Division de la statistique du revenu

75F0002MIF - 00004

Pondérations longitudinale et transversale de l'Enquête sur la Dynamique du travail et du revenu Année de référence 1997

Préparé par :
Isabelle Lévesque
Sarah Franklin

Août 2000

Statistique Canada Statistics Canada

Canada

Des données sous plusieurs formes

Statistique Canada diffuse les données sous formes diverses. Outre les publications, des totalisations habituelles et spéciales sont offertes. Les données sont disponibles sur Internet, disque compact, disquette, imprimé d'ordinateur, microfiche et microfilm, et bande magnétique. Des cartes et d'autres documents de référence géographiques sont disponibles pour certaines sortes de données. L'accès direct à des données agrégées est possible par le truchement de CANSIM, la base de données ordinolingue et le système d'extraction de Statistique Canada.

Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Services aux clients, Division de la statistique du revenu, Statistique Canada, Ottawa, Ontario, K1A 0T6 ((613) 951-7355; (888) 297-7355; revenu@statcan.ca) ou à l'un des centres de consultation régionaux de Statistique Canada :

Halifax	(902) 426-5331	Regina	(306) 780-5405
Montréal	(514) 283-5725	Edmonton	(403) 495-3027
Ottawa	(613) 951-8116	Calgary	(403) 292-6717
Toronto	(416) 973-6586	Vancouver	(604) 666-3691
Winnipeg	(204) 983-4020		

Vous pouvez également visiter notre site sur le Web : http://www.statcan.ca

Un service d'appel interurbain sans frais est offert **à tous les utilisateurs qui habitent à l'extérieur des zones de communication locale** des centres de consultation régionaux.

Service national de renseignements	**1 800 263-1136**
Service national d'appareils de télécommunications pour les malentendants	**1 800 363-7629**
Numéro pour commander seulement (Canada et États-Unis)	**1 800 267-6677**

Renseignements sur les commandes et les abonnements

Les prix ne comprennent pas les taxes de vente

On peut se procurer ce produit n° 75F0002MIF-00004 au catalogue sur internet. Un numéro coûte 0 $CAN. Pour obtenir un numéro de ce produit, les utilisateurs sont priés de se rendre à http://www.statcan.ca/cgi-bin/downpub/freepub_f.cgi.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois et dans la langue officielle de leur choix. À cet égard, notre organisme s'est doté de normes de service à la clientèle qui doivent être observées par les employés lorsqu'ils offrent des services à la clientèle. Pour obtenir une copie de ces normes de service, veuillez communiquer avec le centre de consultation régional de Statistique Canada le plus près de chez vous.

Statistique Canada
Division de la statistique du revenu

Pondérations longitudinale et transversale de l'Enquête sur la Dynamique du travail et du revenu
Année de référence 1997

Publication autorisée par le ministre responsable de Statistique Canada

août 2000

Nº 75F0002MPF - 00004 au catalogue
ISSN 0000-0000

Nº 75F0002MIF - 00004 au catalogue
ISSN 0000-0000

Périodicité : Irr.

Ottawa

This publication is available in English upon request.

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population, les entreprises, les administrations canadiennes et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques précises et actuelles.


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

RÉSUMÉ

L'Enquête sur la Dynamique du travail et du revenu (EDTR), introduite à l'année de référence 1993, est une enquête longitudinale par panel menée auprès des individus. Elle vise à mesurer les changements au niveau du bien-être économique des individus et les facteurs qui peuvent influencer ces changements. L'échantillon de l'EDTR est divisé en deux panels qui se chevauchent, d'une durée de six ans chacun. Les enquêtes longitudinales, comme l'EDTR, sont particulières de par la nature dynamique de la composition de leur échantillon, engendrée directement par la dynamique des familles et des ménages au fil des ans. À chaque année de référence, l'EDTR produit deux ensembles de poids : un ensemble de poids qui est représentatif de la population initiale (l'aspect longitudinal) et un autre qui est représentatif de la population actuelle (l'aspect transversal). Pour la production de poids transversaux, l'EDTR combine deux échantillons indépendants et assigne une probabilité de sélection aux individus qui se sont joints à l'échantillon après la sélection de l'échantillon initial. Les poids longitudinaux tout comme les poids transversaux sont ajustés pour la non-réponse, les valeurs influentes et la confidentialité, et un redressement de l'échantillon est effectué pour représenter la population cible. Le but de ce document est donc de décrire la méthodologie employée à l'EDTR pour pondérer les échantillons longitudinal et transversal, et de présenter les développements importants à venir. Afin de mieux illustrer la stratégie de pondération, les résultats de l'année de référence 1997 sont utilisés.

TABLE DES MATIÈRES

1. Introduction

L'Enquête sur la Dynamique du travail et du revenu (EDTR), introduite à l'année de référence 1993, est une enquête longitudinale par panel menée auprès des individus. Elle vise à mesurer les changements au niveau du bien-être économique des individus et les facteurs qui peuvent influencer ces changements, plus particulièrement les facteurs déterminants au niveau des caractéristiques démographiques, familiales et au niveau de l'activité. Le produit final de l'enquête est un ensemble de fichiers de microdonnées. Pour un aperçu général de l'enquête, se référer à Lavigne et Michaud (1998).

