Techniques d'enquête, décembre 1998
Vol. 24, n° 2, pp. 159-169
Statistique Canada



# Traitement de la non-réponse du cycle deux de l'enquête nationale sur la santé de la population

JEAN-LOUIS TAMBAY, IOANA ŞCHIOPU-KRATINA, JACQUELINE MAYDA, DIANA STUKEL et SYLVAIN NADON[1]

RÉSUMÉ

L'Enquête nationale sur la santé de la population (ENSP) est l'une des trois principales enquêtes-ménages longitudinales que mène Statistique Canada à une grande échelle auprès de la population canadienne. Depuis vingt ans, tous les deux ans, on a suivi un panel constitué d'environ 17 000 personnes. Les données provenant de l'enquête sont utilisées pour des analyses longitudinales, même si l'un des objectifs important est la production d'estimations transversales. Pour chaque cycle, les panélistes fournissent des renseignements détaillés sur leur santé (S) pendant qu'au même moment, pour augmenter l'échantillon transversal, des données socio-démographiques et quelques renseignements sur la santé sont recueillis (G) auprès de tous les membres des ménages. Cette stratégie de collecte présente différents schémas de réponse pour les panélistes après deux cycles: GS-GS, GS-G*, GS-**, G*-GS, G*-G* et G*-**, où «*» indique une portion de données manquantes. Le présent article explique la méthodologie élaborée pour traiter ces types de non-réponse longitudinale de même que la non-réponse d'une perspective transversale. L'utilisation de facteurs de pondération pour la non-réponse et la création de cellules d'ajustement pour la pondération à l'aide de l'algorithme CHAID sont expliquées ici.

MOTS CLÉS: Enquêtes longitudinales; traitement de la non-réponse; algorithmes CHAID.

## 1. INTRODUCTION

En 1996-1997, Statistique Canada a terminé la collecte de données pour le cycle 2 de l'Enquête nationale sur la santé de la population (ENSP). Cette enquête longitudinale a été lancée en 1994 afin d'obtenir des renseignements complets sur l'état de santé de la population canadienne ainsi que sur les facteurs déterminants en matière de santé. La population de l'enquête comprend des résidents de ménages et d'établissements de santé de l'ensemble du pays. Dans les provinces, le questionnaire rempli par les ménages se compose de deux principaux volets qui sont administrés à l'aide d'interviews assistées par ordinateur. Le volet Généralités sert à recueillir des données socio-démographiques et des renseignements de base sur la santé pour chaque membre du ménage. Le volet Santé sert à obtenir des données plus détaillées sur la santé du membre du ménage choisi pour participer au groupe longitudinal.

Même si l'ENSP est une enquête longitudinale, ses objectifs incluent la production d'estimations transversales périodiques (Catlin et Will 1992). La méthodologie de collecte des données reflète les besoins à la fois sur les plans longitudinal et transversal. Les panélistes choisis pour le cycle 1, sont suivis tous les deux ans pendant une période de 20 ans. Les personnes résidant avec les panélistes durant ces périodes fournissent des données pour le volet Généralités qui sont utilisées pour l'estimation transversale. Le champ d'observation de la composante transversale de l'enquête se rétrécissant avec le temps, il faut procéder régulièrement à une «bonification» de l'échantillon. La première bonification est prévue pour le cycle 3, en 1998.

Le présent document explique la méthodologie utilisée dans le Cycle 2 pour traiter la non-réponse au niveau des ménages et des personnes (avec indicateur pour la non-réponse à une question). La méthodologie se fonde sur la repondération des répondants au sein des sous-populations appelées cellules de pondération afin de tenir compte de la non-réponse. La repondération est une méthode couramment utilisée pour le traitement de la non-réponse à une question. Le biais et la variance découlant de cette méthode ont été étudiés par Thomsen (1973), Oh et Scheuren (1983), Kalton et Kasprzyk (1986) et Little (1986), entre autres. Si les cellules de pondération sont définies de telle sorte que la non-réponse survient presque totalement de façon aléatoire au sein de chaque cellule, alors le biais dû à la non-réponse peut devenir négligeable. Dans la même veine, David, Little, Samuhel et Triest (1983) ont étendu à la non-réponse la théorie élaborée par Rosenbaum et Rubin (1983) dans le contexte du jumelage du degré de propension. Leurs résultats supposent que la repondération peut rajuster le biais de la non-réponse lorsque les cellules de pondération sont composées en fonction de la propension à répondre.

On trouve à la section 2, une vue d'ensemble du plan d'échantillonnage et des données de sortie pour les deux premiers cycles. La section 3 explique les stratégies de traitement de la non-réponse et leurs résultats. Quant à la section 4, on y trouve les remarques finales. Il faut noter que la méthodologie présentée se rapporte aux échantillons de ménages des provinces; cette méthodologie ne couvre par les échantillons des territoires et des établissements.

[1] Jean-Louis Tambay, Ioana Şchiopu-Kratina, Jacqueline Mayda, Diana Stukel et Sylvain Nadon, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, 16ième étage, Immeuble R.H. Coats, Parc Tunney, Ottawa, (Ontario), Canada, K1A 0T6.

## 2. VUE D'ENSEMBLE DU PLAN D'ÉCHANTILLONNAGE DE L'ENSP ET DES DONNÉES DE SORTIE

### 2.1 Plan d'échantillonnage du cycle 1

Le premier échantillon de ménages a été constitué en 1994 à l'aide de l'outil de sélection d'un échantillon élaboré pour l'Enquête sur la population active (EPA) et, dans la province de Québec, en utilisant les logements qui ont participé à une enquête sur la santé menée par Santé Québec l'année précédente. Dans les deux cas, les ménages ou les logements ont été choisis au hasard dans des échantillons stratifiés de groupes sélectionnés selon la méthode de la probabilité proportionnelle à la taille. Les groupes ont été organisés en échantillons répétés et d'une période de collecte pour tenir compte du caractère saisonnier et pour permettre l'estimation de la variance. Il y a eu deux périodes de collecte estivale (juin et août) et deux périodes de collecte hivernale (novembre et mars 1995).

La figure 1 illustre le mécanisme de sélection du panel appliqué à l'extérieur du Québec. Les ménages de l'échantillon ont été désignés aléatoirement comme des ménages «adultes» ou «enfants», et comme admissibles ou non au filtrage, avant la collecte. Le filtrage a provoqué une augmentation dans le panel du nombre d'habitants provenant de ménages plus nombreux qui seraient sous-représentés si l'on choisissait seulement un membre par ménage, particulièrement les enfants et les adolescents. Les ménages admissibles au filtrage ont été rejetés de l'échantillon si aucun de leur membre n'avait moins de 25 ans. Le filtrage n'a pas été utilisé au Québec puisque les données provenant de l'enquête provinciale sur la santé permettaient l'application de différents taux de sous-échantillonnage par type de ménage et par taille.

| Type d'unité de sondage | Caractéristique du ménage | Sélection du panel limitée à: |
|---|---|---|
| Ménages «enfants» admissibles au filtrage (AF) | Aucun membre < 25 ans | s.o. – ménage rejeté |
| | Pas d'enfant, quelques membres < 25 ans | N'importe quel membre |
| | Présence d'enfants | Enfants |
| Ménages «enfants» non AF | Pas d'enfants présents | N'importe quel membre |
| | Présence d'enfants | Enfants |
| Ménages «adultes» | Tous | Membres de plus de 12 ans |

