Techniques d'enquête, juin 1985
Vol. 11, no. 1, pp. 1–15
Statistique Canada

Méthodes de compensation de la non-réponse

R. PLATEK et G.B. GRAY[1]

RÉSUMÉ

La non-réponse à une question ou à un questionnaire est un phénomène presque inévitable dans les enquêtes et les recensements. Plus elle a de l'ampleur, plus elle risque d'influer sur les estimations de l'enquête. Il importe donc de contenir ses effets à toutes les étapes où elle peut fausser les estimations. À des degrés divers, il est possible de minimiser l'ampleur de la non-réponse aux stades de la conception, de l'exécution et du traitement. Les problèmes de non-réponse se reflètent sur les formules d'estimation relatives à diverses statistiques à cause des imputations et des corrections de poids, sans compter les poids attribués dans une enquête aux estimations de moyennes, de totaux ou d'autres statistiques. Les formules peuvent être décomposées en divers éléments comme les erreurs de réponse, l'effet de la correction de poids pour tenir compte de la non-réponse au questionnaire et l'effet de la substitution pour non-réponse. Les auteurs analysent l'effet des diverses étapes (conception, exécution et traitement) sur les éléments des estimations.

MOTS CLÉS: Non-réponse; imputation; estimation.

1. INTRODUCTION

Dans la collecte de données auprès d'unités échantillonnées, il y a toujours un certain nombre d'unités qui négligent de répondre à une question ou au questionnaire malgré tous les efforts faits pour que cela ne se produise pas. Le problème que pose le traitement de la non-réponse et des données manquantes qui en découlent est double. Tout d'abord, il faut déterminer dans quelle mesure on peut réduire l'erreur quadratique moyenne des données d'enquête par des visites de rappel, des lettres de rappel, etc. sans que cela n'accroisse indûment le coût de l'enquête et, ensuite, il faut déterminer le genre d'ajustements requis pour les données manquantes découlant des cas persistants de non-réponse, afin de réduire le biais dû à la non-réponse.

Pour réduire au minimum le taux de non-réponse au questionnaire, les bureaux régionaux ou l'administration centrale des enquêtes doivent tenter de contacter à plusieurs reprises les unités choisies jusqu'à ce qu'une personne responsable puisse répondre au questionnaire d'enquête. De telles mesures sont prises dans le cas des interviews sur place ou des interviews téléphoniques. En ce qui concerne les enquêtes par la poste, ce genre de démarche se traduit par l'expédition répétée du questionnaire d'enquête aux unités non répondantes. Cela peut parfois donner lieu à un suivi téléphonique ou à un suivi à domicile. Bien qu'il subsistera toujours des cas de non-réponse, on devrait faire tous les efforts possibles pour réduire leur nombre au minimum. Ainsi, il y aura toujours des non-répondants qui résisteront aux demandes répétées des intervieweurs, peu importe les mesures que ceux-ci prendront. Il est donc nécessaire d'appliquer une méthode d'imputation pour tenir compte des données manquantes. Dans le présent document, nous nous penchons sur les problèmes que soulève le contrôle de la non-réponse à l'étape de la conception et de l'exécution d'une enquête, puis nous analysons certaines méthodes de compensation de la non-réponse à l'étape du traitement. Dans cette deuxième partie, nous examinons l'applicabilité de ces méthodes de même que les questions d'ordre pratique et méthodologique qui s'y rattachent.

[1] R. Platek et G.B. Gray, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada, 4e étage, Imm. Jean-Talon, Parc Tunney, Ottawa (Ontario), Canada K1A 0T6.

La non-réponse à une question est un problème souvent plus difficile à traiter que la non-réponse au questionnaire, sur laquelle a porté surtout notre propos jusqu'à maintenant. Les deux principaux facteurs qui peuvent réduire le taux de non-réponse à une question sont un questionnaire bien conçu et des intervieweurs très compétents, c'est-à-dire recrutés et formés selon les critères appropriés. Un questionnaire mal conçu peut amener un intervieweur ou un répondant à s'interroger sur la façon exacte de remplir ce questionnaire. La non-réponse à une question est donc possible malgré la bonne volonté de l'intervieweur ou du répondant. En outre, il se peut que des répondants consentent à répondre à certaines questions seulement. Quelle que soit la cause des réponses manquantes, le problème de leur substitution reste entier. En règle générale, un organisme d'enquête est peu disposé à rejeter les renseignements qu'il vient de recueillir à moins, bien sûr, que les réponses fournies aux principales questions ne semblent tout à fait erronées ou illogiques. On a donc recours habituellement à des méthodes d'imputation pour les réponses manquantes tout en conservant les réponses déjà obtenues.

Une enquête ou un recensement doivent produire diverses statistiques permettant d'expliquer des phénomènes sociaux, de déterminer des programmes socio-économiques, etc. Ces statistiques comprennent des moyennes, des totaux, des ratios, des distributions, des percentiles et des graphiques. On suppose que ces statistiques sont fondées sur un univers de N unités appartenant à la population visée, où N peut être connu ou non.

Il est possible de démontrer que toutes ces statistiques peuvent être exprimées en fonction de totaux ou de comptages. Ainsi, le reste de cet article est consacré à l'analyse des données manquantes dans la mesure où celles-ci influent sur les estimations des totaux et des comptages dans les enquêtes. Certaines observations porteront aussi sur les recensements.

2. FORMULE D'ESTIMATION

Lorsqu'il y a non-réponse à une question ou au questionnaire, le total de la caractéristique y peut être estimé au moyen de l'expression générale suivante:

$$\tilde{Y} = \sum_{i=1}^{N} t_i \pi_i^{-1} \left\{ \delta_i \left[\delta_{iy} y_i + (1 - \delta_{iy}) z_{iy} \right] + (1 - \delta_i) z_i \right\}, \text{ où} \qquad (2.1)$$

t_i = 1 ou 0, selon que l'unité i est choisie ou non;
π_i = probabilité de sélection de l'unité;
δ_i = 1 ou 0, selon que l'unité i répond ou non;
δ_{iy} = 1 ou 0, selon que l'unité répondante i répond ou non à la question ou à la caractéristique y;
y_i = réponse observée pour la caractéristique y lorsque $\delta_{iy} = \delta_i = 1$; y_i peut être ou non égal à Y_i, la valeur réelle;
z_{iy} = valeur imputée pour la non-réponse à une question, lorsque $\delta_i = 1$, $\delta_{iy} = 0$.
z_i = valeur imputée pour la non-réponse au questionnaire, lorsque $\delta_i = 0$.

L'estimation ci-dessus peut se rapporter à une catégorie a d'unités; dans ce cas, la variable auxiliaire β_{ia} insérée après π_i^{-1} dans l'équation, prend la valeur 1 ou la valeur 0 selon que l'unité i appartient ou non à la catégorie a (par exemple, catégorie d'âge selon le sexe).