Originalement, l'objectif principal de l'EDTR visait à fournir des données longitudinales afin de produire des estimations et des analyses du point de vue longitudinal. L'aspect transversal de l'enquête est maintenant devenu tout aussi important que l'aspect longitudinal, entre autre en raison de l'intégration de l'Enquête sur les finances des consommateurs (EFC), une enquête transversale, à l'EDTR (Cotton et coll., 1999). La combinaison de ces deux enquêtes s'est faite à partir de l'année de référence 1998. Pour plus de renseignements sur l'EFC, se référer à Statistique Canada, 1997.

La pondération d'une enquête longitudinale peut parfois s'avérer un grand défi méthodologique, tant par ses aspects longitudinal et transversal, que par la nature dynamique de la composition du panel (engendrée directement par la dynamique des familles et des ménages au fil des ans). De nombreux défis se sont ajoutés au cours des années du point de vue de la pondération et plusieurs autres sont à venir : la combinaison de panels qui se chevauchent, l'intégration de l'EFC, le changement de logiciel utilisé pour la modélisation de la non-réponse, l'augmentation du nombre de post-strates pour les groupes d'age et sexe et l'ajout de nouvelles post-strates (nombre d'entités économiques, groupes de revenu), les changements au niveau du traitement des données, etc. Toutes ces considérations ajoutent à la complexité des pondérations longitudinale et transversale de l'EDTR.

Le but premier du présent document est de présenter les différentes étapes de la pondération, tant longitudinale que transversale, telles qu'utilisées lors de la production des poids pour l'année de référence 1997. Le deuxième objectif est de soulever certaines préoccupations et questions relatives à quelques-unes des étapes de la pondération. Et le troisième objectif consiste à informer le lecteur des développements futurs et des changements importants à venir en ce qui concerne la pondération. Le document est divisé en cinq sections. Tout d'abord, la méthodologie de l'enquête est donnée à la section 2. Ensuite, les étapes de la pondération longitudinale, de la détermination des poids initiaux à l'obtention des poids longitudinaux finaux, sont décrites à la section 3. Puis, toutes les étapes de la pondération transversale sont discutées à la section 4. De plus, les considérations futures sont présentées à la section 5. Et finalement, la conclusion est donnée à la section 6.

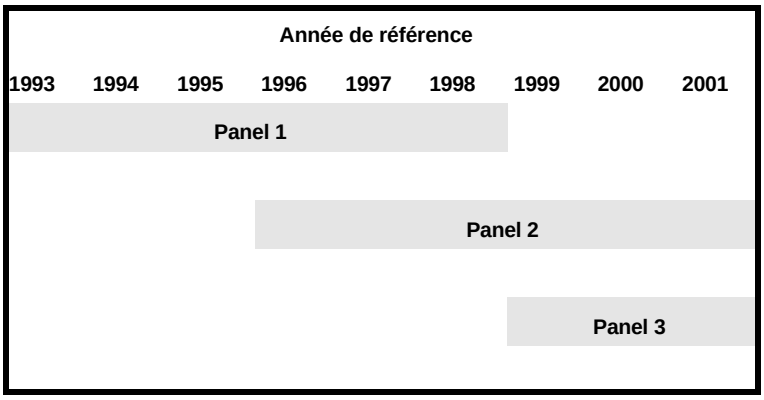
Ce document s'inspire en partie des documents qui ont été écrits jusqu'à maintenant sur la pondération de l'EDTR (Grondin, 1996; Gagnon, 1997; Renaud, 1997). Ces documents portent sur les années de référence 1993 et 1994. Comme plusieurs changements et nouveautés ont été apportés à la méthodologie des pondérations longitudinale et transversale depuis, la rédaction d'un nouveau document s'est alors avérée nécessaire.

2. Méthodologie d'enquête

L'échantillon de l'EDTR est constitué de deux panels d'une durée de six ans chacun (ou six vagues). Le premier panel, sélectionné en janvier 1993, couvre la population des personnes habitant dans les dix provinces canadiennes au 31 décembre 1992. Le deuxième panel a été sélectionné en janvier 1996 et représente la population des personnes dans les dix provinces canadiennes au 31 décembre 1995. La population couverte pour chacun des deux panels exclut les personnes résidant dans une réserve indienne, les membres à temps plein des Forces armées canadiennes habitant dans les casernements militaires ainsi que les pensionnaires dans les établissements institutionnels pour plus de six mois. Dans ce document, tout ce qui fera référence à la population des dix provinces canadiennes implique toutes ces exclusions.

Un nouveau panel est introduit à tous les trois ans pour remplacer le panel le plus âgé. Ceci étant fait dans le but d'améliorer la représentativité de l'échantillon transversal (pour tenir compte des nouveaux ménages qui se sont ajoutés à la population depuis trois ans), de diminuer l'effet de l'érosion et de réduire le fardeau de réponse. La figure 2.1 donne un aperçu de la sélection des panels et du chevauchement entre ceux-ci.

Figure 2.1 : Chevauchement des panels de l'EDTR



Chaque panel de l'EDTR consiste en un sous-échantillon d'environ 15 000 ménages (approximativement 40 000 personnes) provenant de l'Enquête sur la population active (EPA). L'échantillon de l'EPA est prélevé d'une base aréolaire selon un plan de sondage probabiliste à plusieurs degrés. L'EPA fonctionne sur une base de six panels (groupes de renouvellement) d'une durée de six mois chacun dont un est renouvelé à tous les mois. L'unité d'échantillonnage au dernier degré est le logement. Ainsi tous les individus appartenant aux ménages qui occupent les logements choisis font partie de l'échantillon de l'EPA. Pour plus de renseignements sur l'EPA, se référer aux documents décrivant la méthodologie de cette enquête (Singh et coll., 1990; et Gambino et coll., 1998).