**Figure 1.** Mécanisme de sélection du panel à l'extérieur du Québec

La classification en ménages «adultes» et «enfants» a été effectuée pour des raisons opérationnelles: le questionnaire sur la santé, pour les enfants, ne serait pas disponible avant les périodes de collecte hivernales. Pour ce qui est des ménages «adultes», qui pouvaient être interviewés en tout temps, les enfants de moins de 12 ans n'étaient pas admissibles dans ce groupe. Les ménages «enfants», même ceux des groupes de la période estivale, ont été interviewés lors d'une période de collecte hivernale. Si des enfants étaient présents dans ces ménages, alors la sélection était limitée à eux. Pour réduire les distorsions saisonnières sur la charge de travail relative à la collecte de données et sur la représentativité du groupe découlant de ces procédures, un nombre moins grand de ménages ont été classés ménages «enfants» dans les groupes de la période estivale et, sauf une exception mineure, le filtrage a été appliqué seulement aux ménages «enfants».

Les provinces souhaitant améliorer les estimations au niveau infraprovincial pouvaient financer l'ajout d'effectifs à l'échantillon. Trois provinces se sont prévalues de cette possibilité en augmentant la taille de l'échantillon dans des régions ciblées. En Colombie-Britannique, un échantillon additionnel de 800 ménages a été constitué dans un établissement local de santé en utilisant le système de génération aléatoire de numéros de téléphone (GANT). La taille prévue de l'échantillon total dans les provinces était d'environ 23 000 ménages après le filtrage.

Ce qui précède donne une indication générale du plan d'échantillonnage de 1994, et ces renseignements sont suffisants pour les besoins du présent document. Les lecteurs désireux d'obtenir une présentation plus précise de l'échantillon de 1994 peuvent consulter Tambay et Catlin (1995) ou Statistique Canada (1995).

### 2.2 Pondération et résultats pour le cycle 1

Le principal résultat de l'ENSP consiste en un ensemble de fichiers de microdonnées à grande diffusion (FMGD) des réponses individuelles confidentielles (des versions internes de ces fichiers dont certains renseignements sont supprimés pour assurer la confidentialité sont aussi créés). Pour 1994, un FMGD pour le volet Généralités (58 400 dossiers) et un FMGD pour le volet Santé (17 600 dossiers) contenant les données recueillies auprès de chacun des membres des ménages et des membres sélectionnés dans le groupe sans enfants respectivement ont été publiés (Statistique Canada 1995).

Les poids propres à chaque dossier contenu dans un FMGD ont été calculés en appliquant une série d'ajustements à un poids de base représentant les fractions de sondage inverse (FSI) des ménages. Les FSI sont calculées en multipliant les poids des FSI originales ou des échantillons de Santé Québec par l'inverse de la fraction de sondage du sous-échantillon utilisé par l'ENSP. Pour ne pas alourdir le texte, nous décrirons seulement les principaux ajustements utilisés à l'extérieur du Québec.

Les ajustements apportés aux poids pour le FMGD du volet Généralités comprennent: 1) un ajustement pour la non-réponse des ménages; 2) un ajustement pour la procédure de filtrage; 3) un ajustement pour la non-réponse

des personnes [au sein des ménages répondants]; 4) un ajustement simple pour la stratification a postériori. L'ajustement 2) n'a été appliqué qu'aux ménages qui n'ont aucun membre de moins de 25 ans. La formule utilisée est $1/(1 - r_s)$, où $r_s$ représente la fraction de sondage du sous-échantillon pour le filtrage appliqué dans la strate. L'ajustement pour la stratification a postériori a été effectué séparément pour chaque classification croisée province-âge groupe-sexe. Les pondérations résultant de toutes les étapes précédentes sont multipliées par le rapport entre la population connue et la population estimée au sein de la classification croisée. Les chiffres relatifs à la population connue sont en fait des projections découlant du recensement.

Les ajustements apportés relativement à la non-réponse au niveau des ménages et des personnes (11,3 % et 1,4 %, respectivement) ont été appliqués aux unités déclarantes puisque les non-répondants ont été exclus des FMGD. Si $w_i$ est le poids de l'échantillon d'une unité $i$, le poids ajusté de la non-réponse, $w_{\text{aj},i}$ se définit comme $w_{\text{aj},i} = w_i(\sum_{\text{tout}} w_i)/(\sum_{\text{rép}} w_i)$, où les sommes s'effecturont sur toutes les unités d'échantillonnage et toutes les unités déclarantes, respectivement, à l'intérieur des cellules de pondération pour l'ajustement de la non-réponse. En raison d'un manque de renseignements de la part des ménages non répondants, les cellules de pondération pour la non-réponse des ménages ne sont que de simples classifications croisées des strates de l'ENSP et de la saison (c'est-à-dire, les groupes d'été par opposition aux groupes d'hiver). Les cellules de pondération pour la non-réponse des individus, dont le taux est très faible, sont les classifications croisées province-âge-sexe qui ont été utilisées pour l'ajustement de la stratification a postériori.

Les ajustements apportés aux poids pour le FMGD du volet Santé comprennent: 1) un ajustement pour la non-réponse des ménages; 2) un ajustement pour la procédure de filtrage; 3) un ajustement pour le sous-échantillon des ménages «enfants» et des ménages «adultes»; 4) un ajustement pour la sélection des panélistes de l'enquête longitudinale; 5) un ajustement pour la non-réponse des panélistes 6) un ajustement pour la stratification a postériori. Les deux premiers ajustements sont exactement les mêmes que pour le FMGD du volet Généralités. Comme le FMGD du volet Santé n'inclut pas les membres du groupe qui sont des enfants, l'ajustement 3) compense pour les ménages de l'échantillon où des membres non-enfants ne sont pas admissibles comme panélistes. Conséquemment, l'ajustement ne s'applique qu'aux ménages avec enfants et équivaut à $1/r$, où $r$ représente la proportion de ménages « adultes » dans l'échantillon. L'ajustement 4) est l'inverse de la probabilité que l'on ait sélectionné le panéliste. Les ajustements pour la non-réponse des panélistes (3,9 %) et pour la stratification a postériori sont similaires à ceux des FMGD pour le volet Généralités, et utilisent les mêmes classifications croisées province-âge-sexe. Même si les panélistes enfants ne sont pas inclus dans les FMGD du volet Santé, aux fins de l'enquête longitudinale, leur pondération dans l'échantillon est obtenue comme ci-dessus en utilisant $1/(1 - r)$ au lieu de $1/r$ comme à l'étape 3.

### 2.3 Plan d'échantillonnage du cycle 2

Pour le cycle 2, l'accent a davantage été mis sur l'estimation longitudinale: aucune «bonification» de l'échantillon n'a été prévue avant le cycle suivant. L'échantillon principal était donc constitué d'environ 17 000 panélistes et de leur ménage actuel. Les panélistes ont été suivis et on leur a administré les volets Généralités et Santé du questionnaire, alors que les autres membres du ménage n'ont répondu qu'au volet Généralités. Aucun suivi n'a été effectué auprès des ménages non répondants de 1994. En Alberta, au Manitoba et en Ontario d'importants échantillons supplémentaires (ne faisant pas partie de l'échantillon principal) ont été obtenus par génération aléatoire de numéros de téléphone (GANT), ce qui a permis la production d'estimations transversales au niveau infraprovincial. Pour chaque ménage contacté par GANT, un membre de plus de 12 ans a été choisi pour remplir le volet Santé du questionnaire. En Alberta et au Manitoba, dans les ménages contactés par GANT qui avaient des enfants, un enfant a aussi été choisi pour remplir le volet Santé.