Pour ce qui a trait à la non-réponse à une question, z_{iy} est presque toujours une *valeur imputée* explicite pour les données manquantes. On peut obtenir une valeur imputée par la méthode *hot deck*, c'est-à-dire qu'on applique à une question y, laissée sans réponse par l'unité i, une réponse fournie par une autre unité qui ressemble le plus possible à l'unité i, à

l'aide d'une table de décision. On peut aussi obtenir une valeur imputée à partir de données fournies antérieurement par la même unité au cours d'une enquête ou d'un recensement ou à partir de données administratives, si de telles données existent. On peut appliquer des méthodes de régression, la déduction logique et bien d'autres encore. Des erreurs systématiques sont parfois attribuables à de mauvaises opérations de préposés au codage ou à la perforation. Dans ce cas, on peut vouloir remplacer les codes par des valeurs logiques se rapportant à d'autres données contenues dans le questionnaire, au lieu de procéder à une imputation. Quoi qu'il en soit, le but ultime est d'obtenir une valeur imputée ou un code modifié qui se rapproche le plus de la valeur réelle Y_i. Dans les enquêtes permanentes, où les caractéristiques étudiées évoluent peu sur une longue période (par exemple, l'emploi dans certaines industries et certaines professions), on peut considérer les réponses fournies aux enquêtes antérieures par un répondant comme presque aussi valables que celles fournies au cours de l'enquête courante, à plus forte raison lorsque les périodes de référence des deux enquêtes ne sont pas trop éloignées. Il peut en être de même pour les données d'enquête recueillies à un an d'intervalle, comme celles ayant trait à des caractéristiques saisonnières (par exemple, l'industrie de la pêche). Dans les cas de non-réponse au questionnaire, on peut parfois imputer des données d'enquêtes antérieures si le non-répondant en question a participé à ces enquêtes et qu'il présente des caractéristiques stables.

Pour les cas de non-réponse au questionnaire, l'imputation se fait généralement par une correction de poids au moyen de l'inverse du taux de réponse enregistré dans une cellule ou une région. L'estimation du total est donc définie comme suit:

$$\tilde{Y} = \sum_{i=1}^{N} t_i \pi_i^{-1} (wa)_i \delta_i \left[\delta_{iy} y_i + (1 - \delta_{iy}) z_{iy}\right] \tag{2.2}$$

où $(wa)_i$ est le facteur de correction du poids pour l'unité i, qui vise à rééquilibrer l'échantillon amenuisé par la non-réponse au questionnaire. Dans l'expression ci-dessus, on suppose que toutes les questions laissées sans réponse par l'unité répondante i ($\delta_{iy} = 0$) ont déjà fait l'objet d'une imputation au moyen de z_{iy}.

En supposant qu'il puisse manquer des données, on peut obtenir des estimations de la fonction cumulative de distribution (f.c.d.) dans l'échantillon en remplaçant la valeur observée y par la variable auxiliaire $c(y_i, Y)$, celle-ci étant égale à 1 ou à 0 selon que y est inférieur ou supérieur à Y; on fait de même pour z_{iy} et z_i. Les f.c.d. correspondant à (2.1) et (2.2) sont estimées respectivement par les équations (2.3) et (2.4).

$$\tilde{F}(Y) = \frac{1}{\hat{N}} \sum_{i=1}^{N} t_i \pi_i^{-1} \left\{\delta_i \left[\delta_{iy} c(y_i, Y) + (1 - \delta_{iy}) c(z_{iy}, Y)\right] + (1 - \delta_i) c(z_i, Y)\right\} \tag{2.3}$$

où $\hat{N} = \sum_{i=1}^{N} t_i \pi_i^{-1}$ représente le nombre réel ou estimé d'unités dans l'univers, Ainsi, suivant la base de sondage, le plan de sondage et le listage des unités, $\hat{N}$ peut ou non égaler N.

$$\tilde{F}(Y) = \frac{1}{\hat{N}} \sum_{i=1}^{N} t_i \pi_i^{-1} (wa)_i \delta_i \left[\delta_{iy} c(y_i, Y) + (1 - \delta_{iy}) c(z_{iy}, Y)\right] \tag{2.4}$$

Tandis que les valeurs de $\tilde{Y}$ définies en (2.1) et (2.2) sont identiques, peu importe que l'imputation pour la non-réponse au questionnaire se fasse par l'application de moyennes relatives aux répondants ou par la correction de poids, les estimations $\tilde{F}(Y)$ des f.c.d., définies en (2.3) et (2.4) ne le sont pas. Lorsqu'on impute en appliquant les moyennes relatives aux répondants (moyenne globale ou moyennes s'appliquant à des cellules de correction définies pour la compensation de la non-réponse), comme dans l'équation (2.1) ou (2.3), on se trouve à amplifier ces valeurs moyennes dans la f.c.d. estimée, ce qui a pour effet de fausser la

forme de la f.c.d. dans la population. En utilisant le facteur de correction de poids $(wa)_i$ pour gonfler le poids t_i^{-1} de l'échantillon dans l'équation (2.4), on élimine l'effet d'amplification et, du même coup, on obtient une estimation différente mais plus réaliste de la f.c.d.

Lorsqu'il n'y a aucun cas de non-réponse à une question et au questionnaire, les estimations (2.1) et (2.2) se réduisent à l'estimation d'Horvitz-Thompson (1952) qui est sans biais, exception faite des erreurs de réponse. Lorsqu'il manque des données et qu'on procède à une imputation, les estimations (2.1) et (2.2) risquent toutefois d'être biaisées pour des raisons autres que les erreurs de réponse, à moins que les valeurs de z_{iy} et de z_i se rapprochent de celles de y lorsqu'il faut procéder à une imputation pour la non-réponse à une question ou au questionnaire.

Dans la section suivante, les estimations (2.1) et (2.2) sont analysées en fonction de l'erreur de réponse, de l'erreur d'imputation due à la non-réponse à une question, de l'erreur d'imputation due à la non-réponse au questionnaire et de l'effet des facteurs de correction de poids supérieurs à 1.

3. ÉLÉMENTS DE L'ESTIMATION

L'estimation $\tilde{Y}$ définie en (2.1) ou (2.2) peut être décomposée en 5 éléments, dont le premier est l'estimation d'Horvitz-Thompson, laquelle, illustrée au Tableau 1, utilise les valeurs réelles de la caractéristique. La f.c.d. estimée $\tilde{F}(Y)$, définie en (2.4), peut être décomposée de la même façon; cependant, nous n'avons pas jugé bon de traiter ce sujet dans le présent document.

Lorsque le facteur de correction de poids $(wa)_i = 1$, le dernier élément disparaît et la somme des 4 autres ((3.1) à (3.4)), correspond à l'estimation définie en (2.1). Lorsque la non-réponse au questionnaire est compensée par un facteur de correction de poids supérieur à 1, la valeur z_i n'est pas imputée directement pour la valeur manquante et elle est considérée comme nulle dans l'élément (3.4). Dans ce cas, la somme des 5 éléments correspond à l'estimation définie en (2.2) et l'effet négatif de la non-réponse au questionnaire en (3.4) est compensé par l'effet positif de la correction de poids en (3.5).