Les ménages sélectionnés pour l'EDTR sont ceux faisant partie des groupes de renouvellement sortant des panels de l'EPA aux deux premiers mois de la première période de référence de l'EDTR (soit en janvier et février 1993 pour le premier panel, et en janvier et février 1996 pour le deuxième panel). Le fichier de l'EPA de janvier est celui dont l'EDTR se sert pour choisir son échantillon de départ. Parmi tous les ménages faisant partie des groupes de renouvellement sortant de l'EPA en janvier ou février, l'EDTR sélectionne seulement les ménages qui sont répondants à l'EPA en janvier. La dernière interview de

l'EPA sert alors de premier contact pour introduire l'EDTR. Une interview préliminaire est d'abord administrée à chaque personne sélectionnée par l'EDTR. Cette interview recueille des renseignements de base sur l'expérience de travail, les antécédents familiaux et personnels, et le niveau d'instruction.

Pour le premier panel, l'EDTR a mené l'interview préliminaire en même temps que la dernière interview de l'EPA. À cause de contraintes budgétaires, les ménages répondants à l'interview préliminaire ont dû être sous-échantillonnés; l'échantillon longitudinal pour le premier panel ne comprend que les personnes de ces ménages. Un sous-échantillon des non-répondants à cette interview a aussi été choisi pour permettre d'étudier le biais causé par la non-réponse. Même si des données sont recueillies pour ces ménages, elles ne sont pas utilisées pour la production. Cet échantillon ne sera donc pas considéré ici.

Pour le deuxième panel, l'interview préliminaire a été menée en même temps que les interviews de la première vague pour diminuer les frais associés à la collecte des données de l'EDTR. L'échantillon longitudinal pour le deuxième panel comprend alors toutes les personnes auxquelles on a administré une interview préliminaire (répondantes et non-répondantes). Les figures 2.2 et 2.3 donnent un apreçu de la sélection des ménages pour l'EDTR. Plus de détails sur la sélection des panels de l'EDTR sont donnés à la section 3.1.

Figure 2.2 : Sélection des ménages pour l'échantillon longitudinal de l'EDTR - panel 1

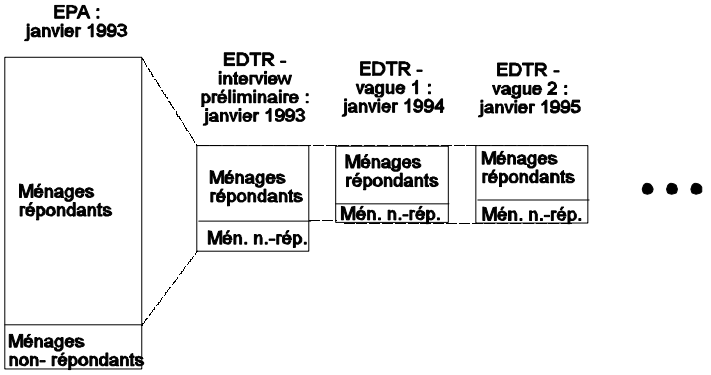
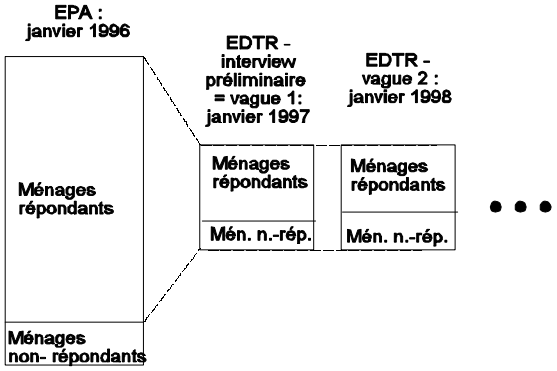


Figure 2.3 : Sélection des ménages pour l'échantillon longitudinal de l'EDTR - panel 2



Chaque personne sélectionnée pour l'EDTR est ensuite interviewée deux fois par année pendant six ans. À chaque année, des informations sont recueillies concernant la composition de la famille et du ménage, l'activité et les revenus de l'année précédente. En janvier, l'information sur l'activité est recueillie alors qu'en mai, on obtient de l'information sur le revenu. Afin de réduire le fardeau de réponse, les répondants peuvent éviter la collecte de mai en préautorisant Statistique Canada à utiliser leur dossier fiscal de Revenu Canada.

À la sélection d'un panel, toutes les personnes faisant partie des ménages choisis pour l'EDTR, indépendamment de l'âge de ces personnes, deviennent des membres de l'échantillon longitudinal du panel. Ces personnes sont considérées comme des membres de l'échantillon longitudinal pendant les six ans de participation du panel, même si elles déménagent, meurent, entrent en institution ou deviennent membres des forces armées à temps plein à l'intérieur des campements militaires. Aucune autre personne ne devient membre de l'échantillon longitudinal pour ce panel. Ainsi, pour chaque panel, l'échantillon longitudinal est constitué au moment de l'introduction du panel et reste tel quel pendant toute la durée du panel. La population ciblée par l'échantillon longitudinal demeurera la même tout au long des six années de participation du panel à l'enquête.