Nous faisons remarquer, qu'aux fins de l'étude transversale, l'échantillon principal ne couvre pas très bien les nouveaux arrivants dans la population comme les nouveaux-nés et les immigrants récents. La population à qui on a administré le volet Généralités du questionnaire est constituée des résidents des ménages où au moins un des membres était admissible au cycle 1; les ménages composés entièrement de récents immigrants (et de leurs nouveaux-nés) sont donc exclus. La population à qui on a administré le volet Santé du questionnaire est constituée de personnes qui étaient admissibles au cycle 1: les immigrants récents et les enfants de moins de deux ans sont exclus de la population cible de l'échantillon principal (ils sont inclus dans la population cible pour la GANT). Pour les deux questionnaires, Généralités et Santé, la stratification après sélection est effectuée en utilisant des chiffres sur la population qui n'excluent pas les récents immigrants. Il en résulte que la population de récents immigrants est implicitement estimée par la population de non-immigrants parce que les dernières pondérations de l'échantillon principal sont ajustées à la hausse pour rendre compte des chiffres précédents. Il s'agit d'une contrainte qui est indiquée dans la documentation des FMGD. D'autres méthodes auraient pu être envisagées comme la stratification a postériori en utilisant uniquement la population non immigrante ou en ajustant d'une façon quelconque seulement les pondérations des immigrants moins récents (qui sont couverts) pour tenir compte des immigrants plus récents (qui ne le sont pas). Ces méthodes auraient été difficiles à appliquer lorsque, dans le cadre du questionnaire Généralités, une distinction doit être établie entre les ménages composés uniquement d'immigrants et les ménages mixtes dont certains membres sont des immigrants.

### 2.4 Données de sortie et pondération du cycle 2

La figure 2 constitue un sommaire des trois principaux fichiers de sortie prévus pour le cycle 2: un FMGD pour l'étude longitudinale; un FMGD pour le volet Santé de l'étude transversale et un FMGD pour le volet Généralités de l'étude transversale. Le FMGD de l'étude longitudinale contient des données sur les volets Généralités et Santé pour les deux cycles pour les 17 000 panélistes [nota: pour des raisons de confidentialité, il est possible que nous ne puissions publier le FMGD de l'étude longitudinale – auquel cas seul un fichier interne de microdonnées serait produit]. Le FMGD du volet Santé de l'étude transversale contient les renseignements de 1996 sur les volets Généralités et Santé pour environ 70 000 panélistes et membres sélectionnés par GANT. Le FMGD du volet Généralités de l'étude transversale contient les renseignements de 1996 sur le volet Généralités pour environ 220 000 membres de l'échantillon principal et du groupe GANT. Les procédés de pondération propres à chaque FMGD, présentés ci-après pour l'échantillon principal, sont décrits de façon plus détaillée dans l'étude de Stukel, Mohl et Tambay (1997).

| Fichier de sortie | FMGD ÉTUDE LONGITUDINALE | FMGD – SANTÉ ÉTUDE TRANSVERSALE | FMGD – GÉNÉRAL ÉTUDE TRANSVERSALE |
|---|---|---|---|
| Contenu | Généralités et Santé | Généralités et Santé | Généralités seulement |
| Échantillons | Principal seulement | Principal et GANT (3 prov.) | Principal et GANT (3 prov.) |
| Unités | Panéliste | Panélistes et GANT | Tous les membres des ménages |
| Taille | ≈ 17 000 dossiers | ≈ 70 000 dossiers | ≈ 220 000 dossiers |
| Pondération (échantillon principale) | 1. Pondération pour l'année de base<br>2. Ajust. panélistes non répondants<br>3. Stratification a postériori | 1. Pondération pour l'année de base<br>2. Ajust. panélistes non répondants<br>3. Intégration échant. principal et GANT<br>4. Stratification a postériori | 1. Pondération pour l'année de base<br>2. Ajust. ménager non répondants<br>3. Ajust. répartition du poids<br>4. Ajust. membres des ménages NR<br>5. Intégration échant. principal et GANT<br>6. Stratification a postériori |

**Figure 2.** Description des fichiers de sortie pour le cycle 2

On obtient les pondérations pour le FMGD de l'étude longitudinale en ajustant la pondération pour l'année de base d'abord pour les non-réponses du groupe de 1996, puis pour la stratification a postériori. La pondération pour l'année de base représente la fraction de sondage inverse pour 1994, y compris tous les ajustements pour le FMGD du volet Santé décrits à la section 2.2 jusqu'à l'ajustement 4) pour la sélection du panel (une correction est nécessaire pour le «retrait»des additions provinciales à l'échantillon de 1994). Le facteur de compensation de la non-réponse est traité dans la prochaine section et y sera décrit. La stratification a postériori est effectuée pour reproduire les chiffres de 1994 pour la population des provinces selon l'âge et le sexe.

Pour ce qui est du FMGD du volet Santé de l'étude transversale, le processus de pondération pour les panélistes (échantillon principal) nécessite trois ou quatre étapes. Habituellement, la pondération de l'année de base est ajustée pour la non-réponse des panélistes, selon les explications qui se trouvent dans la prochaine section, ainsi que pour la stratification a postériori (pour établir la correspondance avec les chiffres de population régionaux ou provinciaux de 1996, selon l'âge et le sexe). Dans les provinces où une partie de l'échantillon provient de la GANT, l'étape supplémentaire consiste à intégrer l'échantillon composé de répondants sélectionnés par la GANT. On obtient l'estimation intégrée par une forme d'adaptation minimale de l'estimateur double Skinner-Rao (Skinner Rao 1996).

Quant au FMGD du volet Généralités de l'étude transversale, le processus de pondération pour l'échantillon principal nécessite cinq ou six étapes. Premièrement, une fois de plus, il faut calculer la pondération de l'année de base. Il faut ensuite procéder à un ajustement pour la non-réponse au niveau des ménages, dont traitera la prochaine section. La prochaine étape consiste à appliquer la méthode de répartition du poids. Cette méthode a été décrite par Ernst (1989) et mise au point plus tard par Lavallée (1995). Le poids du panéliste divisé par le nombre de personnes de son ménage qui étaient admissibles au cycle 1 est attribué à tous les membres du ménage, y compris ceux qui n'étaient pas admissibles au cycle 1 (c'est-à-dire nouveaux-nés, immigrants). Cette méthode est non biaisée pour les estimations des totaux pour la population des ménages où au moins un membre était admissible au cycle 1. La prochaine étape consiste à établir l'ajustement pour la non-réponse d'un membre du ménage. Dans les provinces qui ont eu recours à la GANT, cette étape est suivie de l'intégration de l'échantillon principal avec l'échantillon GANT (cette fois pour tous les âges). La stratification a postériori est effectuée de la même façon que pour le FMGD du volet Santé de l'étude transversale.

