Techniques d'enquête, décembre 1986
Vol. 12, n° 2, pp. 167–188
Statistique Canada

Méthodes de compensation de la non-réponse au U.S. Bureau of the Census

DAVID W. CHAPMAN, LEROY BAILEY, et DANIEL KASPRZYK[1]

RÉSUMÉ

Presque tous les recensements et enquêtes comportent deux types de non-réponses: la non-réponse au questionnaire (non-réponse totale) et la non-réponse à une question (non-réponse partielle). Plusieurs méthodes de compensation de la non-réponse ont été élaborées pour tenter de réduire la distorsion due à la non-réponse. Cet article résume les méthodes de compensation de la non-réponse utilisées au U.S. Census Bureau, et traite particulièrement du problème de la non-réponse au questionnaire. On examine aussi sommairement les travaux de recherche actuels et futurs dans ce domaine.

MOTS CLÉS: Compensation de la non-réponse; imputation; données manquantes; pondération.

1. INTRODUCTION

Le Bureau of the Census a admis depuis longtemps l'importance que peuvent avoir les erreurs de mesure imputables à la non-réponse dans les enquêtes et a toujours prévu des méthodes de compensation de la non-réponse dans les méthodes d'estimation utilisées pour les nombreux et divers recensements et enquêtes qu'il effectue. Cet article a pour objet de donner un aperçu des méthodes de compensation de la non-réponse utilisées par le Census Bureau, principalement dans le cas de la non-réponse au questionnaire. Par non-réponse au questionnaire nous entendons les cas où peu ou aucun renseignement n'a été obtenu, relativement aux principales variables d'enquête, pour l'unité d'échantillonnage visée.

Cet exposé comprend 1) une analyse de la méthode de pondération générale appliquée dans les enquêtes démographiques, 2) un examen de quelques-uns des problèmes liés à la non-réponse dans la Survey of Income and Program Participation (SIPP), 3) une analyse du traitement de la non-réponse au questionnaire dans les recensements et enquêtes à caractère économique, et 4) une section concernant l'imputation pour les gains dans le cas de la Current Population Survey. En plus de présenter les diverses méthodes de compensation de la non-réponse utilisées par le Census Bureau, les auteurs examinent les problèmes particuliers inhérents à ces méthodes et décrivent les travaux et préoccupations courantes du Bureau relativement à la recherche sur la non-réponse.

2. LA NON-RÉPONSE DANS LES ENQUÊTES-ÉCHANTILLONS DÉMOGRAPHIQUES

Il peut arriver, un moment donné, que le Bureau of the Census s'occupe de près d'une trentaine d'enquêtes démographiques périodiques ou spéciales. Ces enquêtes portent sur l'activité, le revenu des particuliers et des familles, les soins médicaux, le transport, les loisirs, la criminalité et d'autres sujets qui reflètent les intérêts courants des citoyens, gouvernements, entreprises et institutions du pays. Les taux de non-réponse au questionnaire dans ces enquêtes varient de trois et quatre pour cent dans la National Crime Survey à plus de 25% (taux relevé pour la National Survey of Natural and Social Scientists and Engineers de 1984).

[1] David W. Chapman et Leroy Bailey, chercheurs principaux (Principal Researchers), Statistical Research Division, U.S. Bureau of the Census, Washington D.C. 20233. Daniel Kasprzyk, adjoint spécial, Office of the Chief, Population Division, U.S. Bureau of the Census, Washington D.C. 20233.

En dépit des différences dans le contenu et la nature des enquêtes démographiques du Census Bureau ainsi que de l'importance de la non-réponse au questionnaire, la correction de poids à l'intérieur des classes (Oh et Scheuren 1983), aussi appelée compensation par cellule, est la principale technique utilisée pour compenser la non-réponse au questionnaire. On n'applique pas la même méthode d'établissement des classes de compensation pour la pondération. Dans certaines enquêtes, les données auxiliaires connues et utilisées pour définir les classes de pondération se limitent à des renseignements de base de nature géographique et à l'information tirée du plan de sondage, tandis que pour d'autres enquêtes on peut disposer d'un volume considérable de données démographiques et économiques.

Les facteurs de compensation de la non-réponse établis pour les enquêtes démographiques du Census Bureau sont habituellement l'inverse du taux de réponse pondéré ou non de l'enquête. Dans un petit nombre d'enquêtes, ce facteur est légèrement modifié de manière qu'on puisse tenir compte des renseignements fournis dans les enquêtes de suivi menées auprès de sous-échantillons de personnes qui n'avaient pas répondu à l'enquête. Puisque la méthode généralement utilisée par le Census Bureau dans les cas de non-réponse à une enquête est essentiellement la même pour toutes ses principales enquêtes démographiques, nous donnons, à la section 2.1, une description générale de la méthode de compensation de la non-réponse utilisée pour la National Crime Survey (NCS), à titre d'exemple d'une application "typique" de la pondération au Census Bureau. La section 2.2 présente une analyse d'autres méthodes de compensation possibles ainsi que les travaux de recherche en cours sur la non-réponse aux enquêtes démographiques.

2.1 L'enquête nationale sur la criminalité (National Crime Survey)

L'échantillon de la NCS est un échantillon aléatoire composé d'environ 72,000 ménages à l'échelle nationale. Cet échantillon est divisé en six panels et chaque panel est interviewé un mois donné et à nouveau à des intervalles de six mois sur une période de trois ans. L'enquête vise à mesurer la criminalité ainsi que le nombre de fois où les membres d'un ménage âgés de 12 ans ou plus ont été victimes d'agression (y compris le viol), de cambriolage, de vol simple, de vol d'auto et de vol. [Pour une description détaillée de la NCS, voir U.S. Department of Commerce, Bureau of the Census (1977).]

Pour produire les estimations trimestrielles de la NCS, il faut tout d'abord gonfler les données-échantillons en les multipliant par l'inverse des probabilités de sélection correspondantes. Les cas où il n'y a pas eu de contact et les refus représentent entre trois et quatre pour cent des logements occupés inclus chaque mois dans l'enquête. Des corrections sont apportées pour tenir compte de ces unités: on applique des facteurs de compensation aux données pondérées sur les répondants dans les classes de pondération. Ces classes sont définies de telle sorte que les répondants et les non-répondants dans chaque classe ont des caractéristiques semblables. Afin de réduire l'effet de la compensation de la non-réponse sur la variance des estimations de l'enquête, il faut généralement combiner certaines des plus petites classes de pondération avec d'autres classes avant de pouvoir effectuer une compensation finale de la non-réponse. Il faut également combiner les classes lorsque les facteurs d'ajustement deviennent trop importants pour au moins une des classes. [Voir Hanson (1978).] On discute plus en détail de la combinaison de classes à la section 4.

Comme la méthode utilisée pour la NCS en est une où les répondants remplissent eux-mêmes le questionnaire, on se préoccupe de l'importance de la non-réponse à l'intérieur des ménages. Aussi, on a établi un ensemble distinct de cellules de pondération pour compenser la non-réponse à l'intérieur des ménages. Ces cellules ou classes de pondération, ainsi que celles qui servent à la compensation de la non-réponse des ménages, sont décrites dans les tableaux 1 à 3. Les taux de non-réponse des ménages et de non-réponse à l'intérieur des ménages pour la NCS de 1984 figurent au tableau 4.

Tableau 1

Cellules de correction pour la non-interview dans les cas de non-réponse à l'intérieur des ménages

Lien avec les membres du ménage	Personnes selon l'âge, selon la race du chef							
	Noir				Non-noir			
	12-34	25-44	45-64	65+	12-24	25-44	45-64	65+
Chef de ménage								
Épouse du chef								
Toutes les autres personnes								

Tableau 2

Cellules de correction pour la non-interview des ménages, pour la NCS, dans le cas des Standard Metropolitan Statistical Area (SMSA)

Race	Principale ville de la SMSA	Reste de la SMSA	
		Région urbaine	Région rurale
Blanc			
Non-blanc			

Tableau 3

Cellules de correction pour la non-interview des ménages, pour la NCS, dans le cas des régions autres qu'une SMSA

Race	Région urbaine	Région rurale	
		Non agricole	Agricole
Blanc			
Non-blanc			

Pour illustrer l'estimateur d'un total de la NCS, posons une probabilité de sélection, $\pi_i = 1, 2, \ldots, N$, appliquée à chacune des N unités dans la population. On suppose que parmi les n unités d'échantillonnage, n_R sont des unités répondantes. L'estimateur de la NCS pour l'ensemble de la population, peut être exprimé par la formule suivante, une fois la compensation de la non-réponse au questionnaire effectuée:

$$\hat{Y}_{\text{NCS}} = \sum_{j=1}^{M} \sum_{k=1}^{P} (z_j u_k)^{-1} \sum_{\ell=1}^{n_{Rjk}} \frac{y_{jk\ell}}{\pi_{jk\ell}}.$$

où pour les unités visées dans la k^e classe de pondération à l'intérieur des ménages et la j^e classe de pondération des ménages,

$y_{jk\ell}$ = valeur du répondant ℓ dans l'échantillon
n_{Rjk} = nombre de répondants dans l'échantillon
n_{jk} = nombre de cas dans l'échantillon
z_j = le taux de réponse des ménages estimé
u_k = le taux de réponse à l'intérieur des ménages estimé,
$\pi_{jk\ell}$ = probabilité de sélection du répondant ℓ dans l'échantillon
P = nombre total de classes de pondération pour la non-réponse à l'intérieur des ménages
M = nombre total de classes de pondération pour la non-réponse des ménages.