Du point de vue longitudinal, l'unité d'intérêt est l'individu; le ménage pouvant difficilement servir d'outil d'analyse longitudinale à cause de sa nature dynamique dans le temps. Du point de vue transversal, on s'intéresse aussi bien à l'individu qu'au ménage. Comme l'EDTR s'intéresse non seulement aux caractéristiques des personnes longitudinales mais aussi aux caractéristiques du ménage, on interviewe toutes les personnes habitant avec au moins un individu longitudinal. L'échantillon transversal pour une année donnée est alors constitué de toutes les personnes longitudinales dans le champ de l'enquête au 31 décembre de l'année de référence correspondante, et de toutes les personnes qui habitent avec elles à ce moment précis de l'année. Une personne est considérée dans le champ de l'enquête au 31 décembre d'une année de référence donnée (dans le champ de l'enquête transversalement) si elle habite dans une des dix provinces canadiennes, ne réside pas dans une réserve indienne ou un établissement institutionnel depuis plus de six mois, et n'est pas un membre à temps plein des Forces armées canadiennes habitant dans un casernement militaire, à ce moment de l'année. Toute personne longitudinale qui est dans le champ de l'enquête transversalement fera aussi parti de l'échantillon transversal. Toutes les personnes interviewées qui ne font pas partie de l'échantillon longitudinal sont appelées cohabitants. Il est à noter que l'ajout des cohabitants à l'échantillon transversal aide à améliorer la représentativité de l'échantillon transversal (Latouche, Michaud, 1997).

À chaque année, des poids longitudinaux et transversaux sont produits pour satisfaire ces besoins. La pondération longitudinale vise donc à produire des estimations qui sont représentatives de la population des dix provinces canadiennes au moment de la sélection de l'échantillon longitudinal alors que la pondération transversale permet de produire des estimations représentatives de la population des dix provinces canadiennes en date du 31 décembre d'une année de référence donnée. Par exemple, les estimations produites à partir de l'échantillon longitudinal du premier panel sont représentatives de la population des dix provinces canadiennes au 31 décembre 1992 et celles produites avec l'échantillon longitudinal du deuxième panel font référence à la population des dix provinces canadiennes au 31 décembre 1995. Les estimations produites à partir de l'échantillon transversal pour l'année de référence 1997 représentent la population des dix provinces canadiennes au 31 décembre 1997.

En janvier 1993, 39 745 personnes ont été sélectionnées pour faire partie de l'échantillon longitudinal du premier panel. En janvier 1996, 43 547 personnes ont été choisies pour faire partie de l'échantillon longitudinal du deuxième panel. L'échantillon transversal pour

l'année de référence 1997 comprend 81 090 individus, dont 70 372 personnes longitudinales.

Le tableau 2.1 donne un aperçu des changements dans la constitution des échantillons longitudinaux au cours du temps. Le tableau 2.2 présente la constitution des échantillons transversaux pour chacune des années de référence traitées jusqu'à maintenant (1993 à 1997 inclusivement).

Tableau 2.1 : Constitution des échantillons longitudinaux des deux panels (nombre de personnes)

Année de référence		1993	1994	1995	1996	1997
Panel 1	**Personnes contactées à l'interview**					
	Dans le champ de l'enquête au 31 décembre d'une année de référence donnée[1]	39 456	36 241	34 336	33 159	31 802
	Habitant en dehors des dix provinces canadiennes	28	37	45	150	280
	En institution	81	119	278	256	280
	Décédée	180	408	657	908	1 134
	Duplicata, erreur	0	0	1	1	1
	Personnes non contactées à l'interview[3]					
	Refus ferme, personnes dans les ménages non dépistés[2]	0	2940	4428	5271	6 248
	Total	39 745	39 745	39 745	39 745	39 745
Panel 2	**Personnes contactées à l'interview**					
	Dans le champ de l'enquête au 31 décembre d'une année de référence donnée[1]	-----	-----	-----	41 767	38 366
	Habitant en dehors des dix provinces canadiennes	-----	-----	-----	126	270
	En institution	-----	-----	-----	40	120
	Décédée	-----	-----	-----	234	466
	Duplicata, erreur	-----	-----	-----	0	2
	Personnes non contactées à l'interview[3]					
	Refus ferme, personnes dans les ménages non dépistés	-----	-----	-----	1 380	4 323
	Total	-----	-----	-----	43 547	43 547

1 Habitant toujours dans les dix provinces canadiennes, n'est pas décédé, ne réside pas dans un établissement institutionnel depuis plus de six mois.

2 On ne considère pas dans cette catégorie les cas de refus à l'interview préliminaire de l'EDTR et à l'interview de l'EPA.

3 Avant l'année de référence 1997, cette catégorie comprend les cas de refus ferme ou les cas non dépistés à la collecte d'une année de référence donnée. À partir de l'année de référence 1997, les cas inclus dans cette catégorie font tous référence à la collecte précédente.