## 3. STRATÉGIE RELATIVE À LA NON-RÉPONSE DANS L'ÉCHANTILLON PRINCIPAL DU CYCLE 2

La présente section explique la stratégie adoptée pour le traitement de la non-réponse du cycle 2 de l'échantillon principal (non GANT). L'ajustement pour la non-réponse a encore une fois été effectué à l'aide de la méthode de la cellule de pondération sauf que, cette fois-ci, les données du cycle 1 étaient disponibles, ce qui a permis de créer des cellules plus homogènes en regard de la propension à répondre. Dans la section 3.1, on définit les non-répondants de l'ENSP. Dans la section 3.2, on explique deux méthodes générales utilisées pour créer des cellules de pondération, dont celle choisie pour l'ENSP. La stratégie d'ajustement pour la non-réponse est expliquée dans la section 3.3 alors qu'à la section 3.4, on décrit la création des cellules de pondération de la non-réponse.

### 3.1 Définition des non-répondants et des unités non admissibles

La première étape du traitement de la non-réponse consiste en sa définition. Pour le cycle 2, les questionnaires ont été complètement remplis par 89 % de l'échantillon principal et partiellement remplis par 3 % de l'échantillon. Le reste de l'échantillon se compose de refus (3,1 %), de cas où le panéliste n'a pu être retracé (1,7 %), de décès (1,7 %) de départs du Canada ou de placements en établissement (0,4 %), ainsi que d'autres types de non-réponse comme les absences temporaires et les circonstances spéciales (0,7 %). Parmi les ménages répondants, le taux de non-réponse au niveau des personnes est très faible: 1,8 % pour le questionnaire Généralités et 1,1 % pour le questionnaire Santé. Nous définissons d'abord les cas qui sont des non-réponses aux fins des études longitudinale et transversale.

Aux fins de l'étude longitudinale, un décès est considéré comme une réponse valide à l'enquête. Les panélistes qui sont décédés avant le cycle 2 voient leur état enregistré comme tel et la portion 1996 des données qui les concernent codée comme «sans objet» dans le fichier de données de l'étude longitudinale. Les panélistes qui sont déménagés dans un établissement ou dans les territoires ont fait l'objet d'un suivi et leurs réponses ont été utilisées aux fins de l'étude longitudinale. Les panélistes qui ont quitté le pays n'ont pas fait l'objet d'un suivi, mais ils ont été traités comme des non-répondants à l'étude longitudinale, même s'il aurait été plus exact aux fins de certaines analyses de les considérer comme ne faisant plus partie du champ de l'enquête. Nous avons choisi ce traitement parce que si ces personnes revenaient au Canada, elles seraient de nouveau admissibles.

Aux fins de l'étude transversale, tous les cas présentés au paragraphe précédent ont été traités comme des situations ne faisant pas partie du champ de l'enquête. Ce choix est acceptable parce que des outils d'enquête distincts pour les établissements et pour les territoires ont été utilisés pour la couverture transversale de ces populations en particulier, qui font partie du champ de l'enquête. Les unités non admissibles ne font pas partie de l'ENSP, mais, comme elles représentent d'autres unités de ce genre, elles ont été traitées aux fins de la pondération comme des répondants à toutes les étapes de l'ajustement de la pondération, sauf celle de l'intégration et celle de la stratification a postériori.

Les refus et les cas où des questionnaires étaient manquants pour des raisons autres que celles données précédemment ont été définis comme des non-réponses. Comme nous le verrons, une distinction a été faite ultérieurement entre les non-répondants complets ou partiels afin de tenir compte de différents utilisateurs.

### 3.2 Méthode de création de cellules de pondération pour l'ajustement de la non-réponse

Il existe deux approches statistiques à la création de cellules de pondération de la réponse qui font appel à la modélisation par segmentation et à la régression logistique. On trouve un exemple de cette dernière approche dans Hirabayashi, Little et Rubin (1992). Les auteurs ont établi des estimations de l'impôt anticipé à partir de déclarations hâtives en créant des cellules de pondération pour l'ajustement à partir de niveaux de propension à remettre une déclaration hâtive. La régression logistique permettait d'estimer la propension des déclarants à soumettre précocement leur déclaration. L'Enquête longitudinale sur la dynamique du travail et du revenu (EDTR) constitue un autre exemple d'application de la régression logistique (Grondin 1996). On calcule la régression des indicateurs de réponse des unités d'échantillonnage selon des caractéristiques connues (dichotomiques). On a créé les cellules d'ajustement de la non-réponse par recoupement des unités d'échantillonnage, à l'aide de toutes les covariables significatives. Afin de tenir compte des exigences quant à la taille minimale des cellules et aux taux de réponse, on a dû fusionner des cellules, à partir de celles partageant toutes les covariables sauf les moins significatives.

Dans l'approche de modélisation des segments, on génère un arbre de décision à partir des données; pour ce faire, on divise les ensembles de données en séquence de façon que, à chaque noeud, le prédicteur le plus significatif pour la variable de la réponse serve de point de départ à la division suivante. La division se poursuit jusqu'à ce qu'il ne soit plus possible de trouver une variable significative pour effectuer une autre division, ou qu'il ne soit plus possible de respecter les exigences quant à la taille minimale de la cellule. L'une des premières applications de la méthode de modélisation des segments pour l'ajustement de la non-réponse portait sur la Panel Study of Income Dynamics (Institute for Social Research 1979). En raison de ses avantages, énoncés ci-après, on a adopté pour l'ENSP l'approche de modélisation des segments fondée sur l'algorithme CHAID (détection automatique d'interactions du chi carré), inventé par Kass (1980). L'algorithme CHAID utilise des tests du $\chi^2$ pour diviser les prédicteurs catégoriques et retenir la division la plus significative à chaque étape. La division, en deux catégories ou plus, n'est pas exécutée de la même façon pour les prédicteurs ordonnés et non ordonnés. L'algorithme CHAID a été appliqué à l'aide du logiciel Knowledge Seeker (ANGOSS Software, 1995). Il faut noter que le logiciel applique l'algorithme aux prédicteurs continus en les traduisant tout d'abord en variables discrètes ordonnées.

Diverses études traitent des avantages et des inconvénients des approches logistiques et CHAID (se reporter par exemple à Kalton et Kasprzyk 1986). Le fondement théorique de la régression logistique est bien connu de nombreux analystes, et la programmation est possible dans plusieurs logiciels d'analyse statistique normalisée. Cette méthode permet entre autres d'obtenir des estimations individuelles de la propension à répondre, que l'on applique directement pour ajuster les poids des répondants. Cependant, pour arriver à une vitesse d'exécution raisonnable du

programme, il vaut mieux limiter le nombre de variables et de termes d'interaction. La fusion des cellules peut aussi s'avérer complexe, comme ce fut le cas pour l'EDTR, déjà mentionnée. Au contraire, l'algorithme CHAID permet de traiter un grande nombre de covariables et, en raison de la structure en arbre de décision, il est facile d'établir les interactions. Qui plus est, il est facile d'incorporer les exigences quant à la taille minimale des cellules comme paramètres d'exécution de programme. Parmi les principaux inconvénients, citons le fondement théorique moins familier (le logiciel Knowledge Seeker est vendu comme un outil d'«exploration des données»), la documentation limitée et le peu de logiciels d'exécution. À ceux-là s'ajoute l'impossibilité d'incorporer le plan d'échantillonnage pour adapter les modèles logistiques aux données de l'enquête, comme c'est le cas pour certains programmes statistiques tels que SUDAAN et PC CARP. Pour pallier à ces contraintes dans l'ENSP, on a inclus comme prédicteurs des caractéristiques liées au plan d'échantillonnage (se reporter à la section 3.4).