Tableau 1:
Éléments de l'estimation $\tilde{Y}$

$\tilde{Y} = \sum_{i=1}^{N} t_i \pi_i^{-1} Y_i$	..	Estimation sans biais fondée sur un taux de réponse de 100% et des valeurs réelles	(3.1)
$+ \sum_{i=1}^{N} t_i \pi_i^{-1} (y_i - Y_i)$	..	effet de l'erreur de réponse	(3.2)
$+ \sum_{i=1}^{N} t_i \pi_i^{-1} \delta_i (1 - \delta_{iy})(z_{iy} - y_i)$	..	effet de la non-réponse à une question	(3.3)
$+ \sum_{i=1}^{N} t_i \pi_i^{-1} (1 - \delta_i)(z_i - y_i)$	..	effect de la non-réponse au questionnaire	(3.4)
$+ \sum_{i=1}^{N} t_i \pi_i^{-1} [(wa)_i - 1] \delta_i [\delta_{iy} y_i + (1 - \delta_{iy}) z_{iy}]$	..	effet de la correction de poids pour la non-réponse au questionnaire	(3.5)

a) Erreur de réponse.

La somme des deux premiers éléments de l'estimation Y (voir 3.1 et 3.2) est égale à l'estimation d'Horvitz-Thompson qui sert à estimer le total lorsque le taux de non-réponse est nul. La réponse y_i fournie par l'unité peut différer de la valeur réelle Y, ce qui entraîne une erreur de réponse pour l'unité i. À défaut de pouvoir l'éliminer, on peut diminuer l'erreur de réponse de deux façons seulement: par la formation adéquate des intervieweurs et par un questionnaire bien conçu, dénué de tout élément qui risquerait de déconcerter l'intervieweur ou le répondant et où les caractéristiques sont clairement définies et les questions formulées en termes explicites.

Lorsque, pour un taux de non-réponse nul, les erreurs de réponse pondérées échantillonnées de (3.2) ne s'annulent pas, l'estimation du total (Y) est entachée d'une erreur de réponse et, si l'on considère l'espérance mathématique E_1 et E_3 (voir Platek et Gray (1983)) pour tous les échantillons et les réponses possibles, on découvre que cette estimation comporte un biais dû à la réponse (BR) et une variance de réponse, sans compter la variance d'échantillonnage (VE). La variance de réponse peut être divisée en variance de réponse simple (VRS) et en variance de réponse corrélée (VRC).

Le biais dû à la réponse et tous les éléments de variance (VE, VRS et VRC) se rapportant à l'estimation ci-dessus sont calculés dans Platek et Gray (1983), paragraphe 2.2, p. 257-258.

On analyse normalement les erreurs de réponse au moyen d'un programme de réinterview de comparaison, en vertu duquel on interviewe à nouveau un sous-échantillon des unités répondantes et on compare, pour la même période de référence de l'échantillon, les données obtenues avec celles de la première interview afin de déterminer laquelle des deux interviews produit les réponses les plus exactes. Les programmes de réinterview de comparaison sont prévus dans l'enquête sur la population active du Canada et la U.S. Current Population Survey (CPS), deux enquêtes mensuelles comparables visant à mesurer le chômage, l'emploi, etc.

Par exemple, Poterba et Summers (1984) présentent au Tableau 2 les résultats d'une réinterview de comparaison menée en mai 1976 dans le cadre de la CPS; le sous-échantillon était composé de 3 329 hommes et de 3 750 femmes. En comparant les réponses fournies par ce sous-échantillon avec celles fournies à la première interview, on connaît la situation réelle d'un individu vis-à-vis de l'activité, de telle sorte qu'il est possible de savoir si cet individu a répondu correctement à la question lors de la première enquête de la CPS. Il est donc possible de connaître le nombre d'unités du sous-échantillon qui sont effectivement occupées et qui avaient été classées dans l'une ou l'autre des trois catégories (occupé, en chômage, inactif) dans l'enquête originale. À l'aide des trois valeurs obtenues, il est possible d'estimer, selon l'activité réelle, la proportion (ou la probabilité) de réponses justes et erronées (voir le tableau suivant).

Ainsi, d'après les résultats de la réinterview de comparaison, environ 87.20% des hommes effectivement en chômage ont été classés comme tels lors de la première enquête, tandis que 12.80% d'entre eux (soit 0.0474 + 0.0806) ont déclaré faussement être occupés ou inactifs. En conséquence, si y désigne la caractéristique *en chômage*, c'est-à-dire que $Y = 1$ lorsque le répondant i est une personne de sexe masculin qui est effectivement en chômage, la probabilité que $y_i = 1$ est de 0.8720 alors que la probabilité que $y_i = 0$ est de 0.1280.

Dans l'enquête sur la population active du Canada, l'échantillon visé par la réinterview de comparaison pour la période de janvier à novembre 1984 comprenait 7 148 personnes. Le Tableau 3 ci-dessous donne la probabilité que ces répondants aient déclaré être occupés, en chômage ou inactifs dans l'EPA régulière suivant leur activité réelle révélée par la réinterview de 1984.

Ainsi, la probabilité qu'une personne effectivement sans emploi soit classée comme telle dans l'EPA est estimée à 0.8691, ce qui est sensiblement comparable à la probabilité correspondante dans la CPS (0.8602). De même, la probabilité qu'une personne effectivement

Tableau 2

Probabilité qu'une personne se soit déclarée occupée, en chômage ou inactive dans la CPS régulière, selon la situation réelle vis-à-vis de l'activité fondée sur les résultats de la réinterview de mai 1976.

	Activité déclarée dans l'enquête régulière de la CPS		
Activité réelle	Occupé	En chômage	Inactif
Total[1]			
Occupé	0.9905	0.0016	0.0079
En chômage	0.0356	0.8602	0.1041
Inactif	0.0053	0.0025	0.9923
Hommes[2]			
Occupé	0.9922	0.0013	0.0065
En chômage	0.0474	0.8720	0.0806
Inactif	0.0062	0.0048	0.9890
Femmes[3]			
Occupé	0.9892	0.0019	0.0089
En chômage	0.0194	0.8442	0.1363
Inactif	0.0049	0.0015	0.9936

[1] Taille d'échantillon = 7 079
[2] Taille d'échantillon = 3 329
[3] Taille d'échantillon = 3 750

Source: Données établies à partir du "General Labour Force Status in the CPS Reinterview menée selon l'activité déclarée lors de l'interview originale. Les deux sexes. Total. Après comparaison. Mai 1976, Bureau of the Census (données non publiées).