Tableau 4

Taux de non-interview dans la NCS - 1984

	Average 1984	janv.	fév.	mars	avril	mai	juin
Non-interview des ménages							
Nombre total de chefs de ménage interviewés	11,769	11,916	11,925	11,743	11,809	11,918	9,482
Total	430	446	540	481	446	388	348
Taux	3.5	3.6	4.3	3.9	3.6	3.2	3.5
Personne à domicile	0.9	0.8	1.1	0.9	0.9	0.7	1.0
Absent temporairement	0.6	0.6	0.6	0.8	0.6	0.4	0.7
Refus	1.9	2.1	2.6	2.2	2.2	2.0	1.9
Autre	0.1	0.2	0.2	0.1	0.1	0.1	0.1
Non-interviews à l'intérieur des ménages							
Total	685	655	751	701	806	804	697
Taux	2.5	2.6	3.0	2.8	3.0	2.9	3.2
		juil.	août	sept.	oct.	nov.	déc.
Non-interview des ménages							
Nombre total de chefs de ménage interviewés		9,869	9,446	9,895	9,350	9,692	9,410
Total		411	409	337	406	387	346
Taux		4.0	4.2	3.3	4.2	3.8	3.5
Personne à domicile		0.9	0.9	0.6	1.0	1.2	1.0
Absent temporairement		1.0	1.0	0.6	0.6	0.4	0.4
Refus		2.1	2.3	2.0	2.4	2.1	2.1
Autre		0.1	0.1	0.1	0.3	0.3	0.1
Non-interviews à l'intérieur des ménages							
Total		709	678	666	728	735	803
Taux		3.1	3.1	2.9	3.4	3.3	3.7

Les hypothèses suivantes sont faites implicitement pour la formation des classes de pondération pour la non-réponse dans la NCS ainsi que pour celles d'autres enquêtes démographiques:

1. Il existe une corrélation "importante" entre les principales variables d'enquête et les covariables utilisées pour définir les classes voisines pour la non-interview.
2. Dans chaque classe de pondération pour la non-réponse des ménages, $E(\bar{y}_{Rj}) = E(\bar{y}_{\bar{R}j})$, où $\bar{y}_{Rj}$ et $\bar{y}_{\bar{R}j}$ sont les moyennes pour les répondants et les non-répondants dans l'échantillon, respectivement, dans la j^e classe de pondération.
3. Les moyennes des classes de pondération diffèrent, c'est-à-dire, $E(\bar{y}_{Rj}) \neq E(\bar{y}_{Rj'})$, $j \neq j'$.

(Des hypothèses semblables sont aussi faites implicitement pour les classes de compensation de la non-réponse à l'intérieur des ménages.)

Le choix des classes de pondération selon cette méthode est limité par le fait que des mesures pour les variables de la classe de pondération (covariables) doivent être produites (soit avant, soit au cours de l'enquête) pour les répondants ainsi que pour les non-répondants. Cela limite essentiellement les caractéristiques utilisées pour définir les classes par rapport à celles concernant la géographie, la race, le domicile dans une ville ou non, les caractéristiques du logement et les niveaux de plan de sondage. Les moyens qu'offre cette méthode pour réduire les biais dépendent, en partie, de la mesure dans laquelle les classes de pondération pour la non-réponse, dans la NCS, confirment les trois hypothèses données plus haut. Aucun résultat définitif relatif à ce sujet n'est connu actuellement, mais des travaux de recherche sont en cours et il semble que d'autres études empiriques sont justifiées.

2.2 Autres méthodes possibles de compensation

Il existe un certain nombre d'autres méthodes qui pourraient être utilisées, à la place de la pondération, pour corriger la non-réponse (voir par exemple Little, 1986, chapitre 5). Cependant, il n'existe pas de résultats définitifs qui montrent qu'une de ces autres méthodes offre des avantages appréciables. Les sections 2.2.1 et 2.2.2 décrivent sommairement deux autres méthodes qui font présentement l'objet d'analyses en vue de leur application dans des enquêtes démographiques.

2.2.1 Estimations distinctes pour des types de non-réponse différents

Dans les enquêtes démographiques, on peut diviser les non-répondants en quatre catégories: refus (REF), absent (ABS), autre logement occupé (ALO), ou logement où il a été impossible d'obtenir une réponse à cause de circonstances spéciales. Pour décrire ces cas de non-réponse, on utilise l'expression non-interview de type A. Le groupe ABS peut être divisé en ménages ou personnes dont l'absence prolongée de leur domicile empêche toute interview au cours de la période prévue pour les interviews (ABS_L) (absence de longue durée), et le groupe des personnes dont le retour est prévu au cours de la période d'enquête (ABS_C) (absence de courte durée).

Les auteurs ne connaissent pas de données qui démontrent que les quatre groupes de non-réponse sont généralement semblables. En fait, les résultats de la Current Population Survey du Census Bureau et ceux de l'Enquête sur la population active, effectuée au Canada, semblent indiquer que les ménages ABS_C sont vraisemblablement plus petits, plus jeunes et qu'ils comprennent une plus grande proportion de personnes occupées que les autres groupes. Le groupe ABS_L est habituellement plus âgé et son taux d'emploi est relativement faible. Le groupe interviewé peut mieux correspondre au groupe REF et ALO. [Voir Palmer et Jones (1967) et Paul et Lawes (1982).] Il est concevable qu'un traitement distinct des quatre groupes de non-réponse puisse mieux compenser les cas de non-réponse que ce que fait la méthode courante. Cette possibilité fait l'objet d'une étude par un groupe de recherche sur les méthodes de compensation de la non-réponse pour la NCS.

2.2.2 Pondération avec des probabilités de réponse

Plusieurs techniques de pondération où l'on applique le concept de probabilité de réponse ont été proposées. La plupart de ces techniques sont basées sur des définitions établies par Politz et Simmons (1949) qui groupent les répondants dans l'échantillon selon les estimations de leurs probabilités de répondre. Le facteur utilisé pour gonfler les données-échantillons afin d'obtenir les groupes de pondération est l'inverse de la probabilité de réponse estimée. La méthode de Politz-Simmons comporte des limites importantes, comme le fait qu'elle ne peut être appliquée au traitement des refus. Cependant, un certain nombre de développements et d'applications de la procédure ont été faits récemment, notamment par Anderson (1978) et

Thomsen et Sirling (1983). Il se peut que ces méthodes soient applicables aux enquêtes périodiques pour lesquelles beaucoup de rappels sont effectués.

Des recherches sont en cours sur la construction de modèles qui pourraient servir, pour plusieurs enquêtes démographiques, à l'estimation des probabilités de réponse des unités pour lesquelles les "variables indépendantes" ont des valeurs semblables. On étudie la faisabilité et les mérites de calculer des facteurs de compensation de la non-réponse ainsi que de construire des classes de pondération basées sur de tels modèles (désignés parfois stratification de la propension à répondre). [Voir Rosenbaum et Rubin (1983) et Little et Samuhel (1983).] De plus, des travaux se poursuivent en vue d'élaborer des méthodes plus objectives de pondération de l'échantillon, et ce afin de contrôler les erreurs liées à la non-réponse.

3. L'ENQUÊTE SUR LE REVENU (SURVEY OF INCOME AND PROGRAM PARTICIPATION)

L'enquête sur le revenu aux États-Unis ou Survey of Income and Program Participation (SIPP) est un nouveau programme d'enquête national permanent mené auprès des ménages par le U.S. Bureau of the Census. L'objet de la SIPP est d'améliorer la mesure de l'information sur la situation économique des ménages et des particuliers aux États-Unis. C'est l'aboutissement d'un vaste programme d'élaboration, l'Income Survey Development Program (ISDP), dans le cadre duquel on a examiné les définitions, les méthodes, les questionnaires, les périodes visées par l'enquête et autres éléments du genre. [Pour une description du ISDP, voir Ycas et Lininger (1981).] On s'attend que les données obtenues dans le cadre de la SIPP serviront à l'analyse du système de transfert fédéral, à l'estimation du coût des programmes si des modifications sont apportées aux critères d'admissibilité, à l'évaluation des effets des changements apportés aux programmes sur certains sous-groupes de la population ainsi qu'à l'examen des modifications apportées au régime fiscal.