Deux études empiriques comparant les approches logistique et CHAID pour l'analyse de la non-réponse obtiennent des résultats divergents. Ainsi, Rizzo, Kalton et Brick (1996) n'ont pas conclus à des différences notables entre les deux approches appliquées à la Survey of Income and Program Participation. Par contre, Dufour, Gagnon, Morin, Renaud et Särndal (1998), dans une étude simulée de l'EDTR, ont trouvé que l'application de l'approche CHAID pour l'ajustement de la non-réponse produisait un biais moins important.

### 3.3 Ajustement de la non-réponse dans l'échantillon principal

Il a fallu établir une méthode d'ajustement de la non-réponse pour chacun des FMGD, soit les fichiers de données longitudinales, Généralités (transversales) et Santé (transversales). Le FMGD Généralités sera traité en premier.

Comme le démontre la figure 2, la stratégie de pondération appliquée au FMGD Généralités a exigé d'ajuster séparément la non-réponse aux niveaux des ménages et des personnes. Quand on a créé des cellules d'ajustement de la non-réponse au niveau des ménages, on a considéré les caractéristiques des panélistes ainsi que celles du ménage comme étant des prédicteurs de non-réponse, pour trois raisons. Premièrement, étant donné que le panéliste repré sentait le lien avec le ménage dans le cycle 2, on considérait que ses caractéristiques avaient permis de trouver le ménage et d'obtenir une réponse (cette personne a souvent permis d'établir le premier contact). Deuxièmement, quelques caractéristiques personnelles du panéliste, telles que la race, constituent jusqu'à un certain point des caractéristiques du ménage. Enfin, l'utilisation des caractéristiques du panéliste n'était pas incompatible avec la nécessité qui nous incombait d'utiliser divers renseignements pour construire les cellules de pondération. Si les caractéristiques du panéliste ne sont pas significatives, alors l'algorithme CHAID ne les retient tout simplement pas.

Pour le volet Généralités, la non-réponse au niveau individuel indique qu'une information est disponible pour certains membres seulement d'un ménage, en raison d'un refus de l'un des membres ou d'une absence temporaire. Le taux de non-réponse au niveau individuel étant de 1,8 % seulement, il a été décidé que la création de cellules de pondération en fonction des catégories province, âge et sexe (comme dans le Cycle 1) suffirait amplement.

Contrairement au FMGD Généralités, on a pu combiner l'ajustement de la non-réponse aux niveaux des ménages et des personnes pour les FMGD longitudinal et Santé, parce qu'il est appliqué aux réponses d'une même personne (le panéliste). On a donc créé un seul ensemble de cellules d'ajustement.

Pour ce qui est du FMGD longitudinal, on a noté que les éléments de données provenaient à la fois des volets Généralités et Santé, mais que les taux de réponse différaient d'un volet à l'autre. Cet écart a entraîné des schémas de réponse différents pour les cycles 1 et 2, comme suit: GS-GS; GS-G*; G*-GS; G*-G*. À ces différences s'ajoutent les schémas de non-réponse longitudinale GS-** et G*-** (les lettres correspondent aux réponses pour les volets dans chaque cycle, et l'astérisque correspond à une non-réponse). Afin de maximiser l'utilité des données, on a décidé d'ajuster de deux façons la non-réponse longitudinale. L'une des méthodes d'ajustement est appliquée au schéma «réponse longitudinale complète» GS-GS. Autrement dit, tous les autres schémas de réponse sont considérés comme étant des non-réponses. L'autre méthode d'ajustement s'applique à la «réponse longitudinale partielle», pour les cas où on a obtenu au moins les données générales pour chacun des cycles (schémas GS-GS, GS-G*, G*-GS et G*-G*). Le groupe de données Réponse complète sera utile aux chercheurs qui souhaitent analyser un ensemble de données longitudinales complètes couvrant tout le contenu du questionnaire. Le groupe de données Réponse partielle peut servir aux chercheurs qui s'intéressent surtout aux types de variables du questionnaire général. On le constate dans le tableau ci-dessous, la Réponse longitudinale partielle dépasse de 3 % seulement la Réponse longitudinale complète.

**Tableau 1**
Schémas de la réponse longitudinale

| Type de réponse | | Schéma de réponse Cycles 1-2 | Nombre de dossiers |
|---|---|---|---|
| Complète | Partielle | | |
| ■ | ■ | GS-GS | 15 670 |
| | ■ | GS-G* | 110 |
| | ■ | G*-GS | 366 |
| | ■ | G*-G* | 22 |
| | | GS-** | 1 014 |
| | | G*-** | 94 |
| | | Total | 17 276 |

À la lumière des données du tableau, il faut construire des cellules d'ajustement pour cinq types de réponse (ou de non-réponse) dans le cycle 2, comme suit:

- FMGD Généralités – réponse des ménages
- FMGD Généralités – réponse des personnes
- FMGD Santé – réponse combinée
- FMGD longitudinal – réponse complète
- FMGD longitudinal – réponse partielle

Seulement trois groupes de cellules d'ajustement ont été créées pour ces types de réponse. Les cellules d'ajustement créées pour la réponse au niveau des ménages dans le FMGD Généralités ont aussi servi pour les réponses partielles dans le FMGD longitudinal, puisque la réponse partielle au niveau des ménages correspondait presque toujours à une réponse partielle au niveau des personnes dans le fichier longitudinal (sauf pour 53 individus). De même, les cellules d'ajustement créées pour la réponse complète dans le FMGD longitudinal ont été appliquées à la réponse dans le FMGD Santé. Bien qu'on ait obtenu 366 réponses de plus pour le dernier type (schéma G*-GS), on a considéré que le même mécanisme de réponse était à l'oeuvre. Le troisième groupe de cellules d'ajustement a été appliqué à la réponse au niveau des personnes dans le FMGD Généralités. On a utilisé les catégories province – âge – sexe, comme ce fut le cas au cycle 1.

Incidemment, bien que les mêmes cellules d'ajustement puissent être utilisées pour différents groupes de données, on effectue des calculs distincts pour chaque type de groupe de données. Ainsi, les 366 dossiers assortis d'un schéma de réponse G*-GS sont traités comme répondants dans le processus d'ajustement des poids pour le FMGD Santé, mais comme non-répondants dans le processus d'ajustement des poids pour l'analyse de la réponse complète dans le FMGD longitudinal.