Tableau 3

Nombre de personnes et probabilité de déclaration de l'activité (entre parenthèses) selon la vraie caractéristique (janvier-novembre 1984)

Charactéristique réelle (Réinterview de comparaison)	EPA régulière			
	Occupé	En chômage	Inactif	Total
Occupé	4 082 (0.9831)	19 (0.0046)	51 (0.0123)	4 152
En chômage	8 (0.0122)	571 (0.8691)	78 (0.1187)	657
Inactif	28 (0.0120)	30 (0.0128)	2 281 (0.9752)	2 339
Total	4 118	620	2 410	7 148

occupée ou inactive soit classée comme telle dans l'EPA est estimée à 0.9831 et à 0.9752 respectivement pour 1984, ce qui est légèrement inférieur aux probabilités correspondantes dans la CPS (0.9905 et 0.9923). On ne peut expliquer ces écarts pour le moment. De toute façon, les erreurs de réponse sont vraisemblablement plus importantes au niveau national qu'au niveau des petites régions. Par exemple, le biais dû à la réponse au niveau national peut être supérieur à l'erreur due à l'échantillonnage, tandis que l'estimation pour une petite région peut afficher un biais qui est comparable, en pourcentage, à celui obtenu à l'échelle nationale, mais beaucoup moins fort que l'erreur d'échantillonnage.

b) Non-réponse à une question et erreur d'imputation

Le troisième élément (3.3) de l'estimation $\tilde{Y}$ au Tableau 1 définit l'écart entre $\tilde{Y}$ et l'estimation voulue $\hat{Y}$ qui est dû à l'imputation faite pour tenir compte de la non-réponse à une question, lorsque la valeur imputée $z_{iy} = y_i$ et que les écarts pondérés échantillonnés $(z_{iy} - y_i)$ pour l'ensemble des unités échantillonnées qui font l'objet d'imputation pour la non-réponse à une question ne s'annulent pas. On considère qu'il y a non-réponse à une question lorsqu'une unité refuse de répondre à certaines questions du questionnaire ou que l'intervieweur ou le répondant (dans le cas de l'autodénombrement) laisse par mégarde des questions sans réponse. La deuxième cause de non-réponse à une question est identique à celle qui est à l'origine des erreurs de réponse, c'est-à-dire un questionnaire comportant des questions complexes et des définitions ambiguës ou un questionnaire confus où, suivant les réponses aux questions-filtres, on peut difficilement s'y retrouver.

Lorsqu'on a des cas de non-réponse à une question, on peut adopter une stratégie d'imputation comme celles qui ont été décrites précédemment; cette stratégie se traduit presque toujours par une substitution explicite. Dans l'analyse de microdonnées, il est essentiel d'obtenir une valeur z_{iy} qui se rapproche le plus possible de la valeur réelle Y_i ou, du moins, de ce qu'aurait été la valeur observée y_i si l'unité avait répondu à la question qui définit la caractéristique y. Il n'y a malheureusement aucun moyen de connaître le degré de concordance de z_{iy} et de y_i, si ce n'est par une réinterview de l'unité ou une analyse de sources externes ou de données d'enquêtes antérieures (qui peuvent être difficiles à obtenir). Un autre inconvénient de la non-réponse à une question et, par conséquent, de l'imputation est le faux sentiment de sécurité que celle-ci procure à l'utilisateur de données qui ignore peut-être qu'une valeur z_{iy} a été substituée à une réponse authentique au niveau des microdonnées. La valeur imputée z_{iy} tendra à s'écarter plus, dans un sens comme dans l'autre, de la valeur réelle Y_i, que l'erreur de réponse probable y_i si l'unité en question répond à la caractéristique. Il peut y avoir des exceptions. Malheureusement, il est normalement impossible de déterminer au niveau des microdonnées si z_{iy} constitue une valeur moins précise que y_i. Même s'il peut arriver que l'erreur d'imputation soit inférieure à l'erreur de réponse probable, cela réduira davantage la qualité des statistiques publiées à cause de la présence d'autres éléments de variance.

Les cas de non-réponse à une question et les erreurs de réponse dans l'EPA sont souvent repérés grâce au programme mensuel Module de contrôle sur le terrain qui permet d'analyser les questionnaires qui ont été rejetés à la vérification pour une ou plusieurs questions. En cours d'analyse, toutefois, on établit rarement la distinction entre les erreurs de réponse et les cas de non-réponse à une question, si l'on ne se donne pas la peine d'examiner à fond chaque questionnaire. L'erreur la plus courante est l'inscription d'un mauvais code plutôt que l'absence d'une réponse à une question. De nombreuses questions sont divisées en cinq ou six sous-questions, et il peut arriver qu'une erreur de codage soit considérée comme un cas de non-réponse à une sous-question et comme une erreur de réponse à une autre sous-question de la même question. Comme l'analyse du module du contrôle sur le terrain porte

sur les questions et non sur les sous-questions, il est difficile de définir de façon catégorique le taux de divergence des réponses. Celui-ci a trait à un sous-ensemble de questionnaires pour lesquels une question particulière, disons la question n° q, est pertinente selon des questions-filtres et des tables de décisions. Supposons que dans un échantillon de m questionnaires la question n° q soit pertinente pour $m_q \leq m$ questionnaires. Le taux de divergence est donc la proportion des m_q questionnaires qui ont été rejetés à la vérification à cause de la non-réponse à une question ou d'une erreur de codage. Il reste à savoir si le sous-ensemble m_q devrait inclure les questionnaires qui comptent des espaces remplis par erreur, ceux qui comptent des espaces laissés en blanc par erreur ou simplement ceux où les réponses sont codées correctement ou incorrectement. Malgré cet aspect ambigu de la définition, l'analyse des taux de divergence pour une cinquantaine de questions, pour l'année civile 1984, devrait indiquer une limite supérieure de l'erreur partielle dans les estimations des statistiques fondées sur les questions. Le Tableau 4 ci-dessous donne certains taux de divergence pour les questions définies au Tableau 4a, pour 1984.