La SIPP a été lancée en octobre 1983 comme programme d'enquête permanent. Elle comportait un panel d'environ 21,000 ménages de logements occupés admissibles à participer à une interview dans 174 unités primaires d'échantillonnage (U.P.É.) choisies pour représenter les membres de la population des États-Unis qui n'habitent pas dans un établissement. (À compter de 1985, un nouveau panel est inclus dans l'enquête, en février de chaque année; le panel de 1985 comprenait 14,500 ménages admissibles à une interview.)

Chaque ménage est interviewé une fois tous les quatre mois, pendant environ 2½ ans, dans le but de recueillir suffisamment de données pour pouvoir effectuer une analyse longitudinale tout en fixant une période relativement courte pour la déclaration du revenu mensuel. La période de référence des principales questions de l'enquête correspond aux 4 mois qui précèdent l'interview. Ce plan de sondage permet d'effectuer huit interviews par ménage et aussi de produire des estimations transversales à partir des réponses de plus d'un panel.

Pour faciliter le travail sur le terrain et le traitement, chaque panel de l'échantillon est divisé en quatre sous-échantillons comprenant approximativement le même nombre de ménages; ce sont des groupes de renouvellement. Un groupe de renouvellement est interviewé chaque mois. Ainsi, un cycle d'interviews, faites à l'aide du même questionnaire, prend quatre mois consécutifs. Les taux de non-interview cumulatifs des ménages sont donnés au tableau 5 pour les panels de la SIPP de 1984.

Au moment de la visite de l'interviewer, on demande à chaque personne présente, âgée de 15 ans ou plus, de fournir des renseignements à propos d'elle-même; on demande à un membre du ménage de fournir les renseignements pour les absents. Une caractéristique importante du plan de sondage de la SIPP, c'est que toutes les personnes qui font partie d'un ménage échantillonné au moment de la première interview demeurent dans l'échantillon même

Tableau 5

Taux de non-interview cumulatifs des ménages pour les panels de la SIPP de 1984

Cycle	Perte d'effectifs de l'échantillon
1	4.9%
2	9.4%
3	12.3%
4	15.4%
5	17.4%
6	19.4%
7	21.0%
8	22.0%
9	22.3%

si elles déménagent au cours des deux années et demie suivantes. Pour des raisons d'ordre opérationnel et budgétaire, les interviews sur place sont effectuées aux nouvelles adresses seulement si ces dernières se trouvent à moins de 100 milles d'une unité primaire d'échantillonnage de la SIPP. Plus de 96% de la population des États-Unis habite dans les régions géographiques établies de cette façon. Pour les personnes qui ont déménagé à l'extérieur de cette limite de 100 milles, on essaie d'effectuer une interview téléphonique.

Après la première interview, l'échantillon de la SIPP est un échantillon de personnes, qui comprend toutes les personnes qui habitaient l'unité d'échantillonnage au moment de la première interview. Les personnes âgées de 15 ans ou plus qui, par la suite, partagent des pièces d'habitation avec les personnes qui faisaient partie de l'échantillon au début sont aussi interviewées afin qu'on puisse définir le contexte économique global des personnes qui faisaient partie de l'échantillon initial.

Pour plus de renseignements sur le plan de sondage, le contenu et les opérations relatives à la SIPP, voir l'ouvrage de Nelson, McMillen et Kasprzyk (1985).

3.1 Compensations de la non-réponse dans la SIPP

Les données recueillies dans le cadre de la SIPP peuvent être étudiées de deux points de vue; elles peuvent servir à des études transversales ou à des études longitudinales. Dans le premier cas, chaque interview effectuée pour la SIPP est traitée comme une enquête transversale distincte qui fournit des estimations ponctuelles. Pour des exemples de ces estimations, voir U.S. Department of Commerce, Bureau of the Census (1984a). Du point de vue longitudinal, les données sont recueillies à plusieurs moments donnés et chaque enregistrement de l'enquête est considéré non comme un ensemble d'observations sans liens entre elles, mais comme un ensemble de variables entre lesquelles il existe une dépendance logique à deux moments ou plus. Comme le traitement informatique et l'estimation statistique sont effectués en fonction des analyses longitudinales, ces opérations sont basées sur l'utilisation de données recueillies dans deux interviews ou plus.

Puisqu'on peut considérer la SIPP dans une perspective d'analyses tant longitudinales que transversales, les fichiers de microdonnées à grande diffusion comprennent des fichiers de données transversales produits cycle par cycle ainsi que des fichiers longitudinaux. Cela suppose donc deux systèmes distincts de traitement de la non-réponse à l'enquête.

3.1.1 Compensations de la non-réponse transversale au questionnaire

Pour la SIPP, la compensation de la non-réponse transversale au questionnaire est semblable à la façon dont les corrections pour la non-interview sont effectuées dans les autres enquêtes périodiques du Census Bureau. Les variables utilisées pour définir les cellules de correction pour les cas de non-interview des ménages dans le premier cycle d'interview de la SIPP sont les suivantes [voir U.S. Department of Commerce, Bureau of the Census (1983 et 1984b)]:

1. Région de recensement - Nord-est, Midwest, Sud, Ouest.
2. Résidence - région métropolitaine type (Standard Metropolitan Statistical Area, SMSA), hors d'une SMSA.
3. Localité/hors localité - pour les unités qui ne sont pas dans une SMSA.
   Principale ville / autres agglomérations - pour les unités dans une SMSA.
4. Race de la personne-repère - Noir, Non-Noir.
5. Mode d'occupation - propriétaire de son domicile, locataire.
6. Taille du ménage - 1, 2, 3, 4 personnes ou plus.
7. Groupe de renouvellement - 1, 2, 3, 4.

Chaque classe de pondération doit satisfaire à deux conditions: 1) elle doit comprendre au moins 30 unités non pondérées et 2) le facteur de compensation de la non-réponse, dans le cas d'une classe de pondération, doit être inférieur ou égal à 2.0. Pour un groupe de renouvellement donné, la procédure d'agrégation visant à satisfaire à ces deux conditions est appliquée indépendamment pour chacune des quatre combinaisons de mode d'occupation selon la race. (Pour le premier cycle, il n'y avait pas de facteur de compensation de la non-réponse *à l'intérieur* des ménages.)

Dans les cycles ultérieurs de la SIPP, le facteur de compensation de la non-réponse tient compte des non-interviews des unités qui ont déménagé et qui ne peuvent être retracées ou qui ont déménagé à plus de 100 milles d'une U.P.É. de la SIPP et qu'on ne peut joindre par téléphone, ainsi que des unités qui refusent de répondre, etc. Les rajustements sont effectués pour chaque mois de la période de référence ainsi que pour le mois d'interview, afin de tenir compte de l'augmentation du nombre de non-interviews qui est due au fractionnement des ménages dans l'échantillon. La procédure est semblable à celle qui est utilisée pour déterminer le facteur de compensation de la non-réponse des ménages au 1er cycle; cependant, les variables utilisées pour définir les classes de pondération diffèrent. Ces variables sont:

1. Race (Blanc, Non-Blanc) et origine hispanique (hispanique, non hispanique) de la personne-repère: a) la personne-repère est blanche et non hispanique, et b) autres.
2. Genre de ménages - trois catégories: a) femme responsable du ménage, sans conjoint, avec ses propres enfants de moins de 16 ans, b) membre responsable du ménage âgé de 65 ans ou plus, et c) autres.
3. Niveau de scolarité de la personne-repère: a) moins de 8 années, b) de 8 à 11 années, c) de 12 à 15 années, et d) 16 années ou plus.
4. Genre de revenu reçu (d'après la plus récente interview menée auprès des membres du ménage) - deux catégories: a) ménages ayant reçu un revenu d'au moins une des sources suivantes - Supplemental Security Income; Black Lung Payments; Aid to Families with Dependent Children; General Assistance, Indian Assistance, Cuban Assistance ou Refugee Assistance; paiements pour la garde d'un enfant en foyer nourricier; Women's, Infants', and Children's Nutrition Program; Food Stamps; et Medicaid; et b) autres.
5. Éléments d'actifs - deux catégories: a) ménages dans lesquels au moins un membre détient des éléments d'actifs autres qu'un compte d'épargne ou qu'un compte de chèques qui porte intérêt, et b) tous les autres.

6. Mode d'occupation : a) propriétaire de son domicile et b) locataire.
7. Logements sociaux ou allocations de logement – les locataires sont divisés en deux catégories a) ceux qui habitent des logements sociaux ou auxquels le gouvernement verse des allocations de logement et b) ceux qui n'habitent pas des logements sociaux ou auxquels le gouvernement ne verse pas d'allocation de logement.
8. Taille du ménage: 1, 2, 3, 4 personnes ou plus.

