible, elle est transférée dans le questionnaire du mois courant.

L'utilisation des données du dernier mois est justifiée en grande partie par le fait que la corrélation des caractéristiques de certaines estimations d'un mois à l'autre est assez élevée.

7. Si aucune information directement reliée à la fréquentation scolaire n'est disponible, alors une valeur de réponse est imputée à partir d'un autre enregistrement similaire. Ici, l'enregistrement similaire est sélectionné à partir des mêmes UPE, cheminement et groupe âge-sexe sur la base de leur disponibilité, c'est-à-dire le premier enregistrement dans le fichier qui satisfait les critères de sélection. Compte tenu de la façon dont les enregistrements sont reçus pour le traitement, il est supposé que la sélection du premier enregistrement disponible est une bonne approximation de la méthode de sélection aléatoire. Il convient de mentionner que, si cette recherche d'enregistrement échoue, les conditions sont relâchées afin d'inclure le groupe âge-sexe voisin. La complétude et la cohérence interne du questionnaire pour les ménages répondants est maintenant assurée grâce à l'utilisation des méthodes d'imputation pertinentes.

J'ai donné un exemple des procédures d'imputation utilisées dans le cas de l'Enquête sur la population active au Canada. Les procédures utilisées dans d'autres enquêtes auprès des ménages s'appuient fortement sur les méthodes appliquées pour l'Enquête sur la population active et l'Enquête sur les finances des consommateurs. Pour cette dernière, plusieurs strates d'imputation sont définies et la méthode de stratification est principalement fondée sur une technique élaborée par Morgan et Sonquist.

8 Conclusion

L'article donne un bref aperçu du concept de non-réponse, de plusieurs sources de non-réponse et de diverses méthodes d'ajustement pour tenir compte de celle-ci. Le résultat des divers moyens de corriger la non-réponse est l'obtention d'un « ensemble de données épuré », c'est-à-dire

un ensemble de valeurs cohérentes qui sont disponibles pour chaque unité dans l'échantillon.

Il existe plusieurs méthodes de correction de la non-réponse, mais l'élaboration d'un fondement méthodologique solide semble faire défaut pour la plupart d'entre elles. L'élaboration d'une théorie intégrée de l'imputation devient de plus en plus importante étant donné le nombre croissant d'enquêtes et les difficultés à obtenir des réponses complètes. Cependant, il sera toujours accordé une importance primordiale au contrôle de l'ampleur de la non-réponse sur le terrain plutôt qu'aux ajustements par diverses techniques pour en tenir compte.

Bibliographie

[1] Ashraf, A. et Macredie, I., "Edit and Imputation in the Labour Force Survey," *Imputation and Editing of Faulty or Missing Survey Data*, 114-119, articles choisis présentés au 1978 American Statistical Association meeting.

[2] Bailar, J.C. III et Bailar, B.A. (1978), "Comparison of Two Procedures For Imputing Missing Survey Values", *Imputation and Editing of Faulty or Missing Survey Data*, 65-75, articles choisis présentés au 1978 American Statistical Association meeting.

[3] Fellegi, I.P. (1964) "Response Variance and its Estimation", *Journal of the American Statistical Association*, Volume 59. 1016-1041.

[4] Ford, Bary, L. (1976) "Missing Data Procedures: A Comparative Study," American Statistical Association, *Proceedings of Social Statistics Section*.

[5] Gower, A.R. "Non-response in the Canadian Labour Force Survey" *Techniques d'enquête* (Statistique Canada) Volume 5, Number 1 (Juin 1979) 29-55.

[6] Hansen, M.H., Horvitz, W.N., et Madow, W.G. (1953), "*Sample Survey Methods and Theory*", Volume II, Théorie, 139-141, John Wiley and Sons, Inc.

[7] Koch, G. (1973), "An Alternative Approach to Multivariate Response Error Model for Sample Survey Data with Applications to Estimators Involving Subclass Means", *Journal of the American Statistical Association*, Volume 68, 906-913.

[8] Kovar, Mary Grace et Robert P. Wright (1973), "An Experiment with Alternative Respondent Rules in NHIS", ***Proceedings of Social Statistics Section***, American Statistical Association.

[9] Madow, William G., rapport non publié sur la non-réponse (1980).

[10] Morgan, J.A. et Sonquist, J.N. (1964), Monograph 35 Survey Research Centre, University of Michigan.

[11] Platek, R. "Some Factors Affecting Nonresponse" ***Techniques d'enquête*** (Statistique Canada) volume 3, numéro 2 (Décembre 1977), 191-214.

[12] Platek, R., Singh, M.P. et Tremblay, V., "Adjustment for Nonresponse in Surveys", ***Survey Sampling and Measurement***, Academic Press, New York, 1978, 157-175.

[13] Platek, R. et Gray, G.B. "Methodology of Adjustments for Nonresponse", article sollicité présenté à la 42 Session of International Statistical Institute, Manila, Philippines, Décembre 1979.

[14] Statistique Canada, "Imputation for Household Survey in Statistics Canada", article soumis par le Canada au meeting of European Statisticians, Genève, Mars 1978.