### 3.4 Création de cellules d'ajustement de la pondération

On a créé des groupes distincts de cellules d'ajustement de la pondération pour chaque province. Tout d'abord, il a fallu déterminer les variables à prendre en compte. Avec l'approche CHAID, le nombre de variables à considérer importe peu, et on a tenu compte de différents types de variables. Bien entendu, on a pris en compte les caractéristiques du ménage, ou du logement, ainsi que les caractéristiques personnelles du panéliste. Afin d'être en mesure d'incorporer le plan d'enquête dans l'analyse, nous avons aussi tenu compte de certaines caractéristiques liées à ce plan ou au poids d'échantillonnage. Parmi celles-ci se trouvent les variables géographiques telles que le code Région métropolitaine de recensement (RMR) ou l'indicateur urbain/rural, les variables d'échantillonnage particulières au cycle 1, telles que les indicateurs de ménages visés par le filtrage et le type de ménage «adulte/enfant», de même que les caractéristiques liées à l'application de ces variables d'échantillonnage, telles que la présence dans un ménage d'un membre âgé de moins de 25 ans ou d'un enfant. La taille du ménage a été considérée comme étant une caractéristique du ménage, et on l'a aussi liée au poids de l'échantillon. À la lumière des expériences passées, on a décidé d'inclure, outre la caractéristique revenu du ménage, une caractéristique fictive indiquant si le revenu du ménage avait été déclaré dans le cycle 1. Les changements d'adresse pouvant conduire à une incapacité de retracer la provenance d'une non-réponse, il aurait été utile d'intégrer un identificateur de changement d'adresse. Cependant, dans certains cas de non-réponse et d'absence de contact, il était difficile de déterminer si le panéliste avait vraiment déménagé. Enfin, nous avons utilisé pour l'analyse une variable «personne ayant déménagé», qui nous permet de savoir si le panéliste avait changé de province entre les cycles. Cette modalité représente un pis-aller parce que la valeur par défaut est «non». Des caractéristiques personnelles tirées du volet Santé de l'enquête, telles que le statut fumeur/consommateur d'alcool, le niveau d'indice de santé et l'échelle de santé/souffrance morale n'ont pas été utilisées parce qu'elles étaient inconnues pour quelque 500 panélistes.

On trouve ci-après la liste des variables qui ont été utilisées. La valeur de l'indicateur de non-réponse, soit la variable dépendante, a été assignée en fonction de la définition donnée au concept de non-réponse.

#### VARIABLES D'ÉCHANTILLONNAGE/GÉOGRAPHIQUES

| | |
|---|---|
| PROVINCE | Analyse effectuée à l'échelle provinciale |
| CMA | Région métropolitaine de recensement (0 s'il ne s'agit pas d'une RMR) |
| URBAN | Indicateur urbain/rural |
| REJECT | Indicateur qu'une unité (ménage) est sujette au filtrage |
| ACFLAG | Classe d'échantillonnage Adulte/enfant pour l'unité |

#### CARACTÉRISTIQUES DU LOGEMENT/MÉNAGE

| | |
|---|---|
| DWELL | Type de logement (10 catégories) |
| OWNER | Indicateur propriétaire/locataire |
| FAMTYP | Type de famille (personne seule, famille monoparentale, couple marié, autre) |
| INC | Suffisance du revenu du ménage (5 niveaux) |
| INCNR | Indicateur de non-réponse pour le revenu |
| INCSRC | Principale source de revenu (6 catégories) |
| *HHSIZE | Taille du ménage |
| UND25 | Indicateur membre de moins de 25 ans |
| KIDS | Indicateur enfant de moins de 12 ans |

## CARACTÉRISTIQUES PERSONNELLES DU PANÉLISTE

| | |
|---|---|
| SEX | Sexe |
| AGE | Âge en années |
| AGE16 | Indicateur personne âgée de 16 ans et plus |
| MARIT | État civil |
| FAMID | Identificateur famille à l'intérieur du ménage (A, B, C, ...) |
| RACE | Blanc, Noir, Autochtone ou autre |
| BORN | Lieu de naissance (Canada, É.-U./Mexique Amérique S./Afrique, Europe/Australie, Asie) |
| AGIMM | Âge au moment de l'immigration pour les immigrants |
| *MOVED | Indicateur changement de province (voir le texte) |
| *EDUC | Niveau de scolarité le plus élevé (12 catégories) |
| *STUDNT | Indicateur étudiant |
| MACT | Activité principale (8 catégories) |
| *NUMJOB | Nombre d'emplois occupés l'an dernier (cycle 1) |
| RESTR | Indicateur pour activité restreinte |
| *CAUSE | Cause principale de restriction (12 catégories) |
| CONSUL | Nombre de consultations auprès d'un médecin |
| INHOSP | Indicateur pour séjour de 1 nuit à l'hôpital |
| *CHRONIC | Nombre de maladies chroniques |

* Indique que la variable n'a jamais été significative lors de l'établissement de classes.

La figure 3 énonce les variables choisies par l'algorithme CHAID en vue de créer des cellules d'ajustement de la non-réponse pour la réponse au niveau des ménages et pour la réponse longitudinale complète dans chaque province. À des fins de confidentialité, on ne peut donner le détail sur les tailles des cellules individuelles et les taux de réponse (certaines des variables utilisées sont considérées comme étant sensibles et ne figurent pas dans les FMGD). Cependant, on trouvera de l'information sommaire sur la construction des cellules aux tableaux 2 et 3.

Les résultats varient d'une province à l'autre. Comme il était prévu, les provinces où l'échantillon est plus important, telles que l'Ontario, le Québec, la Colombie-Britannique et l'Alberta, ont permis de tracer des arbres de décision «plus riches». Les tailles des cellules ainsi que les taux de réponse varient aussi passablement. Dans le tableau 2, qui porte sur la réponse au niveau des ménages, une seule cellule est attribuée au Manitoba, et une seule cellule contient 88 % de l'échantillon du Nouveau-Brunswick. De même, dans le tableau 3, une grande partie de l'échantillon de Terre-Neuve est placée dans une seule des deux cellules. On constate pour quelques provinces des cellules avec des taux de non-réponse autour de 40 %.

La figure 3 montre les différences entre les caractéristiques des classes de pondération pour les différentes provinces, mais aussi pour les deux types de non-réponse à l'intérieur des provinces. Dans toutes les provinces, sauf en Alberta, l'algorithme CHAID utilise des caractéristiques distinctes pour les deux types de non-réponse, dès le premier ou le deuxième niveau d'embranchement. Quelques caractéristiques sont dominantes aux premiers niveaux d'embranchement dans de nombreux arbres, pour les deux types de non-réponse, soit le niveau de suffisance du revenu du ménage (INCNR), l'indicateur de non-réponse quant au revenu (INCNR), la race (RACE), et le lieu de naissance (BORN).