Les variables utilisées pour la compensation de la non-réponse des ménages à la deuxième interview de la SIPP et aux interviews subséquentes diffèrent des variables utilisées pour le premier cycle parce que, après la première interview, on dispose de données additionnelles qu'on applique au traitement de la non-réponse dans les interviews suivantes. Cinquante-trois classes de pondération ont été créées à partir de ces variables et c'est le mode d'occupation qui a été la principale variable de fractionnement de l'échantillon. [Pour une description de ces classes de pondération, voir U.S. Department of Commerce, Bureau of the Census (1984c).] Bien qu'on ait défini une méthode d'agrégation des cellules selon laquelle les cellules qui possèdent des caractéristiques semblables liées à la pauvreté sont fusionnées, il y a peu d'agrégation car les facteurs de compensation de la non-réponse sont calculés pour trois groupes de renouvellement (le cycle informatique de la SIPP) plutôt que pour un groupe de renouvellement, comme dans le cas de la première interview.

Il existe une technique de compensation de la non-réponse à l'intérieur des ménages pour le deuxième cycle et ceux qui suivent. Pour appliquer cette technique, qui est fondée sur la méthode "hot deck", il faut reproduire l'enregistrement complet d'un répondant dont on suppose que les caractéristiques sont semblables à celles du non-répondant.

3.1.2 Compensations de la non-réponse longitudinale

Comme les personnes déclarées habitant à l'adresse incluse dans l'échantillon au moment de la première interview constituent l'échantillon de la SIPP pour les cycles qui suivent le premier, c'est en fonction des individus ou des personnes qu'il est le plus utile et logique de décrire la nature du problème de la non-réponse longitudinale dans la SIPP. L'enregistrement des microdonnées de chaque personne est un enregistrement élargi qui contient des variables exprimant souvent la même mesure à un moment différent. Ainsi, dans une enquête par panel de n cycles, il existe 2^n présentations possibles des non-interviews pour une personne échantillonnée. On trouve au tableau 6, les présentations des non-interviews des personnes qui faisaient partie de l'échantillon initial pour les cinq premières interviews (cycles) du panel de 1984, ces renseignements sont adaptés de Kalton, McMillen et Kasprzyk (1986).

Le premier fichier longitudinal de microdonnées de la SIPP contiendra douze mois (trois interviews) de données provenant du panel de la SIPP de 1984, et la personne est la principale unité d'analyse. L'échantillon des cas à pondérer pour ce fichier comprendra seulement les personnes pour lesquelles trois interviews ont été effectuées. Les personnes échantillonnées pour lesquelles seulement une ou deux interviews ont été effectuées seront traitées comme des non-répondants. Les données qu'elles auront fournies aideront à définir les classes de compensation de la non-réponse.

Comme le premier fichier longitudinal de microdonnées concerne uniquement les personnes qui ont répondu aux trois interviews, la question de la correction pour la non-réponse est, à toutes fins pratiques, identique à celle pour le fichier transversal. Cependant, deux facteurs de compensation de la non-réponse sont appliqués aux poids de l'échantillon initial. [Voir Kobilarcik et Singh (1986).] Le premier facteur de compensation tient compte des ménages classés comme des cas de non-interview au premier cycle. Le second facteur tient compte des personnes qui n'ont pas répondu aux trois interviews.

Tableau 6

Présentations des interviews des personnes incluses dans l'échantillon initial pour les cinq premières interviews du panel de la SIPP de 1984

Présentation des réponses	Pourcentage
Réponse à toutes les interviews (5 interviews)	
Présentation: XXXXX	79.1
Cas apparents de perte d'effectifs	13.8
Présentations: XXXXO	3.8
XXXOO	3.1
XXOOO	3.2
XOOOO	3.7
Les première et cinquième interviews ont été complétées, mais au moins une des interviews intermédiaires manque	4.1
Présentations: XXXOX	1.6
XOXXX	0.6
XXOXX	1.2
XXOOX	0.1
XOXOX	0.1
XOOOX	0.3
XOOXX	0.2
La cinquième interview manque et au moins une des deuxième à quatrième interviews manque aussi	0.7
Présentations: XOXXO, XOXOO, XOOXO, XXOXO	
A quitté l'univers (décédé, placé dans un établissement, habite dans des casernements des Forces armées, a déménagé outre-mer)	2.3
Total	100.0
	(25,128)

Pour le premier facteur de compensation, on peut utiliser seulement les variables des ménages, qui ont été observées à la première interview. Les facteurs de compensation sont calculés séparément à l'intérieur des cellules définies par les variables suivantes:

a. Région de recensement
b. Domicile (région métropolitaine, région non métropolitaine)
c. Race de la personne-repère
d. Mode d'occupation (propriétaire, locataire)
e. Taille du ménage

Le deuxième ensemble de facteurs de compensation est appliqué au niveau des personnes. Les facteurs sont calculés à l'intérieur des cellules définies par les caractéristiques suivantes:

a. Revenu mensuel du ménage
b. Situation du ménage de la personne sur le plan de la participation au programme
c. Situation vis-à-vis de l'activité
d. Race
e. Nombre d'années d'études terminées
f. Type d'actifs détenus par le ménage de la personne

Les cellules sont regroupées chaque fois qu'elles ne contiennent pas trente personnes échantillonnées ou lorsque le facteur de compensation de la non-réponse est supérieur à 2.

À mesure que l'enquête progresse, des méthodes plus perfectionnées de compensation de la non-réponse longitudinale seront élaborées en fonction des données fournies par les répondants qui fournissent une partie des renseignements demandés (c.-à-d. pour les personnes échantillonnées qui répondent à certaines, et non à toutes, les interviews des différents cycles auxquels elles sont supposées participer). Le traitement des cas de réponse partielle n'est pas évident. Les données manquantes dues au fait que des personnes n'ont pas répondu à une interview ou plus peuvent être considérées soit comme une non-réponse par une personne, cas généralement traité par des corrections de poids, soit comme une non-réponse à une question, cas généralement traité par imputation. Par exemple, on pourrait considérer le cas d'une personne dont la présentation des réponses serait de type (R, NR, R) comme un cas de non-réponse à une question car l'interview manquante est encadrée par des interviews achevées; mais on pourrait considérer le cas d'une personne dont la présentation des réponses serait de type (NR, R, NR) comme un cas de non-réponse au questionnaire, et le traiter de la même façon qu'un cas de présentation des réponses de type (NR, NR, NR). Nous devons cependant admettre que même dans le cas où la présentation des réponses est du type (R, NR, R) selon la personne, il est encore possible de trouver quatre types de présentations des réponses selon la question. On peut donc considérer de nombreuses possibilités au moment de l'élaboration des méthodes de compensation de la non-réponse pour la base de données longitudinale de la SIPP. Cette question est traitée par Kalton (1986) et par Kalton, Lepkowski, et Lin (1985).

3.2 Recherches relatives à la SIPP

Des travaux entrepris récemment dans deux domaines pourraient aider à la prise de décisions futures concernant la compensation de la non-réponse. Tout d'abord, à compter de la quatrième interview, le questionnaire de la SIPP contient une section sur le "cycle manquant" ("Missing Wave"). Cette section comprend un petit nombre de questions sur l'activité, les sources de revenu et la possession ou non-possession d'actifs, qui doivent être posées aux répondants dans le cycle courant qui n'ont pas répondu au questionnaire dans le cycle précédent. Les répondants qui n'ont pas répondu à deux interviews consécutives ou plus ne sont pas autorisés à remplir la section sur le "cycle manquant".On peut donc, en accordant plus d'importance à la collecte des données, même si cela comporte une légère augmentation du fardeau de déclaration, réduire le problème de la non-réponse due à une personne à un problème de non-réponse à une question. Il faudra cependant évaluer la qualité des données chronologiques avant d'utiliser ces données.

Le second domaine dans lequel des travaux sont en cours est celui des techniques générales de traitement de la non-réponse d'une personne au cours d'un cycle de la SIPP. Graham Kalton et ses collègues du Survey Research Centre doivent 1) comparer les techniques d'imputation et de pondération longitudinales pour le traitement de la non-réponse d'une personne au cours d'un cycle, 2) évaluer les modèles d'imputation et de pondération en fonction de l'analyse du changement entre les cycles et du groupement entre les cycles et 3) définir des critères provisoires pour le choix d'une méthode de traitement de la non-réponse d'une personne au cours d'un cycle. On trouve une analyse de ces questions et d'autres qui seront traitées plus loin dans Kalton (1986), et Kalton et Miller (1986).