Ainsi, pour une question aussi simple que la question 10 (le répondant a-t-il travaillé la semaine dernière, oui ou non), le taux de divergence n'est que de 0.2%, soit beaucoup moindre que l'erreur type à l'échelle nationale. Pour des questions plus complexes comme 12, 36, 41, 54 et 77, le taux de divergence se situe en moyenne au-dessus de 10% et fluctue dans un intervalle de 2 à 6 points de part et d'autre de la moyenne au cours de l'année. On élimine les divergences par des méthodes du genre *hot deck*, l'utilisation de réponses fournies à la dernière enquête (si celles-ci sont disponibles) ou par la déduction logique appliquée à d'autres données du questionnaire. Par conséquent, il est souvent possible qu'une divergence soit interprétée comme le résultat d'une erreur de réponse et non d'une erreur d'imputation, de sorte qu'on devrait considérer les taux de divergence comme une limite supérieure de la fréquence globale des erreurs d'imputation dans les questions.

c) Non-réponse au questionnaire et correction de poids

En ce qui a trait à la non-réponse au questionnaire, les deux éléments de Y définis par (3.4) et (3.5) doivent être analysés ensemble puisque la non-réponse au questionnaire est normalement compensée par une correction de poids $((wa)_i)$ plutôt que par une substitution directe de z_i à une valeur d'unité manquante. On calcule normalement les facteurs de correction de poids au moyen de l'inverse des taux s'appliquant à des cellules de correction; celles-ci se divisent en deux grandes catégories: les régions de compensation et les classes de pondération. Les premières sont souvent des régions géographiques liées au plan de sondage, comme une strate, une unité primaire d'échantillonnage, une grappe, un groupe de strates ou même un échantillon complet. Les classes de pondération sont définies par des strates formées a posteriori (c'est-à-dire, créées après l'échantillonnage) d'après des renseignements dont pouvaient disposer les répondants et les non-répondants compris dans l'échantillon. Pour ce qui a trait aux non-répondants, les renseignements peuvent être obtenus des unités qui ont répondu partiellement au questionnaire et dont on connaît certaines caractéristiques, peu importe que la caractéristique estimée se trouve ou non parmi celles-ci. On peut, par ailleurs, consulter des sources externes ayant trait aux non-répondants. On peut calculer l'inverse des taux de réponse soit pour des régions de compensation ou des classes de pondération et les utiliser comme facteurs de correction de poids pour compenser les cas de non-réponse dans les cellules.

Il existe plusieurs genres de facteurs de correction de poids pour compenser la non-réponse au questionnaire, le plus courant étant le taux de réponse inverse qui est défini comme le rapport de la taille de l'échantillon à la taille de l'échantillon répondant dans une cellule de

Tableau 4
Taux moyen de divergence par question (définie au Tableau 4a)

Question	Taux moyen de divergence	Intervalle de variation des taux en 1984 (minimum-maximum)
10	0.2%	0.2% chaque mois
12	12.3%	10.4% à 14.3%
14	6.7%	5.7% à 8.4%
16	0.4%	0.3% à 0.5%
17	6.6%	2.0% à 9.9%
30	0.4%	0.3% à 0.5%
32	7.0%	3.0% à 11.6%
33	4.3%	1.8% à 6.0%
36	10.6%	8.1% à 12.7%
40	4.1%	1.5% à 6.8%
41	12.1%	6.2% à 19.7%
54	10.1%	7.9% à 12.1%
76	<0.1%	0.0% à 0.1%
77	15.0%	11.8% à 17.3%

Source: Rapport interne présenté à P.D. Ghangurde par Karen Switzer, 4 mars 1985, "Some Findings on the Field Edit Module (FEM) Reports from 1984".

Tableau 4a
Libellé des questions

10)	La semaine dernière, (le répondant) a-t-il(elle) travaillé à un emploi ou à une entreprise peu importe le nombre d'heures? Oui ou Non.
12)	Si la réponse à la question 11 "Avait-il(elle) plus d'un emploi la semaine dernière?" est oui, était-ce dû à un changement d'employeur? Oui ou Non.
14)	Pourquoi ... travaille-t-il(elle) habituellement moins de 30 heures par semaine?, si la réponse à la question 13) indique un nombre d'heures travaillées inférieur à 30.
16)	La semaine dernière, combien d'heures ... a-t-il(elle) été absent(e) du travail pour une raison quelconque (jour férié, vacances, maladie, conflit de travail, etc.)? (Ne pas oublier d'inscrire 00 s'il y a lieu).
17)	Quelle était la raison principale de cette absence? (10 codes possibles)
30)	La semaine dernière, ... avait-il(elle) un emploi ou une entreprise auxquel il(elle) n'a pas travaillé? Oui ou Non.
32)	A partir de la fin de semaine dernière, dans combien de semaines ... doit-il(elle) commencer à travailler à son nouvel emploi? (Si on a répondu Oui à la question 31), "La semaine, dernière, ... avait-il(elle) un emploi devant commencer à une date future déterminée?")
33)	Pourquoi ... s'est-il(elle) absenté(e) du travail la semaine dernière? (8 codes possibles)
36)	Même que 14), sauf qu'elle concerne les personnes en chômage au lieu des personnes occupées.
40)	Au cours des 4 dernières semaines, ... s'est-il(elle) cherché(e) un autre emploi? Oui ou Non.
41)	Qu'a fait ... au cours des quatre dernières semaines pour se trouver un autre emploi? (8 codes possibles, 1 à 3 codes différents dans 1, 2 ou 3 cases).
54)	"Quelle est la raison principale pour laquelle ... a laissé cet emploi?" Neuf codes possibles si l'on a répondu oui à la question 50) " ... a-t-il(elle) déjà travaillé à un emploi ou à une entreprise?" (dans le cas des personnes souffrant d'incapacité permanente) et si l'on a répondu aux questions 51) à 53) qui concernent la date du dernier emploi et le statut d'employé (à temps plein ou à temps partiel). On omet la question 54) si la date du dernier emploi se situe avant la date pré-imprimée au n 52).
76/77)	Indiquent, relativement à l'emploi principal (76) et à un autre emploi (77), quel est le statut du travailleur et si c'est le même qu'au mois précédent.

correction. Par conséquent, si la population d'une cellule est composée de N unités et que cette cellule est représentée par un échantillon de n unités, où :

$n_b = \sum_{i\epsilon b} t_i$ la taille de l'échantillon de la cellule b, cette valeur pouvant ou non être une constante selon la définition de la cellule;

$\hat{N}_b = \sum_{i\epsilon b} \pi_i t_i$, une estimation de la taille de la population de la cellule b; N_b ne devrait normalement pas être connue, sauf s'il s'agit d'un recensement;

$m_b = \sum_{i\epsilon b} t_i \delta_i$ = nombre d'unités répondantes dans la cellule b, c'est-à-dire la taille de l'échantillon répondant,

alors, $(wa)_i = n_b/m_b$ lorsque i appartient à la cellule de correction b. (3.6)