**Tableau 2**
Caractéristiques des cellules d'ajustement de la réponse (réponse au niveau des ménages)

| Prov. | $N^{bre}$ Unités | $N^{bre}$ NR | Tailles cellules | | | % par cellules NR | | |
|---|---|---|---|---|---|---|---|---|
| | | | min. | max. | moy. | min. | max. | moy. |
| T.-N. | 1 082 | 40 | 354 | 728 | 541 | 1,4 | 4,8 | 3,7 |
| Î.P.-E. | 1 037 | 51 | 81 | 478 | 259 | 3,0 | 13,6 | 4,9 |
| N-É. | 1 085 | 55 | 46 | 374 | 217 | 0,7 | 10,9 | 5,1 |
| N.-B. | 1 125 | 59 | 32 | 986 | 281 | 2,6 | 34,4 | 5,2 |
| QC | 3 000 | 133 | 123 | 2 363 | 750 | 1,8 | 12,1 | 4,4 |
| Ont. | 4 307 | 315 | 44 | 1 038 | 308 | 0,9 | 25,8 | 7,3 |
| Man. | 1 205 | 50 | 1 205 | 1 205 | 1 205 | 4,1 | 4,1 | 4,1 |
| Sask. | 1 168 | 59 | 37 | 626 | 167 | 1,6 | 35,3 | 5,1 |
| Alb. | 1 544 | 116 | 32 | 837 | 221 | 3,9 | 36,7 | 7,5 |
| C.-B. | 1 723 | 149 | 82 | 678 | 246 | 5,2 | 29,0 | 8,6 |

**Tableau 3**
Caractéristiques des cellules d'ajustement de la réponse (réponse longitudinale complète)

| Prov. | $N^{bre}$ unités | $N^{bre}$ NR | Tailles cellules | | | % par cellules NR | | |
|---|---|---|---|---|---|---|---|---|
| | | | min. | max. | moy. | min. | max. | moy. |
| T.-N. | 1 082 | 73 | 35 | 1 047 | 541 | 6,2 | 22,9 | 6,7 |
| Î.P.-E. | 1 037 | 80 | 41 | 453 | 207 | 4,1 | 26,8 | 7,7 |
| N.-É. | 1 085 | 96 | 236 | 555 | 362 | 6,5 | 14,3 | 8,8 |
| N.-B. | 1 125 | 86 | 59 | 819 | 281 | 4,8 | 16,8 | 7,6 |
| QC | 3 000 | 211 | 42 | 2 202 | 375 | 2,5 | 37,8 | 7,0 |
| Ont. | 4 307 | 470 | 34 | 619 | 196 | 0,0 | 38,0 | 10,9 |
| Man. | 1 205 | 91 | 186 | 763 | 402 | 5,6 | 15,1 | 7,6 |
| Sask. | 1 168 | 83 | 90 | 339 | 195 | 0,0 | 28,9 | 7,1 |
| Alb. | 1 544 | 148 | 41 | 866 | 172 | 1,1 | 39,0 | 9,6 |
| C.-B. | 1 723 | 192 | 33 | 408 | 191 | 4,5 | 37,3 | 11,1 |

Dans la figure 3a, le revenu du ménage et les variables liées (INCNR et INCSRC), le statut propriétaire/locataire (OWNER), la race, le lieu de naissance et le type de logement (DWELL) ont toutes été utilisées trois fois ou plus pour la création des classes de pondération de la non-réponse au niveau des ménages. On remarque aussi que dans cinq des neuf provinces, une caractéristique personnelle des panélistes

a été sélectionnée au premier embranchement par l'algorithme CHAID. Cette constatation vient confirmer qu'il a été judicieux de prendre en compte les caractéristiques personnelles pour ajuster la non-réponse au niveau des ménages.

Dans la figure 3b, Non-réponse longitudinale complète, la Région métropolitaine de recensement (CMA), l'état civil (MARIT) et le sexe, malgré leur importance moindre que les caractéristiques de revenu, de race et de lieu de naissance, ont été utilisées le plus souvent (cinq fois chacune).

Comme il a déjà été souligné, les variables d'échantillonnage telles que les indicateurs filtrage (REJECT) et ménages «Adulte/Enfant» (ACFLAG) ont été prises en compte afin que soit incorporé le plan d'enquête dans les analyses exécutées avec l'algorithme CHAID. Bien que ces variables aient été sélectionnées une fois seulement chacune, les caractéristiques du ménage utilisées dans le plan, telles que la présence d'enfants (KIDS) et de membres de moins de 25 ans (UND25) ont été sélectionnées à l'occasion. La taille du ménage n'a pas été utilisée, mais le type de famille (FAMTYP), une variable liée à la taille, a été sélectionné deux fois.

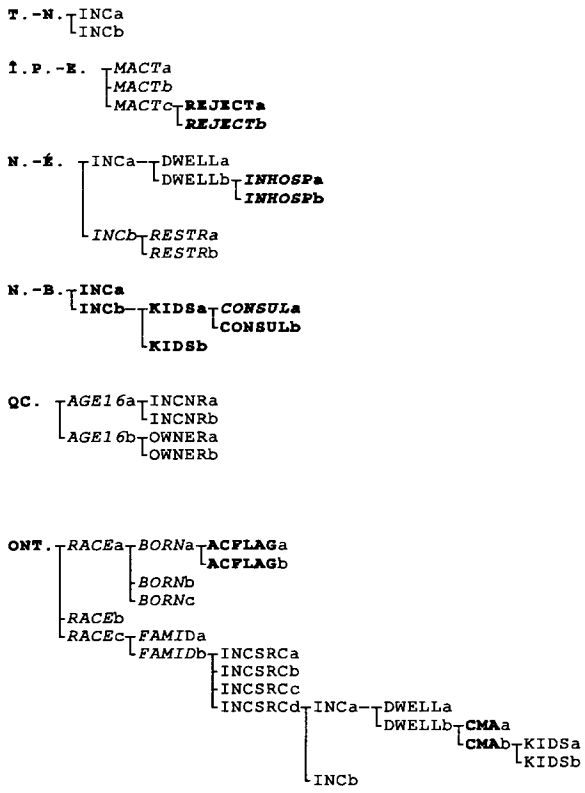
Les cellules d'ajustement produites par l'algorithme CHAID ont été révisées, mais très peu de modifications ont été apportées. À l'intérieur de chacune des cellules, le poids des unités déclarantes a été réparti au prorata, afin d'obtenir le poids total pour les unités de déclaration et de non-déclaration. L'ampleur de l'ajustement au poids des unités non déclarantes n'a jamais excédé 1,83.

## 4. CONCLUSION

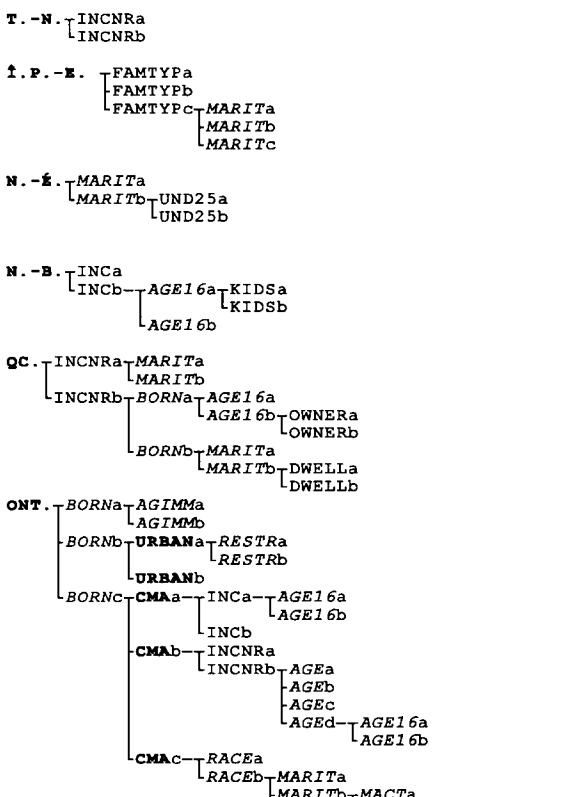
Le présent document explique la stratégie utilisée pour le traitement de la non-réponse dans les volets longitudinal et transversal du cycle 2 de l'ENSP. L'approche adoptée permettait de tenir compte de considérations pratiques telles que la nécessité de se doter d'un moyen facile, valide sur le plan statistique, pour définir les cellules d'ajustement de la pondération et de produire un ensemble de données plus utile (par exemple, nous avons ajusté séparément la réponse longitudinale complète et partielle). Il fallait de plus éviter de compliquer indûment le processus; c'est pourquoi nous avons utilisé à divers escients les mêmes cellules d'ajustement de la pondération. Nous avons choisi l'algorithme CHAID au détriment de la régression logistique parce que la première approche nous donnait plus de liberté quant au nombre et à la diversité des variables pouvant être prises en compte. Ainsi, nous avons pu tenir compte de nombreuses variables d'échantillonnage et caractéristiques, pour en retenir quelques-unes seulement. Il nous semblait ainsi que, nous pourrions tirer profit de ces caractéristiques dans le traitement de la non-réponse.