Enfin, des recherches sont prévues pour plusieurs autres sujets, notamment 1) l'attribution d'une valeur numérique au choix des variables utilisées pour déterminer les classes de pondération; 2) l'évaluation de la robustesse des estimations de l'enquête relatives à la population et à certains sous-groupes selon diverses méthodes de compensation de la non-réponse et diverses techniques d'agrégation des cellules des classes de pondération; 3) l'étude de la possibilité d'effectuer des compensations distinctes de la non-réponse selon le genre de non-interview; 4) l'étude de l'effet de la suppression de données déclarées dans une enquête afin de

simplifier le problème des données manquantes dans la SIPP; et 5) l'évaluation des méthodes de compensation de la non-réponse longitudinale établies pour le premier fichier longitudinal de recherche de la SIPP.

4. MÉTHODES DE COMPENSATION DE LA NON-RÉPONSE POUR LES RECENSEMENTS ET ENQUÊTES ÉCONOMIQUES

Le Bureau of the Census effectue six recensements économiques toutes les cinq années; les plus récents sont ceux qui portent sur l'année 1982. Ces six recensements économiques sont désignés par l'appellation des secteurs commerciaux suivants:

1) Commerce de détail
2) Commerce de gros
3) Services
4) Fabrication
5) Industries extractives
6) Construction

En plus des recensements économiques, le Census Bureau effectue un recensement des administrations publiques (Census of Governments) et un recensement de l'agriculture qui sont menés les mêmes années que les recensements économiques afin de faciliter le traitement et de faire le couplage des données. Le Census Bureau effectue aussi un certain nombre d'enquêtes mensuelles, trimestrielles et annuelles dans presque tous ces secteurs économiques.

Comme dans le cas des enquêtes démographiques, il y a un certain taux de non-réponse au questionnaire des recensements et enquêtes économiques. Dans la plupart des cas, les données omises sont imputées d'après a) des réponses déjà fournies par le non-répondant, b) des données tirées de dossiers administratifs et c) des liens établis entre diverses données élémentaires. Plutôt que de déclarer le pourcentage d'unités qui n'ont pas répondu, on indique habituellement le niveau de non-réponse dans un recensement ou une enquête économique sous forme de pourcentage d'un ou de plusieurs totaux imputés pour des questions. On désigne ces pourcentages les taux d'imputation.

Des explications sur les méthodes de compensation de la non-réponse au questionnaire utilisées pour cinq des six recensements économiques sont données à la section 4.1. Dans la section 4.2, on traite des méthodes de compensation de la non-réponse dans trois enquêtes économiques et, à la section 4.3, des méthodes équivalentes utilisées dans le recensement de l'agriculture. Les travaux de recherche et d'évaluation portant sur les méthodes de compensation de la non-réponse dans les recensements et enquêtes économiques sont examinés à la section 4.4.

Pour des explications plus détaillées sur les méthodes de compensation de la non-réponse dans ces recensements et dans plusieurs enquêtes connexes, voir Bailey, Chapman et Kasprzyk (1985).

4.1 Les recensements économiques

La base de sondage utilisée pour les recensements économiques est la Standard Statistical Establishment List (SSEL), un fichier informatique tenu à jour par le Census Bureau. La SSEL est un répertoire de tous les établissements employeurs déclarés par les entreprises employeuses à établissements multiples dans le cadre de la Company Organization Survey (COS) du Census Bureau, ainsi que de toutes les entreprises à établissement unique qui ont produit une déclaration d'impôt auprès de l'IRS. La COS est une enquête annuelle portant

sur les entreprises à établissements multiples. Les entreprises qui emploient au moins 50 personnes participent à cette enquête chaque année, tandis que les entreprises qui emploient moins de 50 personnes y participent tous les trois ans. On envoie à chaque entreprise visée une liste des établissements qu'elle a déclarés au cours de l'enquête la plus récente et on lui demande de la mettre à jour. L'entreprise doit aussi déclarer, pour chaque établissement, les effectifs au cours du premier trimestre de l'année précédente ainsi que le montant total de la paye versée l'année précédente. Pour les recensements économiques, on envoie à chaque établissement (par le biais de son siège social) qui figure dans le fichier SSEL, sauf les petites entreprises à établissement unique, un questionnaire de recensement qui a été conçu selon les caractéristiques des diverses classes de la classification américaine des activités économiques (code SIC).

Bien que les méthodes de compensation de la non-réponse au questionnaire pour les six secteurs commerciaux se ressemblent, il existe certaines différences importantes. Dans la description suivante des méthodes de compensation utilisées pour cinq des six recensements économiques, nous avons groupé les secteurs commerciaux pour lesquels on utilise essentiellement la même méthode:

a) Commerce de détail, commerce de gros, services
b) Fabrication, industries extractives

4.1.1 Commerce de détail, commerce de gros, services

Ces trois recensements économiques sont souvent considérés comme formant ensemble le recensement du commerce. Pour ces secteurs commerciaux, on recueille, relativement à l'année de recensement, des données sur les ventes et les recettes, l'emploi et la paye. Le taux d'imputation pour les ventes et les recettes varie de 10 à 15% dans le commerce de gros et de détail, et il est d'environ 20% dans les services.

Les réponses, pour tout établissement qui ne fournit pas les données requises dans le recensement, sont généralement imputées à partir des données fournies dans la déclaration d'impôt et stockées dans les fichiers du Internal Revenue Service (IRS). Pour les renseignements sur la paye, l'IRS dispose, à partir des déclarations d'impôt, de données pour quatre trimestres pour chaque numéro d'identification de l'employeur (numéro EI). Une entreprise peut avoir plusieurs numéros EI. Pour obtenir les données sur la paye d'une entrepise particulière, il faut additionner les chiffres de paye pour chacun des numéros EI utilisés par cette entreprise. On peut aussi obtenir, à partir des dossiers de l'IRS, le nombre d'employés par trimestre pour chaque numéro EI; ces chiffres peuvent aussi être groupés par entreprise. Pour les ventes et les recettes, l'IRS dispose de diverses déclarations d'impôt utilisées selon que l'entreprise non répondante est une entreprise individuelle, une société en nom collectif ou une société constituée.

L'imputation se complique du fait que l'unité de recensement est différente de l'entité fiscale de l'IRS. Pour le recensement du commerce, l'unité déclarante est l'établissement (c.-à-d. un local d'affaires). Pour l'IRS, l'unité déclarante est un numéro EI. Il se peut qu'un établissement ou plus produisent leurs déclarations d'impôt selon le même numéro EI. Si une entreprise non répondante n'a qu'un local d'affaires (c.-à-d. s'il s'agit d'une entreprise à établissement unique), il n'y aura qu'un seul numéro EI et l'imputation se fait sans problème particulier. Cependant, dans le cas d'une entreprise non répondante à établissements multiples, l'imputation est plus complexe puisque, en général, nous ne disposons pas de données de l'IRS sur chaque établissement. Dans un tel cas, on doit tout d'abord déterminer la structure de l'entreprise en se reportant au SSEL afin d'obtenir une liste de tous les établissements de cette entreprise et tous les numéros EI que cette dernière utilise. On obtient le total de chaque élément d'information sur une entreprise en additionnant les chiffres déclarés pour tous les numéros EI utilisés par cette entreprise, comme nous l'avons men-

tionné plus haut. Le total relatif à l'entreprise est ensuite réparti entre les établissements proportionnellement au total obtenu à partir des données les plus récentes tirées d'une enquête annuelle ou mensuelle. S'il n'existe pas de données de ce genre, on procède à une répartition égale. S'il y a non-réponse pour seulement une partie des établissements d'une entreprise à établissements multiples, les données sur les établissements non répondants sont imputées selon les ratios qui existaient l'année précédente.

4.1.2 Fabrication, industries extractives

Dans ces deux recensements économiques, on obtient des renseignements généraux sur le nombre d'employés, les heures travaillées et les niveaux de production selon les codes SIC à quatre chiffres. Les taux d'imputation varient d'environ 10 à 15%. Les méthodes de compensation de la non-réponse au questionnaire varient selon le genre d'entreprise qui n'a pas répondu (c.-à-d., entreprise à établissement unique ou entreprise à établissements multiples) et selon qu'on dispose ou non d'un enregistrement pour l'année précédente. Ainsi, il y a quatre cas différents de non-réponse. Les méthodes de compensation de la non-réponse dans ces quatre cas sont:

1) Entreprise à établissement unique, les données de l'année antérieure sont tirées du Annual Survey of Manufactures.

   Dans ce cas, on obtient le chiffre de paye annuel à partir des déclarations d'impôt et on le compare au chiffre déclaré l'année précédente. Cette comparaison permet de déterminer le pourcentage de variation par rapport à la période antérieure. On applique ensuite ce pourcentage à tous les éléments d'information de l'enregistrement antérieur pour obtenir un enregistrement imputé (sauf pour l'emploi et la valeur des livraisons quand ces dernières valeurs peuvent être obtenues de l'IRS.)