Avant de définir d'autres facteurs de correction de poids, nous allons analyser plus à fond ce facteur fréquemment utilisé qu'est le taux de réponse inverse non pondéré dans une cellule, décrit en (3.6). Avec $(wa)_i = n_b/m_b$, l'estimation du total définie en (2.2) peut être réécrite comme un cas particulier de (2.1), où z_i est défini comme suit:

$$z_i = \hat{T}_b/\pi_i^{-1} m_b, \tag{3.7}$$

où $\hat{T}_b = \sum_{i\epsilon b} \pi_i^{-1} t_i \delta_i [\delta_{iy} y_i + (1 - \delta_{iy}) z_{iy}]$, le total pondéré de l'échantillon d'unités répondantes dans la cellule b. Lorsqu'une cellule présente des poids d'échantillons égaux, la valeur imputée z_i se réduit à la valeur moyenne des m_b répondants dans la cellule. Si l'on substitue l'expression (3.7) à z_i dans l'équation (2.1), on peut montrer que l'estimation est similaire à celle obtenue dans l'équation (2.2) si l'on considère que $(wa)_i = (n_b/m_b)$. On peut, par conséquent, considérer l'imputation pour la non-réponse au questionnaire comme une substitution de $z_i = \hat{T}_b/(\pi_i^{-1} m_b)$ dans l'équation (2.1) ou comme une correction des poids de l'échantillon dans l'équation (2.2) au moyen du facteur $(wa)_i = n_b/m_b$. En ce qui concerne la correction de poids, on poserait $z_i = 0$ dans l'expression (3.4), ce qui aurait pour effet de maintenir les cinq éléments de $\tilde{Y}$. Par ailleur, on pourrait utiliser la valeur imputée z_i telle qu'elle est définie dans l'équation (2.1) et on poserait alors $(wa)_i = 1$ dans l'expression (3.5), ce qui rendrait nul le dernier élément de $\tilde{Y}$. Par conséquent, si l'on veut analyser l'effet de la correction de poids ($(wa)_i > 1$), il faut considérer dans une même bloc l'élément négatif (3.4) et l'élément positif (3.5); mais si l'on veut analyser l'effet de la valeur imputée implicite z_i, définie en (3.7), seul l'élément (3.4) peut être pris en considération.

Le facteur de correction de poids (n_b/m_b) est utilisé dans l'EPA. Les cellules de correction de l'EPA sont des U.P.É. liées au plan de sondage dans les unités non autoreprésentatives (UNAR) et des strates (sous-unités) d'îlots urbains contigus dans les unités autoreprésentatives (UAR). Le tableau 5 donne, par région et par genre d'unité, le nombre de cellules, la moyenne non pondérée des facteurs de correction de poids et leur distribution de fréquence dans les intervalles 1-1.01, 1.01-1.02, ..., 1.10 et plus pour l'enquête de janvier 1983.

Le facteur de correction de poids moyen pour le Canada (1.0348) est inférieur à ce qu'on s'attendrait d'obtenir avec un taux de non-réponse d'environ 5%. Cette moyenne apparemment faible s'explique par le fait que certains non-répondants, qui ont toutefois répondu au questionnaire du mois précédent, sont considérés comme des répondants aux fins du calcul du taux de réponse inverse. Cette mesure vise de 20 à 30% des non-répondants à chaque mois.

Tableau 5
Nombre de cellules de correction, et moyenne et distribution de fréquence des facteurs de correction de poids selon la région et le genre d'unité. Janvier 1983

Région	Genre d'unité	Nombre de cellules	Moyenne $(wa)_i$	Nombre de cellules dans les intervalles $(wa)_i$										
				1-1.01	1.01-1.02	1.02-1.03	1.03-1.04	1.04-1.05	1.05-1.06	1.06-1.07	1.07-1.08	1.08-1.09	1.09-1.10	1.10+
Atl.	UANR	254	1.0250	143	6	22	21	13	13	9	7	8	2	10
Atl.	UAR	123	1.0246	58	5	11	15	14	4	3	6	4	1	2
Qué.	UNAR	126	1.0550	72	2	8	10	10	6	8	6	0	1	3
Qué.	UAR	185	1.0265	106	0	7	8	23	11	4	5	7	3	11
Ont.	UNAR	120	1.0333	58	1	10	11	11	8	4	2	2	2	11
Ont.	UAR	252	1.0416	116	1	13	24	21	16	9	9	8	10	25
Pr.	UNAR	328	1.0348	167	5	17	22	23	24	15	12	10	8	25
Pr.	UAR	149	1.0306	40	23	23	20	13	8	7	3	5	4	3
C.B.	UNAR	85	1.0468	38	3	7	8	8	2	5	1	1	1	11
C.B..	UAR	119	1.0412	46	4	7	15	10	7	7	7	3	3	10
Can.	UNAR	913	1.0358	478	17	64	72	65	53	41	28	21	14	60
Can.	UAR	828	1.0337	366	33	61	82	81	46	30	30	27	21	51
Canada		1 741	1.0348	844	50	125	154	146	99	71	58	48	35	111

Si on ne connaît pas les caractéristiques des non-répondants, il n'est pas possible de déterminer avec exactitude la valeur maximale que devrait avoir le facteur de correction de poids si l'on veut éviter un biais excessif et une augmentation de la variance attribuable à une taille réelle d'échantillon moindre. Si on fixe arbitrairement cette valeur maximale à 1.05 pour l'EPA (niveau que les spécialistes d'enquête posent parfois par hypothèse), on observe qu'environ le quart des unités de compensation au Canada (441 sur 1741) avaient un facteur de correction de poids de 1.05 ou plus en janvier 1983. Dans beaucoup d'autres enquêtes comme celles portant sur les revenus et les dépenses, le taux de non-réponse est généralement plus élevé et les facteurs de correction de poids de presque toutes les cellules dépasseraient vraisemblablement le seuil critique si celui-ci était maintenu par hypothèse à 1.05.

Il existe d'autres genres de facteurs de correction de poids dans les cellules. Par exemple, on pourrait exclure de la cellule b, telle qu'elle est définie plus haut, les unités qui n'ont pas répondu à toutes les questions du questionnaire. Supposons que la cellule b compte m_{bQ}, unités qui ont répondu à toutes les questions du questionnaire. Pour les $(m_b - m_{bQ})$ unités répondantes de la cellule qui n'ont pas répondu à toutes les questions, le facteur de correction de poids $(wa)_i = 1$, et pour les m_{bQ} unités qui ont répondu à toutes les questions, le facteur de correction de poids est défini par:

$$(wa)_i = [n_b - (m_b - m_{bQ})] / m_{bQ}, \text{ which exceeds } n_b/m_b. \tag{3.7a}$$

Dans ce qui suit, nous tentons de démontrer qu'il est préférable de ne pas appliquer de facteur de correction de poids, c'est-à-dire, $(wa)_i = 1$, aux unités de la cellule qui n'ont pas répondu à toutes les questions du questionnaire mais d'appliquer un facteur de correction supérieur à n_b/m_b (3.7a) aux unités qui ont répondu à toutes les questions. Les question-

naires qui ne sont pas entièrement remplis renferment probablement des erreurs de réponse et des erreurs d'imputation tandis que les questionnaires entièrement remplis ne renferment que des erreurs de réponse. De plus, en appliquant un facteur de correction de poids plus grand aux unités qui ont répondu à toutes les questions, on peut obtenir des estimations entachées d'une erreur quadratique moyenne plus faible que si l'on appliquait le même facteur de correction de poids à toutes les unités répondantes de la cellule. À notre connaissance, les facteurs de correction de poids que nous venons de décrire n'ont pas été utilisés jusqu'à maintenant, mais il serait peut-être bon que l'on envisage de le faire s'il est démontré que la réduction du biais compense l'augmentation de variance causée par l'existence de poids différents.