Tableau 2.2 : Constitution des échantillons transversaux, par panel (nombre de personnes)

Année de référence	1993	1994	1995	1996	1997
Panel 1					
Personnes longitudinales	39 456	36 241	34 336	33 159	31 802
Cohabitants	2 062	3 640	4 620	5 768	6 655
	41 518	39 881	38 956	38 927	38 457
Panel 2					
Personnes longitudinales	-----	-----	-----	41 767	38 366
Cohabitants	-----	-----	-----	2 351	4 011
	-----	-----	-----	44 118	42 377
Total	41 518	39 881	38 956	83 045	80 834

L'EDTR a pour mandat de fournir des données longitudinales et transversales sur les caractéristiques reliées à l'activité et au revenu, et de rendre ces données disponibles aux utilisateurs aussi bien à l'intérieur qu'à l'extérieur de Statistique Canada. Pour ce faire, un ensemble de fichiers de microdonnées non traitées pour la confidentialité (fichiers internes) et un ensemble de fichiers de microdonnées traitées pour la confidentialité (fichiers externes) sont produits par l'EDTR. Les fichiers internes sont disponibles seulement pour les utilisateurs de Statistique Canada, ils comprennent alors des poids qui n'ont pas reçu de traitement pour la confidentialité (*poids internes*). Quant aux fichiers externes, produits pour les utilisateurs de l'extérieur de Statistique Canada, ils sont traités pour la confidentialité. Pour cette raison, du bruit est ajouté dans tous les poids fournis sur ces fichiers (*poids externes*).

Après le traitement des données d'une année de référence donnée, les utilisateurs de Statistique Canada peuvent avoir accès à la base de données relationnelle pour faire leurs analyses transversales et longitudinales. Toute information sur cette base de données est confidentielle étant donné qu'elle n'a pas reçu de traitement pour la confidentialité. Les poids internes et externes, de même que certaines autres données traitées pour la confidentialité, sont aussi disponibles sur cette base de données.

Des fichiers transversaux et longitudinaux internes sont aussi produits pour les utilisateurs de Statistique Canada pour faciliter l'accès aux données. Plus de travail est nécessaire pour diffuser les fichiers externes puisque ces derniers doivent être traités pour la confidentialité. L'ensemble des méthodes de contrôle de la divulgation est évaluée et améliorée pour assurer la confidentialité des données fournies. Pour l'année de référence 1997, la stratégie de diffusion initialement prévue consistait à produire un ensemble de fichiers transversaux et longitudinaux de microdonnées à grande diffusion.

Les fichiers transversaux (faisant référence à l'année de référence traitée) comprennent toutes les personnes (longitudinaux et cohabitants) de 16 ans et plus dans les ménages répondants qui sont dans le champ de l'enquête au 31 décembre de l'année de référence ciblée. Les fichiers longitudinaux (comprenant les informations de la première année de référence jusqu'à l'année de référence la plus à jour) incluent toutes les personnes longitudinales, peu importe leur statut. Le contenu des fichiers transversaux et

longitudinaux est très similaire. Une des particularités de cette stratégie de diffusion est qu'elle ne permettait pas la reconstitution des ménages au sein des fichiers de microdonnées diffusés (pour des raisons de confidentialité). Cependant à cause de contraintes de temps, la diffusion de microdonnées pour l'année de référence 1997 n'a pas encore été faite.

3. Pondération longitudinale

Cette section a pour but de décrire la méthodologie utilisée pour pondérer à chaque vague les individus longitudinaux de l'EDTR. Afin de mieux illustrer la méthodologie utilisée, des exemples tirés de la cinquième vague du premier panel et de la deuxième vague du deuxième panel sont présentés. Il est important de rappeler que les poids longitudinaux visent à représenter la population des dix provinces canadiennes au 31 décembre 1992 pour le premier panel et au 31 décembre 1995 pour le deuxième panel.

L'EDTR produit deux ensembles de poids longitudinaux : un ensemble de poids internes et un ensemble de poids externes. Les poids externes sont utilisés par les utilisateurs à l'extérieur de Statistique Canada, du bruit étant ajouté à ces poids afin de rendre plus difficile la construction des familles et des ménages. Il est important de noter que les ensembles de poids longitudinaux sont produits pour les deux panels séparément.

La pondération longitudinale s'effectue selon les six étapes présentées aux sections suivantes : la détermination des poids initiaux (section 3.1), la classification des individus longitudinaux (section 3.2), la modélisation et l'ajustement pour la non-réponse (section 3.3), l'ajustement pour les valeurs influentes (section 3.4), la stratification a posteriori (section 3.5) et l'ajout de bruit (section 3.6). Une description des poids longitudinaux finaux produits est donnée à la section 3.7.

3.1 Détermination des poids initiaux

Pour le premier échantillon de l'EDTR (premier panel), les ménages répondants des deux groupes de rotation sortant de l'EPA au mois de janvier ou février 1993 ont d'abord été choisis pour un total de 20 486 ménages. Parmi ces ménages, 17 659 ménages ont répondu à l'interview préliminaire de l'EDTR, alors que les 2 827 autres ménages ont été non-répondants. Pour des raisons budgétaires, il a été décidé de réduire la taille de l'échantillon initial de l'EDTR à environ 15 000 ménages. Pour ce faire, les ménages répondants à l'interview préliminaire ont été sous-échantillonnés avec un processus de Poisson ayant comme paramètre 0,84. De la même façon, les ménages non-répondants à l'interview préliminaire ont été sous-échantillonnés également avec un processus de Poisson mais cette fois avec un paramètre de 0,06. Ainsi, l'échantillon réduit était formé de 14 832 ménages répondants (39 745 personnes) et de 174 ménages non-répondants (410 personnes).