2) Entreprise à établissement unique, aucune donnée relative à l'année antérieure.

   Dans ce cas, on utilise la paye comme "valeur de départ" pour calculer un ensemble de ratios entre les données du recensement, pour chaque entreprise codée à quatre chiffres dans la SIC. C'est-à-dire qu'on calcule des rapports de telle façon qu'on puisse imputer toutes les données à partir de ces rapports soit directement, soit indirectement, si on obtient le chiffre de paye. Les rapports particuliers sont obtenus à partir de données chronologiques déclarées par les répondants dans la même industrie. La valeur de la paye (valeur de départ) est alors tirée des dossiers fiscaux (IRS) et toutes les autres données sont imputées à partir des rapports obtenus.

3) Entreprise à établissements multiples, on dispose des données de l'année antérieure.

   On calcule tout d'abord, pour chaque entreprise codée à quatre chiffres, un facteur global de croissance entre la période antérieure et la période courante, à partir de sources externes pour chacune des données clés suivantes: paye, emploi, variation des stocks et variation des dépenses en immobilisations. Ces quatre facteurs de croissance sont appliqués aux données correspondantes de l'année précédente sur chaque établissement afin d'obtenir des réponses imputées pour la période en cours. Ces quatre données imputées sont alors utilisées comme "valeurs de départ" pour imputer d'autres données.

4) Entreprise à établissements multiples, aucune donnée de l'année précédente sur l'entreprise.

   Dans ce cas, on obtient les données de base sur la paye et l'emploi de chaque établissement à partir du SSEL dont on a parlé plus tôt à la section 4.1. Comme nous l'avons indiqué, le fichier SSEL contient des données sur l'emploi et la paye relativement à tous les établissements visés par la COS. Puis, en utilisant les données du SSEL comme base, on impute l'enregistrement de données sur chaque établissement à partir des rapports calculés entre les données du SSEL et les autres données du recensement. Cette procédure est analogue à celle utilisée pour le cas 2) ci-dessus.

4.2 Enquêtes économiques

Le Census Bureau effectue un grand nombre d'enquêtes économiques mensuelles, trimestrielles et annuelles, en plus des recensements économiques. Des enquêtes mensuelles ou annuelles sont menées pour la plupart des six secteurs commerciaux du recensement. Les méthodes utilisées pour la compensation de la non-réponse au questionnaire dans le Monthly Retail Trade Survey (enquête mensuelle sur le commerce au détail) et le Truck Inventory and Use Survey (enquête sur le camionnage) sont résumés plus loin. On ne décrit pas la méthode de compensation de la non-réponse au questionnaire utilisé pour l'Annual Survey of Manufacturers (enquête annuelle auprès des fabricants, ASM) parce que c'est virtuellement identique à celle qui est utilisée pour le Census of Manufacturers, dont la description a été donnée à la section 4.1.2. Les taux d'imputation pour l'ASM varient de 5 à 10%.

4.2.1 Enquête mensuelle sur le commerce de détail (Monthly Retail Trade Survey)

L'enquête mensuelle sur le commerce de détail comprend environ 30,000 unités déclarantes dont près de 3,000 sont choisies systématiquement et 27,000 aléatoirement. Les unités choisies systématiquement doivent participer à l'enquête chaque mois, alors que ce n'est le cas que pour un tiers des unités choisies aléatoirement. Il faut donc expédier un questionnaire par la poste à environ 12,000 unités déclarantes chaque mois. Dans le cas d'une entreprise à établissements multiples visée par l'enquête, c'est un sous-échantillon des établissements de cette entreprise qui est choisi pour être inclus dans l'enquête. Les chiffres mensuels de ventes au détail sont les seules données recueillies dans l'enquête. Le taux d'imputation pour ces ventes est d'environ 11%.

Si une entreprise à établissement unique choisie systématiquement ou un établissement sélectionné d'une entreprise à établissements multiples ne fait pas de déclaration un mois donné, on impute une valeur pour les ventes en multipliant les données du mois précédent par une facteur dit "rapport d'éléments identiques". Pour obtenir ce rapport de correction, il faut diviser la somme pondérée des ventes du mois courant par la somme pondérée des ventes du mois précédent pour tous les établissements de la même cellule de correction qui ont déclaré des ventes pour le mois courant et le mois précédent. Les cellules de correction sont généralement définies d'après les trois premiers chiffres (ou les quatre premiers chiffres dans quelques cas) du code SIC, selon le genre d'établissement (c.-à-d. si ce dernier appartient ou non à une grosse entreprise à établissements multiples), et selon la classe de l'échelle des ventes. Le poids utilisé pour chaque unité déclarante dans le calcul du facteur "rapport d'éléments identiques" est l'inverse de la probabilité de sélection de l'unité déclarante.

Si une entreprise à établissements multiples choisie systématiquement ne déclare pas de chiffres de ventes pour aucun de ses établissements, les valeurs des ventes sont imputées pour chaque établissement et pour l'entreprise comme dans le cas précédent: on applique le facteur "rapport d'éléments identiques", pour les cellules de correction correspondantes, aux données des ventes du mois précédent. Si une entreprise de ce genre déclare les ventes mensuelles courantes pour toute l'entreprise, les réponses imputées pour les établissements sont corrigées à l'aide d'un ratio de façon que le résultat soit compatible avec le total déclaré pour l'entreprise.

Pour les entreprises choisies aléatoirement, l'imputation des données se fait comme dans le cas des unités déclarantes choisies systématiquement, sauf qu'une étape supplémentaire est nécessaire car les entreprises choisies aléatoirement produisent une déclaration tous les trois mois seulement. La première étape consiste à imputer le chiffre des ventes du mois précédent d'une entreprise non répondante d'après la réponse fournie trois mois plus tôt. Pour ce faire, on doit multiplier les ventes déclarées trois mois auparavant par le "rapport d'éléments

ce faire, on doit multiplier les ventes déclarées trois mois auparavant par le "rapport d'éléments identiques" basé sur la somme pondérée des ventes au cours du mois précédent et la somme pondérée des ventes trois mois auparavant (cellule par cellule). Une fois faite l'imputation des ventes du mois précédent, la valeur des ventes du moiscourant est établie à partir de la valeur imputée pour le mois précédent, à l'aide de la méthode décrite relativement aux unités déclarantes choisies systématiquement.

Si une entreprise non répondante est incluse dans l'échantillon pour la première fois, on impute les ventes du mois précédent (dans le cas d'une entreprise choisie systématiquement) ou les ventes trois mois plus tôt (dans le cas d'une entreprise choisie aléatoirement) à partir des ventes déclarées au recensement le plus récent, si ces données existent. Lorsque l'entreprise non répondante n'a pas participé au recensement le plus récent, il s'agit alors d'une entreprise qui vient d'entrer dans l'échantillon, pour laquelle les données sur les ventes de deux mois ont généralement été fournies au moment où l'entreprise a été ajoutée à la base de sondage. Ces données sont alors désaisonnalisées et gonflées de façon à produire une valeur annuelle. L'imputation est alors effectuée comme si l'entreprise non répondante avait fourni la valeur de ses ventes lors d'un recensement.

4.2.2 Enquête sur le camionnage (Truck Inventory and Use Survey TIUS)

La TIUS est menée toutes les cinq années et vise à produire des données sur les caractéristiques physiques et l'exploitation des camions dans tout le pays. Ces caractéristiques comprennent le type de remorque (configuration du véhicule), les genres de produits transportés, le genre de carburant utilisé ainsi que le nombre de milles parcourus par véhicule au cours de l'année. L'univers de l'enquête est composé des immatriculations des camions dans les cinquante États et le district de Colombia. L'échantillon comprend environ 120,000 immatriculations de camions. Le taux de réponse des unités choisies pour participer à l'enquête est d'environ 75%.

Pour compenser la non-réponse au questionnaire, on augmente par pondération, séparément pour chaque classe de pondération, les réponses fournies par les répondants afin d'obtenir l'équivalent pour l'échantillon total. Les classes de pondération correspondent aux strates de l'échantillon qui sont composées de classifications combinées selon l'État et le type de caisse (5 catégories). La compensation de la non-réponse par pondération est basée sur le nombre de camions; dans chaque classe (strate) le poids initial de chaque répondant est multiplié par le ratio du nombre de camions dans la strate et de la somme des poids initiaux des répondants dans la strate.

De toutes les enquêtes économiques que nous avons étudiées, seule la TIUS utilise une procédure de correction de poids pour tenir compte de la non-réponse au questionnaire. Dans les autres enquêtes économiques, on dispose généralement d'autres sources de renseignements de base qui permettent de "créer" un enregistrement pour un non-répondant.