En ce qui concerne les unités ayant des probabilités d'échantillonnage différentes, il existe un facteur de correction de poids fondé sur l'échantillon pondéré et les unités répondantes d'une cellule plutôt que sur les échantillons non pondérés. Dans ce cas,

$$(wa)_i = \hat{N}_b / \hat{M}_b, \tag{3.8}$$

où $\hat{M}_b = \sum_{i \epsilon b} \pi_i^{-1} t_i \delta_i$ est le nombre pondéré d'unités répondantes dans la cellule b. Pour ce qui a trait aux facteurs de correction de poids définis en (3.7a), que l'on applique uniquement aux unités qui ont répondu entièrement au questionnaire, nous avons

$$(wa)_i = [\hat{N}_b - (\hat{M}_b - \hat{M}_{bQ})] / \hat{M}_{bQ} \tag{3.9}$$

où $\hat{M}_{bQ} = \sum_{i \epsilon b} \pi_i^{-1} t_i \delta_i \Pi_{q=1}^{Q} \delta_{iq}$, est le nombre pondéré d'unités de la cellule b, qui ont répondu à toutes les questions du questionnaire.

$\delta_{iq} = 1$ ou 0 selon que l'unité i a répondu ou non à la question n° q du questionnaire d'enquête comportant Q questions; par conséquent, $\Pi_{q=1}^{Q} \delta_{iq} = 1$ seulement si l'unité i a répondu à toutes les questions du questionnaire.

On pourrait justifier l'utilisation de (3.9) au lieu de (3.8) par les mêmes raisons qui ont motivé l'utilisation de (3.7a) au lieu de (3.6). Il est nécessaire toutefois d'expliquer pourquoi il est préférable d'utiliser les taux de réponse pondérés au lieu des taux de réponse non pondérés; cette explication est fournie après le Tableau 6.

À partir des équations (3.7a) et (3.9), il serait possible de définir une expression $(wa)_i$ distincte pour chaque question q ou chaque caractéristique y définie par une ou plusieurs questions. Malheureusement, des questions ou des caractéristiques différentes supposeraient nécessairement des facteurs de correction de poids différents dans une cellule de correction, ce qui créerait des incohérences entre diverses caractéristiques figurant dans des tableaux publiés. Pour assurer l'uniformité des poids de l'enquête et des facteurs de correction de poids, il faudrait que $(wa)_i$ dépende uniquement de l'unité et non de la question ou de la caractéristique. L'imputation pourrait néanmoins être permise pour certaines questions mais non pour d'autres, comme les principales questions incluses dans les facteurs de correction de poids (3.7a) ou (3.9), pour autant qu'il y ait cohérence entre les inclusions et les exclusions à l'intérieur de la cellule de correction. Pour calculer un facteur de correction de poids, on pourrait, par exemple, envisager de faire une imputation par déduction logique plutôt que par la méthode *hot deck*, en se fondant sur un questionnaire entièrement rempli.

Pour chacun des facteurs de correction de poids définis ci-dessus (3.6 à 3.9), il est possible de montrer que (2.2) est une forme particulière de (2.1) lorsque z_i correspond à une moyenne pondérée ou non pondérée des répondants. Ainsi, la valeur imputée implicite z_i

pour l'unité non répondante i est indiquée au Tableau 6 pour chacun des 4 facteurs de correction de poids définis ci-dessus. Pour bien comprendre les expressions figurant au Tableau 6, il est nécessaire de définir les termes suivants:

$$\hat{T}_b = \sum_{i=1}^{N_b} t_i \pi_i^{-1} \delta_i [\delta_{iy} y_i + (1 - \delta_{iy}) z_{iy}] = \text{total pondéré de l'échantillon d'unités ayant répondu entièrement au questionnaire, y compris les imputations pour non-réponse à une question mais sans les corrections de poids effectuées au moyen de l'inverse du taux de réponse au questionnaire,} \quad (3.10)$$

$$\hat{T}_{by} = \sum_{i=1}^{N_b} t_i \pi_i^{-1} \delta_i \delta_{iy} y_i = \text{total pondéré de l'échantillon d'unités ayant répondu partiellement ou entièrement au questionnaire en fonction de la caractéristique } y, \quad (3.11)$$

$$\hat{T}_{bQy} = \sum_{i=1}^{N_b} t_i \pi_i^{-1} \delta_i \prod_{q=1}^{Q} \delta_{iq} y_i = \text{total pondéré de l'échantillon d'unités ayant répondu partiellement ou entièrement au questionnaire en fonction de la caractéristique } y, \text{ sans les unités de la cellule qui ont fait l'objet d'une imputation pour non-réponse à une question.} \quad (3.12)$$

Alors, $\hat{T}_{bQy} \leq \hat{T}_{by} \leq \hat{T}_b$.

Tableau 6
Valeur imputée implicite pour les non-répondants au questionnaire, selon le facteur de correction de poids (au niveau de la cellule)

	Facteur de correction de poids	Equation correspondante	Valeur imputée implicite lorsque $i=0$	Description
a)	n_b/m_b	(3.6)	$\hat{T}_b/(\pi_i^{-1} m_b)$	Taux de réponse au questionnaire non pondéré
b)	$\hat{N}_b/\hat{M}_b$	(3.8)	$\hat{T}_b/\hat{M}_b$	Taux de réponse au questionnaire pondéré
c)	$\dfrac{n_b - m_b + m_{bQ}}{m_{bQ}}$	(3.7a)	$\hat{T}_{bQy}/\pi_i^{-1} m_{bQ}$	Taux de réponse au questionnaire non pondéré parmi les unités qui ont répondu à toutes les questions
d)	$\dfrac{\hat{N}_b - \hat{M}_b + \hat{M}_{bQ}}{\hat{M}_{bQ}}$	(3.9)	$\hat{T}_{bQy}/\hat{M}_{bQ}$	Taux de réponse au questionnaire pondéré parmi les unités qui ont répondu à toutes les questions

Note: Dans le cas d'un échantillon auto-pondéré (par exemple, EASSR), la valeur imputée implicite z_i devient la moyenne simple des répondants, pour a) et b), et la moyenne simple des répondants (à l'exclusion de ceux qui n'ont pas répondu à toutes les questions) pour c) et d).

* Voir annexe I pour les calculs.