Pour des raisons budgétaires, seuls les ménages ayant répondu à l'interview préliminaire (14 832) ont été retenus pour faire partie de l'échantillon initial du premier panel de l'EDTR. Une étude a démontré que les impacts d'une telle décision seraient relativement faibles (Durning, 1994). Quoiqu'ils ont été exclus de l'échantillon final, les 174 ménages non-répondants à l'interview préliminaire ont été conservés pour des fins d'analyse. L'ensemble de ces ménages est appelé échantillon de qualité. Cet échantillon ne sera pas pris en considération dans ce document.

La sélection de l'échantillon du second panel de l'EDTR s'est effectuée de manière beaucoup plus simple : tous les ménages qui ont répondu à l'EPA en janvier 1996 et qui faisaient partie des groupes de renouvellement sortant des panels de l'EPA en janvier et février 1996, ont été sélectionnés (16 472 ménages pour un total de 43 547 personnes). Les interviews préliminaires du deuxième panel ont été menées en même temps que les interviews de la première vague (c'est-à-dire en janvier 1997).

Les poids initiaux de la pondération longitudinale correspondent à l'inverse de la probabilité de sélection du ménage. Ainsi tous les membres longitudinaux d'un même ménage (celui auquel ils appartenaient lors de la sélection de l'échantillon) ont le même poids initial.

L'échantillon de l'EDTR étant tiré de l'échantillon de l'EPA, le poids initial sera donc :

$$w_{initial,p_1} = w_{EPA} \times 3 \times 1{,}19$$
$$w_{initial,p_2} = w_{EPA} \times 3$$

où $w_{initial,p_1}$ = poids longitudinal initial pour le panel 1
$w_{initial,p_2}$ = poids longitudinal initial pour le panel 2
w_{EPA} = poids de l'EPA ajusté pour la non-réponse (sous-poids de l'EPA)
3 = inverse de la probabilité de sélection des groupes de rotation de l'EPA (2 groupes sur 6 ont été sélectionnés)
1.19 = inverse de la fraction de sondage du sous-échantillon de répondants à l'interview préliminaire du panel 1 (sélectionné selon un processus de Poisson).

3.2 Classification des individus longitudinaux

À chaque vague, la première étape de la pondération longitudinale est de classer les individus longitudinaux parmi les catégories suivantes : les non-répondants (dans les ménages répondants ou non-répondants), les répondants qui sont dans le champ de l'enquête transversalement et les personnes qui sont hors du champ de l'enquête transversalement. Un ménage répondant est un ménage dans lequel il y a au moins une personne qui ait répondu à au moins une des deux interviews (l'interview sur le travail ou l'interview sur le revenu). Un ménage non-répondant est un ménage dans lequel tous les membres du ménage sont non-répondants aux deux interviews. Une personne est considérée hors du champ de l'enquête transversalement (ou non-éligible transversalement) si, au 31 décembre de l'année de référence, elle était décédée, en institution pour plus de six mois ou ne résidait pas dans une des dix provinces canadiennes.

La classification des individus longitudinaux permet de déterminer les personnes qui devraient recevoir un poids longitudinal non nul. Un poids longitudinal de zéro est attribué aux individus dans les ménages non-répondants. Tous les autres individus qui étaient sélectionnés initialement pour faire partie de l'échantillon longitudinal ont un poids longitudinal non nul, même s'ils sont hors du champ de l'enquête transversalement (puisqu'ils sont considérés comme étant répondants). Il est à noter que du point de vue de la pondération, l'EDTR définit les personnes non-répondantes dans les ménages répondants comme répondantes. Cela est fait pour assurer la cohérence intra-ménage au niveau des poids (les données des personnes non-répondantes dans les ménages répondants sont alors imputées).

La classification des individus longitudinaux permet aussi de déterminer si le poids d'un individu devrait être ajusté pour la non-réponse, en plus d'identifier les individus qui seront utilisés dans le modèle d'ajustement pour la non-réponse. Ce modèle sert à trouver un ajustement adéquat pour gonfler les poids des répondants afin que ces derniers soient représentatifs de l'ensemble des répondants et des non-répondants.

Un ajustement pour la non-réponse est habituellement appliqué aux poids de toutes les unités répondantes de l'enquête. Dans le cas de l'EDTR, un facteur d'ajustement pour la non-réponse est appliqué aux poids de toutes les personnes considérées comme

répondantes, à l'exception des personnes qui sont hors champ transversalement et des enfants dans les ménages répondants. Dans le cas des personnes qui sont hors du champ de l'enquête transversalement, aucun facteur d'ajustement pour la non-réponse n'est appliqué à leur poids puisqu'on surestime déjà leur nombre (Franklin, 1999b). Dans le cas des enfants, leur poids n'est pas non plus ajusté pour compenser pour la non-réponse. La raison pour cela est que trop peu d'information est recueillie sur eux pour qu'il soit possible de modéliser adéquatement la non-réponse. Un ajustement pour compenser pour la non-réponse est fait implicitement lors de la stratification a posteriori.

Le tableau 3.1 présente les différentes catégories de réponse des individus longitudinaux pour la cinquième vague du premier panel et la deuxième vague du second panel. La colonne *Modélisation de la non-réponse* indique quelles catégories d'individus sont utilisées dans le modèle d'ajustement pour la non-réponse. Les trois dernières colonnes indiquent quel traitement sera appliqué aux poids des individus des différentes catégories.