4.3 Recensement de l'agriculture

Le recensement de l'agriculture fournit des données sur les exploitations agricoles (cultures, élevages) et sur les activités connexes, à l'échelle nationale. C'est la source principale de statistiques agricoles et la seule source de données agricoles compatibles et comparables au niveau des comtés, des États et du pays.

La compensation de la non-réponse dans le recensement de l'agriculture est complexe parce qu'on ne peut utiliser le SSEL aussi efficacement dans ce cas que dans les autres secteurs économiques. La liste d'envoi postal pour le recensement de l'agriculture est composée à partir de plusieurs sources qui se répètent. La base de sondage peut contenir par conséquent un certain nombre d'éléments en double et elle contient toujours certaines entités non agricoles. Aussi, la méthode de compensation de la non-réponse doit tout d'abord permettre de déterminer et d'estimer l'importance de la correction à apporter avant qu'on ne puisse l'effectuer.

Pour le recensement de 1982, les non-répondants ont été classés comme "gros" ou "petits" selon que leurs ventes prévues étaient supérieures ou inférieures à $100,000. Un suivi téléphonique intégral a été effectué dans le cas de tous les gros non-répondants. Dans le cas des petits non-répondants, on les a stratifiés d'après d'autres caractéristiques de la liste d'envoi postal. On a fait le suivi par la poste et par téléphone auprès d'un échantillon de ces unités afin d'obtenir des estimations, par strate à l'intérieur de chaque État, du pourcentage de non-répondants qui sont effectivement des exploitations agricoles. Ces estimations ont ensuite été utilisées, avec les données sur les proportions de répondants visés par comté, pour produire des estimations du nombre d'exploitations agricoles non répondantes au niveau du comté pour chaque strate. Les poids d'un échantillon aléatoire de répondants par comté, qui était compatible avec le nombre estimé d'exploitations agricoles non répondantes, ont alors été multipliés par deux. Tous les autres répondants ont conservé leur poids unitaire.

4.4 Recherches relatives à la compensation de la non-réponse dans les enquêtes économiques

Les données de l'IRS tirées des déclarations d'impôt constituent probablement la source de renseignements la plus utile pour la compensation, par imputation, de la non-réponse au questionnaire dans les enquêtes économiques. Il peut exister certaines différences entre les données de l'IRS et celles recueillies dans les recensements économiques à cause des diverses définitions, formules et méthodes utilisées pour la collecte des données. Dans une étude (1984), Dyke a comparé les données administratives (IRS) utilisées pour l'imputation des ventes, des recettes, de la paye et de l'emploi dans le recensement du commerce de 1977 avec les réponses correspondantes obtenues pour un échantillon de suivi de non-répondants. Il a trouvé que, en général, les valeurs déclarées à l'enquête de suivi étaient supérieures à celles tirées des sources administratives. L'importance des écarts variait selon la question et, en outre, les écarts étaient plus prononcés dans le cas des entreprises à établissements multiples. Aussi, comparaisons additionnelles de ce genre s'imposent. Si des écarts systématiques sont notés, il sera alors possible d'élaborer des facteurs de compensation qui pourront être appliqués aux données de l'IRS.

Pour plusieurs recensements et enquêtes, un facteur "rapport d'éléments identiques" est calculé et appliqué aux chiffres d'une période antérieure afin de produire une valeur imputée pour la période courante. Il est possible que ce rapport calculé pour *tous* les cas dans l'échantillon qui ont répondu au questionnaire pour les deux périodes ne s'applique pas très bien aux non-répondants à certaines questions. Bailey (1986) a examiné l'utilisation d'autres moyens que le "rapport d'éléments identiques" pour imputer les valeurs manquantes, telles que la régression linéaire et la régression quadratique, à l'aide de divers ensembles de variables indépendantes.

Pour de nombreuses méthodes d'imputation pour la compensation de la non-réponse dans les enquêtes économiques, les cas dans l'échantillon, à la fois des répondants et des non-répondants, sont classés dans des cellules avant le calcul a) d'un certain type de rapport entre la période courante et la période antérieure, relativement à une question ou b) d'un certain type de relation entre les questions de l'enquête et les questions de base: paye, emploi et recettes. Un projet de recherche visant à étudier d'autres choix possibles pour la définition des cellules dans le cas de l'enquête mensuelle sur le commerce de détail vient d'être réalisé par Huang (1986). Elle a trouvé que, pour certains secteurs d'activité économique, une autre méthode de définition des cellules réduit considérablement l'erreur quadratique moyenne (EQM) des ventes estimées. De plus, elle a comparé la méthode d'imputation utilisée actuellement - les "rapports d'éléments identiques" - et trois autres méthodes, pour ce qui a trait au biais et à l'EQM. Une seule des méthodes évaluées a donné de meilleurs résultats que la méthode utilisée actuellement. Cependant, elle a conclu que les faibles avantages de la méthode optimale pourraient ne pas justifier les exigences additionnelles inhérentes à son application.

5. IMPUTATION POUR LES GAINS DANS LE CAS DE LA CURRENT POPULATION SURVEY

5.1 La méthode "hot deck" hiérarchique

La Current Population Survey (CPS) est une enquête permanente du Census Bureau qui vise environ 60,000 ménages américains chaque mois. La CPS, parrainée par le Bureau of Labor Statistics, cherche principalement à recueillir des renseignements sur la population active et sur l'emploi. Tous les ans, en mars, l'enquête de la CPS comporte un supplément sur le revenu. Entre 11 et 12% des ménages de l'échantillon ne répondent pas aux questions sur le revenu. Une procédure spéciale, la méthode "hot deck" hiérarchique, a donc été élaborée pour traiter les réponses manquantes par imputation.

Avec la méthode "hot deck" hiérarchique, les gains non déclarés sont pris en compte à partir de l'enregistrement des réponses d'une autre unité observée dans l'échantillon - un enregistrement donneur. Le but est de trouver un donneur dont les caractéristique pour cette enquête ressemblent à celles du non-répondant à une question. La première étape dans la recherche des enregistrements donneurs appropriés consiste à séparer tout l'échantillon, à part les cas de non-réponse totale au questionnaire, en cellules d'après une classification multiple basée sur un certain nombre de caractéristiques de l'enquête. Dans chaque cellule, on dresse une liste des répondants et des non-répondants à une question donnée. Des enregistrements donneurs tirés de la liste des répondants sont affectés systématiquement aux non-répondants, avec origine choisie au hasard. S'il y a plus de non-répondants que de répondants dans une cellule pour une question donnée, les réponses de certains ou peut-être de tous les répondants dans la cellule sont utilisées plus d'une fois. Il se peut que, dans certaines cellules, il y ait au moins un non-répondant mais aucun répondant à une question.

Pour éviter de se trouver avec des non-répondants sans enregistrements donneurs correspondants, la définition des cellules et le choix des enregistrements donneurs pour les non-répondants à une question sont effectués plusieurs fois. À chaque étape, on définit moins de cellules qu'à l'étape précédente. À l'étape finale, le nombre de cellules définies est suffisamment petit pour qu'on soit certain qu'il y aura des enregistrements donneurs dans chaque cellule. Pour définir les cellules à chaque étape, on combine les cellules utilisées à l'étape précédente. Chaque non-répondant à une question se voit donc attribuer au moins un enregistrement donneur. L'enregistrement donneur utilisé pour produire la valeur imputée est le premier qui est trouvé au cours des différentes étapes successives.

Le principal avantage que cette méthode hiérarchique offre, c'est qu'on peut définir un très grand nombre de cellules à la première étape, à cause de la série d'étapes successives utilisées. Chaque fois qu'on trouve un enregistrement donneur à la première étape, on apparie le non-répondant à une question ainsi que le donneur en fonction d'un grand nombre de caractéristiques définies pour l'enquête. Dans de tels cas, il est fort probable qu'on puisse avoir une bonne imputation. Dans d'autres cas, les non-répondants à une question ainsi que les donneurs seront appariés en fonction d'un peu moins de caractéristiques. Cette méthode hiérarchique vise à trouver les enregistrements donneurs d'une façon qui maximise le nombre de caractéristiques pertinentes qui sont appariées. Pour une description plus détaillée de cette méthode, voir Welniak et Coder (1980), Oh et Scheuren (1980a), ou David, Little, Samuhel et Triest (1986, section 2).