Le facteur de correction de poids indiqué en c) $(n_b - m_b + m_{bQ})/m_{bQ} = 1 + (n_b - m_{bQ})/m_{bQ}$ est $\geq$ à celui indiqué en a) (n_b/m_b) puisque $m_{bQ} \leq m_b$(voir Tableau 6). Ainsi, pour un taux de réponse donné m_b/n_b dans la cellule, la variance d'une estimation obtenue à l'aide du facteur de correction c) risque d'être plus élevée que celle d'une estimation obtenue à l'aide du facteur de correction a). L'augmentation de la variance peut être compensée ou non par une réduction possible du biais dû à l'imputation dans l'erreur quadratique moyenne globale. Le même raisonnement, s'applique aux taux de réponse pondérés figurant en d) $(\hat{N}_b - \hat{M}_b + \hat{M}_{bQ})/\hat{M}_{bQ}$ par rapport à celui figurant en b) $(N_b/\hat{M}_b)$ puisque $\hat{M}_{bQ} \leq \hat{M}_b$. Dans l'échantillonnage avec ppt, l'utilisation de taux de réponse pondérés par rapport à des taux de réponse non pondérés nous amène à une autre conclusion intéressante. Platek et Gray (1983, p. 264-265), on montré que lorsque les probabilités de réponse et de sélection, c'est-à-dire, α_i et π_i sont interdépendantes, les facteurs de correction de poids formés avec des taux de réponse pondérés tendent à être plus élevés que ceux formés avec des taux non pondérés. Ainsi, lorsqu'il y a une corrélation positive entre α_i et π_i, $E(\hat{N}_b/\hat{M}_b) > E(n_b/m_b)$ parallèlement, $E[(\hat{N}_b - \hat{M}_b + \hat{M}_{bQ})/\hat{M}_{bQ}] > E[(n_b - m_b + m_{bQ})/m_{bQ}]$, où $E = E_1E_2$, est l'espérance mathématique pour l'ensemble des échantillons d'unités et des sous-échantillons d'unités répondantes (voir Platek et Gray (1983), p. 251).

Bref, quel que soit le facteur de correction de poids utilisé pour compenser la non-réponse au questionnaire, il n'est pas certain que les valeurs imputées implicites z_i se rapprochent des valeurs réelles Y_i ou même des valeurs observées probables y_i. Tout ce que nous pouvons faire relativement à cette question, c'est d'espérer que les cellules de correction créées pour compenser la non-réponse au questionnaire permettront d'uniformiser le plus possible les caractéristiques des répondants et des non-répondants dans les cellules. La formation et la délimitation des cellules de correction sont donc fondamentales pour la compensation, peu importe le genre de facteur de correction de poids qui est utilisé.

7. CONCLUSION

Comme nous l'avons vu dans les sections précédentes, il n'y a pas de solution toute faite à la non-réponse, peu importe le genre de non-réponse qui se produit. Il s'agit tout d'abord de réduire au minimum le taux de non-réponse sans que cela n'engendre des frais trop élevés ou ne supprime l'actualité des données d'enquête. Au départ, on devra savoir prévoir les cas de non-réponse et élaborer des stratégies d'imputation. Si la non-réponse se manifeste à peu près de la manière prévue, le traitement des données se déroulera conformément au programme établi, notamment au moyen de substitutions et de corrections de poids appropriées. Il est clair que les enquêtes permanentes ou répétées permettent un déroulement plus ordonné des diverses opérations (collecte des données, publication, etc.) que les enquêtes spéciales ou ponctuelles, où il est difficile pour le concepteur de prévoir tous les obstacles qui peuvent surgir en cours d'enquête, comme les refus inattendus ou le manque d'intérêt aussi bien de la part des intervieweurs que des répondants.

Pour surmonter les problèmes de non-réponse, il est essentiel de faire une étude continue des taux de non-réponse par caractéristique d'enquête (dans le cas de la non-réponse à une question) et des causes de la non-réponse et, si possible, d'analyser, par la même occasion, les probabilités de réponse à une question et au questionnaire afin de pouvoir estimer les biais dus à l'imputation à partir de l'enquête proprement dite. Par ailleurs, il est souhaitable d'analyser le biais dû à l'imputation en approfondissant la recherche sur les estimations fondées sur les modèles et d'améliorer les estimations à l'aide de renseignements additionnels.

ANNEXE

Calcul de la valeur implicite z_i pour l'imputation dans les cas de non-réponse au questionnaire

Pour les cas c) et d) du Tableau 6, l'estimation au niveau de la cellule b est définie:

$$\tilde{Y}_b = \hat{T}_{bQy}\,(wa)_i + (\hat{T}_b - \hat{T}_{bQy}) \tag{A.1}$$

$$= \hat{T}_b + [(wa)_i - 1]\hat{T}_{bQy}$$

Pour ce qui est de c), $(wa)_i - 1 = (n_b - m_b)/m_{bQ}$

$$= \sum_i t_i(1 - \delta_i)/m_{bQ}$$

ou $\tilde{Y}_b = \hat{T}_b + \sum_i t_i\pi_i^{-1}(1 - \delta_i)\hat{T}_{bQy}/\pi_i^{-1}m_{bQ}$ (A.2)

par ailleurs, en mettant (A.2) en équation avec (A.1), compte tenu de la définition de $\hat{T}_b$ donnée en (3.10), on peut voir que la valeur imputée z_i est défini par $\tilde{Y}$, ce qui correspond à l'élément c) du Tableau 6.

De même, lorsque des taux de réponse pondérés sont utilisés, on peut voir que la valeur imputé implicite z_i équivaut à $\hat{T}_{bQy}/\hat{M}_{bQ}$, ce qui correspond à l'élément d) du Tableau 6. On obtient les élément a) et b) du Tableau 6 en posant $m_{bQ} = m_b$ and $\hat{M}_{bQ} = \hat{M}_b$.

BIBLIOGRAPHIE

HORVITZ, D.G., et THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

LESSLER, J.T. (1979). An expanded survey error model. Dans *Incomplete Data in Sample Surveys*, volume n° 3 - Compte rendu des travaux du symposium (éd. W.G. Madow, I. Olkin, et D.B. Rubin), San Diego: Academic Press, 259-270.

PLATEK, R. (1977). Some factors affecting nonresponse. *Techniques d'enquête*, 3, 191-214.

PLATEK, R. (1980). Causes of incomplete data, adjustments and effects. *Techniques d'enquête*, 6, 93-132.

PLATEK, R., et GRAY, G.B. (1978). Non-response and imputation. *Techniques d'enquête*, 4, 144-177.

PLATEK, R., et GRAY, G.B. (1979). Methodology and application of adjustments for nonresponse. Exposé présenté à la 42e séance de l'*Institut internationale de statistique*, Manille, Philippines.

PLATEK, R., et GRAY, G.B. (1983). Part V - Imputation Methodology: Total Survey Error. Dans *Incomplete Data in Sample Surveys*, volume n° 2 - Theory and Bibliographies (éd. W.G. Madow, I. Olkin, et D.B. Rubin), San Diego: Academic Press, 249-333.

POTERBA, J.M., et SUMMERS, L.H. (1984). Response Variation in the CPS: Caveats for the unemployment analyst. *Monthly Labour Review*, mars 1984, résumés de recherche, 37-43.