Tableau 3.1 : Classification des individus longitudinaux pour la cinquième vague du premier panel et la deuxième vague du second panel

Classification des individus		Panel 1 vague 5	Panel 2 vague 2	Modélisation de la non-réponse	$W_{ajust} = 0$	$W_{ajust} = W_{initial}$ (non-gonflés)	$W_{ajust} = W_{initial} / R_{GHR}$ (gonflés)
ENFANTS **0-15 ans (au 31 déc. 1997)**	Dans un ménage non-répondant	1 043	1 057		**Ž**		
	Dans un ménage répondant	5 530	7 986			**Ž**	
	Dans un ménage non-éligible transversalement (en institution, décédé ou en dehors des provinces)	39	65			**Ž**	
ADULTES **16 ans et + (au 31 déc. 1997)**	Non-éligibles transversalement (en institution, décédé ou en dehors des provinces)	1 513	705			**Ž**	
	Répondants	25 819	29 002	U			**Ž**
	Non-répondants dans un ménage non-répondant	5 648	4 407	U	**Ž**		
	Duplicata, erreur	1	2		**Ž**		
	Non-répondants dans un ménage répondant (seront imputés)	151	323				**Ž**
Échantillon de qualité[1]		411	-----		**Ž**		
Total		40,155	43,547				

[1] Cette catégorie contient toutes les personnes longitudinales faisant partie de l'échantillon de qualité, de même que toute personne longitudinale qui se trouve dans un ménage contenant au moins un membre de l'échantillon de qualité (ce qui est le cas d'une personne pour l'année de référence 1997).

3.3 Modélisation et ajustement pour la non-réponse

Une fois les poids initiaux calculés, il faut maintenant calculer un facteur d'ajustement pour la non-réponse pour chaque individu considéré comme répondant, à l'exception de ceux qui sont hors champ transversalement et des enfants dans les ménages répondants. Rappelons ici que les deux panels sont traités séparément. Le facteur d'ajustement pour la non-réponse appliqué à un individu est défini comme étant l'inverse du taux de réponse du groupe homogène de réponse (GHR) dans lequel l'individu se situe. Les GHR sont formés en regroupant ensemble des individus dont les caractéristiques qui expliquent la non-réponse sont similaires. Si les GHR sont définis de telle sorte que la non-réponse soit complètement aléatoire à l'intérieur d'un groupe, alors le biais dû à la non-réponse est négligeable (Tambay et coll., 1998). Pour trouver un ajustement adéquat pour gonfler les poids des répondants afin que ces derniers soient représentatifs de l'ensemble des répondants et des non-répondants, on procède à la modélisation de la non-réponse.

La première étape de la modélisation de la non-réponse consiste à déterminer quelles sont les variables explicatives qui doivent être considérées dans le modèle afin de bien prédire la variable dichotomique *réponse* (dont les modalités sont simplement *répondant* ou *non-répondant*). La manière la plus simple de choisir ces variables est d'utiliser des informations recueillies à l'EPA, c'est-à-dire avant les premières interviews de l'EDTR. L'avantage d'utiliser ces informations est qu'elles sont disponibles pour tous les individus longitudinaux, qu'ils soient répondants ou non à la vague en traitement. Cependant, certaines des variables recueillies sont de nature dynamique (comme le revenu, le niveau d'éducation, l'état matrimonial, l'emploi, la présence d'enfants, etc.) et risquent d'être désuètes à la date de référence de la vague traitée. Mais si on décidait d'utiliser, pour chaque individu, les informations connues les plus à jour, on risquerait de détenir des informations provenant de périodes de référence différentes (à cause de la non-réponse observée à la vague en traitement). En d'autres termes, l'information la plus à jour serait utilisée pour les répondants, alors que pour les non-répondents, de l'information plus ancienne devrait être utilisée. De plus, en considérant l'information la plus à jour pour la modélisation de la non-réponse, il faudrait utiliser le poids associé à la vague précédant la vague en traitement comme poids de départ au lieu d'utiliser le poids initial de l'EDTR. Mais là vient aussi le problème des non-répondants convertis; ce sont des personnes qui sont répondantes deux vagues précédant la vague en traitement, non-répondantes à la vague précédente mais répondantes à la vague en traitement. Comme les non-répondants convertis ont un poids longitudinal nul à la vague précédant celle en traitement, il faudrait trouver une façon de leur assigner un poids de départ. Une des façons serait de se servir du poids calculé deux vagues précédant celle en traitement. Il faudrait alors s'assurer que la méthode utilisée pour assigner des poids de départ qui proviennent de périodes différentes est acceptable du point de vue méthodologique.

La possibilité d'utiliser l'information la plus à jour pour la modélisation de la non-réponse avait aussi été rejetée parce que l'on croyait que l'effet cumulatif de tous ces modèles de non-réponse donnerait des estimations qui seraient trop dépendantes des modèles de non-réponse.

L'EDTR a donc choisi de ne considérer que l'information connue recueillie avant le début de la collecte des données de l'EDTR. Pour le premier panel, cette information a été prise de l'interview préliminaire étant donnée que l'interview préliminaire avait été menée au moment de la dernière interview de l'EPA. Pour le deuxième panel, l'information considérée est celle qui a été recueillie lors de la dernière interview de l'EPA.

On effectue la première étape de la modélisation de la non-réponse (qui est de déterminer les variables explicatives) seulement lors du traitement de la première vague d'un panel

étant donné qu'on fixe dès le départ les variables à considérer dans le modèle de non-réponse. Aux vagues suivantes, les variables choisies lors de la modélisation de la non-réponse à la première vague sont utilisées. Pour déterminer les variables explicatives à considérer pour la modélisation de la non-réponse, les variables catégoriques dont le nombre de modalités (*m*) est supérieur à 2 sont recodées en *m* variables dichotomiques. À partir de toutes ces variables, on établit subjectivement une liste des variables les plus susceptibles d'influencer la réponse.