5.2 Évaluation de la méthode "hot deck" hiérarchique

Certaines études portant sur la méthode "hot deck" hiérarchique utilisée pour la CPS ont été effectuées par Welniak et Coder (1980); Oh et Scheuren (1980a et 1980b); Lillard, Smith et Welch (1982); et David et coll. (1986). Une des faiblesses de la méthode "hot deck" employée pour la CPS qu'on a signalé est le fait que les valeurs des enregistrements donneurs

peuvent être utilisés à maintes reprises, ce qui augmente la variance. On pourrait modifier la méthode de façon qu'on ne puisse utiliser les valeurs des enregistrements donneurs plus d'une ou deux fois; cependant, cette modification n'a pas été apportée. La méthode "hot deck" utilisée pour la CPS repose sur l'hypothèse de la non-réponse sans biais, selon laquelle la distribution des réponses pour une variable d'enquête est la même chez les répondants et les non-répondants qui sont dans la même cellule.

David et coll. (1986) ont élaboré plusieurs méthodes basées sur des modèles, qui pourraient être utilisées à la place de la méthode "hot deck" hiérarchique employée pour la CPS. Ils les ont évaluées ainsi que la méthode "hot deck" hiérarchique par rapport à l'erreur absolue moyenne et à l'erreur relative moyenne.

Ces évaluations ont été faites à partir d'un fichier apparié CPS-IRS. Avec la création de ce fichier, on a tenté d'apparier le fichier de la CPS de mars 1981 aux dossiers fiscaux de l'IRS pour 1980. En dépit des limites apparentes de la méthode "hot deck", cette méthode appliquée à la CPS a permis d'obtenir une erreur absolue moyenne et une erreur relative moyenne plus faibles que lorsqu'on a utilisé des méthodes basées sur un modèle. Cependant, les modèles avaient été construits pour environ 10% seulement de tout l'échantillon de la CPS utilisé pour élaborer la méthode "hot deck".

6. SOMMAIRE ET CHAMPS D'ÉTUDE ÉVENTUELS

Dans cet article on a décrit, principalement à l'aide d'exemples, les méthodes de compensation de la non-réponse s'appliquant aux questionnaires utilisés actuellement dans les enquêtes et recensements de l'U.S. Bureau of the Census. On a mis l'emphase sur le besoin d'effectuer d'autres études empiriques et théoriques tant dans le domaine démographique que dans le domaine économique afin d'établir des lignes directrices plus objectives pour a) concevoir des méthodes de compensation de la non-réponse et b) mesurer les effets de la non-réponse sur les résultats d'enquête, selon diverses conditions d'enquête. On a également mentionné quelque projets de recherche dont le travail est présentement amorcé. Toutefois, l'augmentation de recherche dans ce domaine serait fort souhaitable. Par exemple, jusqu'à quel point peut-on utiliser les données accessoires ainsi que les méthodes de modélisation et d'analyse des données pour déterminer les rapports fonctionnels nécessaires pour fournir une description "raisonnablement" précise de la structure de la réponse/non-réponse qui s'applique à une enquête donnée?

En général, la compensation pour la non-réponse n'est qu'une étape parmi plusieurs permettant de réduire la variance et le biais des résultats des enquêtes et recensements. La réduction de l'impact de la non-réponse que pourraient apporter de telles étapes fera l'objet d'études ultérieures. De plus, il faut poursuivre les travaux de recherche concernant les questions qui reviennent périodiquement telles que l'effet des poids utilisés pour compenser la non-réponse au questionnaire et de l'imputation pour compenser la non-réponse à une question sur les estimateurs de la variance complexes, sur les méthodes de modélisation utilisées pour déterminer les facteurs de compensation appropriés ainsi que l'efficacité de la combinaison de divers types de techniques de compensation de la non-réponse.

REMERCIEMENTS

Les auteurs tiennent principalement à remercier tous ceux qui ont fourni des renseignements concernant les méthodes de compensation de la non-réponse utilisées dans beaucoup d'enquêtes et recensements de l'U.S. Census Bureau. Plusieurs divisions du Census Bureau ont contribué à cet article dont: Agriculture Division, Business Division, Construction Statistics

Division, Economic Surveys Division, Governments Division, Industry Division, Statistical Methods Division et Statistical Research Division.

Les auteurs désirent également remercier Dr. Fritz Scheuren pour ses commentaires pertinents lors de la rédaction de cet article.

Finalement, les auteurs remercient Hazel Beaton, Alice Bell et Valerie Howard pour leur excellent travail de dactylographie.

BIBLIOGRAPHIE

ANDERSON, H. (1978). On nonresponse bias and response probabilities, *Scandinavian Journal of Statistics*, 6, 107-112.

BAILEY, L. (1986). A study of alternative imputation techniques for surveys in the Current Industrial Reports. Rapport interne du Census Bureau, le 24 décembre.

BAILEY, L., CHAPMAN, D.W., et KASPRZYK, D. (1985). Nonresponse adjustment procedures at the Census Bureau: A review. *Proceedings of the Bureau of the Census First Annual Research Conference*, 241-444.

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

DYKE, T.C. (1984). Evaluation of the use of administrative record data for establishments which were non-respondents to the 1977 Census of Wholesale Trade, Retail Trade, or Selected Services. Rapport interne: Statistical Research Division Report Series, Census/SRD/RR-84/08, U.S. Bureau of the Census.

HANSON, R. (1978). The Current Population Survey: Design and Methodology. Technical Paper No. 40, Washington, D.C.: U.S. Bureau of the Census, pp. 55-59.

HUANG, E.T. (1986). Comparison of different imputation procedures in the monthly retail trade survey. *Proceedings of the Survey Research Methods Section, American Statistical Association* (à paraître).

KALTON, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, 2, 303-314.

KALTON, G., et LEPKOWSKI, J., et LIN, T. (1985). Compensating for wave nonresponse in the 1979 ISDP research panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 372-377.

KALTON, G., McMILLEN, D. et KASPRZYK, D.(1986). Nonsampling error issues in the Survey of Income and Program Participation. *Proceedings of the Bureau of the Census Second Annual Research Conference*, 147-164.

KALTON, G., et MILLER, M. (1986). Effects of Adjustments for Wave Nonresponse on Panel Survey Estimates. *Proceedings of the Survey Research Methods Section, American Statistical Association* (à paraître).

KOBILARCIK, E.L., et SINGH, R.P., (1986). SIPP: Longitudinal estimation for persons' characteristics. *Proceedings of the Section on Survey Research Methods, American Statistical Association* (à paraître).

LILLARD, L., SMITH, J.P., et WELCH, F. (1982). What do we really know about wages: The importance of non-reporting and census imputation. *Journal of Political Economy*, 94, 489-506.

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

LITTLE, R.J.A., et SAMUHEL, M.E. (1983). Imputation models on the propensity to respond. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 415-420.

NELSON, D., McMILLEN, D., et KASPRZYK, D. (1985). An overview of the Survey of Income and Program Participation: Update 1. SIPP Working Paper Series No. 8401, U.S. Bureau of the Census.

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

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

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

PALMER, S., et JONES, C. (1967). A look at alternate imputation procedures for CPS noninterview. *Proceedings of the Social Statistics Section, American Statistical Association*, 73-80.

PAUL, E.C., et LAWES, M. (1982). Caractéristiques des ménages répondants et non-répondants dans l'enquête sur la population active. *Techniques d'enquête*, 8, 53-79.

POLITZ, A., et SIMMONS, W. (1949). An attempt to get the 'Not-At-Homes' into the sample without callbacks. *Journal of the American Statistical Association*, 44, 9-31.

ROSENBAUM, P., et RUBIN, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

THOMSEN, I., et SIRLING, E. (1983). On the causes and effects of nonresponse: Norwegian experiences. Dans *Incomplete Data in Sample Surveys,* Vol. 3, (Eds. W.G. Madow, I. Olkin, and D.B. Rubin), New York: Academic Press, 25-29.

U.S. Department of Commerce, Bureau of the Census (1977). National Crime Survey, national sample, survey documentation. Rapport du Bureau of the Census.

U.S. Department of Commerce, Bureau of the Census (1983). Cross-sectional weighting specifications for the first wave of the 1984 panel of the Survey of Income and Program Participation (SIPP). Note de service interne du Census Bureau adressée par C. Jones à T. Walsh, 25 novembre.

U.S. Department of Commerce, Bureau of the Census (1984a). Economic characteristics of households in the United States: Third Quarter 1983. *Current Population Reports, Series P-70, No. 1,* Washington, D.C: U.S. Government Printing Office.

U.S. Department of Commerce, Bureau of the Census (1984b). 1984 SIPP first wave weighting-first stage estimate factors and specifications for collapsing noninterview adjustment calls. Note de service interne du Census Bureau adressée par C. Jones à T. Walsh, 16 février.

U.S. Department of Commerce, Bureau of the Census (1984c). SIPP weighting: Subsequent wave cross-sectional - revised. Note de service interne de l'U.S. Bureau of the Census adressée par C. Jones à T. Walsh, 12 octobre.

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

YCAS, M., et LININGER, C. (1981). The income survey development program: Design features and initial findings. *Social Security Bulletin*, vol. 44, n° 11, novembre.