Par ailleurs, le calendrier très serré auquel nous étions soumis nous a obligés à laisser de côté certaines analyses. Il aurait été intéressant d'approfondir les possibilités offertes par l'algorithme CHAID. Par exemple, comme ce dernier permet d'utiliser une variable de réponse nominale, nous aurions pu classer les unités d'échantillonnage en trois groupes, soit les déclarants vivants, les déclarants décédés ou hors du champ de l'enquête, et les non-déclarants. Nous aurions aussi aimé comparer une analyse effectuée avec l'algorithme CHAID et une autre avec la méthode de la régression logistique. Nous aurions pu utiliser les variables du volet Santé, telles que l'indice de santé ou le statut fumeur/consommateur d'alcool, pour définir les cellules d'ajustement de la pondération, bien que leur utilité aurait été amoindrie du fait de leur absence dans certaines unités (elles sont absentes des schémas de réponse G*-GS, G*-G* et G*-**). Il faudrait reconsidérer la décision d'utiliser les mêmes cellules d'ajustement de la pondération pour différents types de non-réponse. Aurait-il mieux valu utiliser les cellules d'ajustement créées pour la réponse au niveau des ménages pour l'analyse de la non-réponse transversale dans le volet Santé? Pour comparer l'efficacité des diverses stratégies d'ajustement de la non-réponse, il faudrait évaluer leur incidence sur la variance des estimateurs. Nous pourrions de la même façon évaluer l'incidence de l'ajustement de la non-réponse dans le cycle 2 sur le biais de la non-réponse, en utilisant les données du cycle 1 qui sont disponibles pour tous les panélistes. Les estimations portant sur l'ensemble de l'échantillon pourraient être comparées aux estimations ajustées en fonction de la non-réponse, générées à partir des unités déclarantes.

Le cycle 3 présentera de nouvelles difficultés. Un «aplanissement» de l'ensemble de l'échantillon est prévu pour cette année, qui influera sur la stratégie d'estimation transversale et, partant, sur le traitement de la non-réponse. Étant donné que la non-réponse dans le volet longitudinal augmentera, il faudra tenir compte des effets secondaires de l'ajustement de la pondération, tels que la création possible de valeurs aberrantes de pondération. Il deviendra ardu d'établir des groupes de pondération pour différents types d'analyses longitudinales, parce que le nombre de schémas de réponse partielle augmentera. Combien de schémas peut-on traiter de front sans difficulté, et lesquels choisir? Il faudrait évaluer la pertinence des données supplémentaires, telles que le statut «personne déménagée», pour le traitement de la non-réponse. On devra probablement recourir à l'imputation pour la non-réponse dans le cycle 3. Comment pourra-t-on concilier le procédé d'imputation avec la méthode d'ajustement de la pondération dans le traitement de la non-réponse? Il reste beaucoup de travail autour de l'ENSP. Il faut espérer que nous aurons le temps de régler certaines questions en suspens avant que ne s'achève le traitement de la réponse pour le cycle 3.




3A. CLASSES DE RÉPONSE AU NIVEAU DES MÉNAGES




3B. CLASSES DE RÉPONSE LONGITUDINALE COMPLÈTE

(TYPES DE CARACTÉRISTIQUES: **ÉCHANTILLONNAGE/GÉOGRAPHIQUES**, LOGEMENT/MÉNAGE, *PERSONNELLES*)

**Figure 3.** Classes de réponse à l'échelle provinciale, établies pour la non-réponse au cycle 2

## REMERCIEMENTS

Les auteurs tiennent à remercier Douglas Yeo ainsi que les examinateurs pour leurs judicieux commentaires.

### BIBLIOGRAPHIE

ANGOSS SOFTWARE (1995). *Knowledge SEEKER IV for Windows – User's Guide*. ANGOSS Software International Limited.

CATLIN, G., et WILL, P. (1992). Enquête nationale sur la santé de la population: premiers faits saillants. *Rapports sur la santé*, Statistique Canada, 4, 313-319.

CZAJKA, J.L., HIRABAYASHI, S.M., LITTLE, R.J.A., et RUBIN, B.R. (1992). Projecting from advance data using propensity modelling: an application to income and tax statistics. *Journal of Business & Economic Statistics*, 10, 117-131.

DAVID, M.H., LITTLE, R., SAMUHEL, M., et TRIEST, R. (1983). Imputation models based on the propensity to respond. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 168-173.

DUFOUR, J., GAGNON, F., MORIN, Y., RENAUD, M., et SÄRNDAL, C.-E. (1998). Measuring the impact of alternative weighting schemes for longitudinal data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. À paraître.

ERNST, L. (1989). Weighting issues for household and family estimates. Dans *Panel Surveys*, (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley and Sons, 135-159.

GRONDIN, C. (1996). Pondération longitudinale – Première vague de l'EDTR. Note interne, division des méthodes d'enquête sociales, Statistique Canada.

INSTITUTE FOR SOCIAL RESEARCH (1979). A Panel Study of Income Dynamics: Procedures and Tape Codes – 1978 Interviewing Year – Wave XI – A Supplement. The University of Michigan.

KALTON, G., et KASPRZYK, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-17.

KASS, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29, 119-127.

LAVALLÉE, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 27-35.

LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de Statistique*, 54, 137-157.

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

RIZZO, L., KALTON, G., et BRICK, J.M. (1996). Comparaison de quelques méthodes de correction de la non-réponse d'un panel. *Techniques d'enquête*, 22, 43-53.

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

SKINNER, C.J., et RAO, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 433, 349-356.

STATISTIQUE CANADA (1995). Enquête nationale sur la santé de la population 1994-95 public use microdata file. No. 82F0001XCB au catalogue.

STUKEL, D.M., MOHL, C.A., et TAMBAY, J.L. (1997). Weighting for cycle two of Statistics Canada's National Population Health Survey. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 111-116.

TAMBAY, J.L., et CATLIN, G. (1995). Plan d'échantillonnage de l'Enquête nationale sur la santé de la population. *Rapports sur la santé*, Statistique Canada, 7, 31-42.

THOMSEN, I. (1973). A note on the efficiency of weighting subclass means to reduce the effects of nonresponse when analyzing survey data. *Statistik Tidskrift*, 11, 278-283.