Pour chacune des variables de cette liste, des régressions logistiques sont produites en prenant une variable à la fois. D'après les résultats de ces régressions logistiques univariées, les variables qui sont significatives sont retenues. En tout, 57 et 43 variables dichotomiques ont été conservées pour le premier et le deuxième panels respectivement. La liste complète de ces variables est fournie par panel à l'annexe A. Comme mentionné précédemment, les mêmes variables sont conservées d'une vague à l'autre d'un même panel; les seuls changements habituellement effectués sont des modifications aux modalités de certaines variables (comme les variables de groupe d'âge et de revenu).

La deuxième étape est de dériver les GHR en considérant les variables retenues à l'étape précédente. Avant l'année de référence 1996, l'EDTR utilisait la régression logistique pour déterminer les GHR en utilisant le type de sélection STEPWISE de la procédure LOGISTIC de SAS. Depuis l'année de référence 1996, l'EDTR utilise la méthode de modélisation par segmentation pour déterminer les GHR, en se servant du logiciel Knowledge Seeker (version 4.2.2). Cette méthode est préférée à la régression logistique puisqu'elle est plus efficace pour réduire le biais dû à la non-réponse (Dufour et coll., 1998).

La modélisation par segmentation utilise un processus itératif pour partitionner le fichier de données et ainsi former l'arbre de décision. Le premier noeud de l'arbre de décision est tout simplement l'ensemble complet des données. La première itération définit le premier embranchement de l'arbre. Ce dernier est créé en déterminant la variable qui influence le plus significativement la variable réponse parmi les variables retenues pour la modélisation. Une mesure statistique reflétant l'importance de la variable pour discriminer la non-réponse est calculée pour chaque variable; et d'après cette mesure, la variable qui est la plus importante significativement est choisie. Le premier embranchement est ainsi formé, créant un noeud à chaque extrémité. Pour chaque noeud formé, une nouvelle itération est effectuée en répétant le même mécanisme. De nouveaux embranchements et de nouveaux noeuds sont alors créés. Si, pour un noeud donné, aucune variable n'est significative pour expliquer la réponse, alors aucun autre embranchement n'est créé. La formation de cette branche de l'arbre est alors complétée. Le processus itératif se termine lorsque toutes les branches de l'arbre sont achevées.

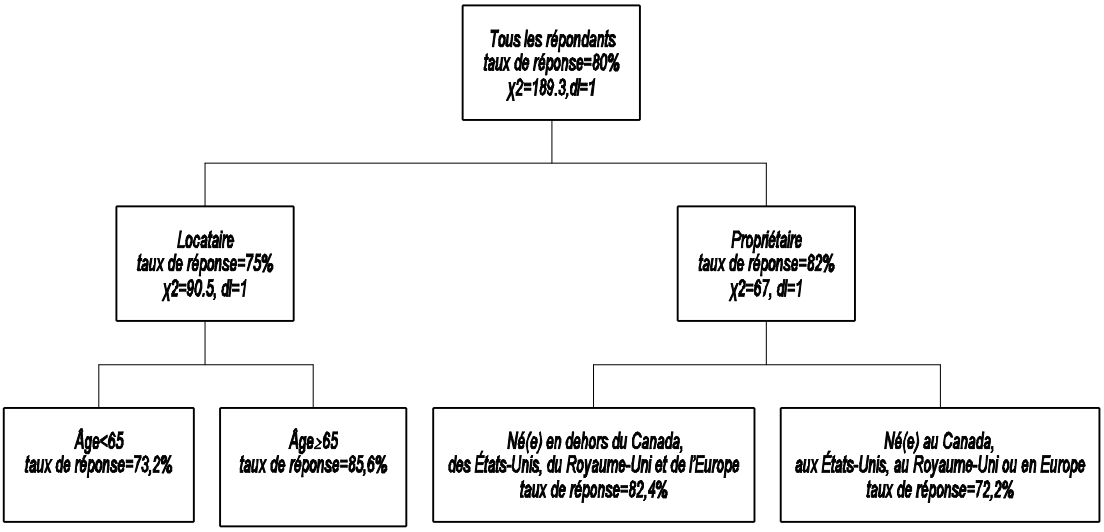
Plusieurs méthodes peuvent être employées pour déterminer la variable qui influence le plus la variable réponse. L'EDTR utilise l'algorithme CHAID ("Chi-Square Automatic Interaction Detection"), qui sélectionne la variable avec la valeur de la statistique du chi-deux de Pearson la plus élevée.

Une des raisons pour laquelle la modélisation par segmentation est préférable à la régression logistique est qu'elle est plus flexible. La régression logistique force le modèle à être symétrique : une fois que les variables significatives sont déterminées, les GHR sont formés en prenant toutes les combinaisons possibles de l'intersection entre ces variables. En forçant le modèle à être symétrique, de petites cellules ou des cellules qui ne sont pas significatives peuvent être formées, ce qui peut affecter la fiabilité des estimations. La modélisation par segmentation permet au modèle d'être asymétrique : à chaque embranchement de l'arbre de décision, on détermine quelle variable est la plus significative;

ce qui garantie que chaque RHG formé est significatif et résulte habituellement en un modèle asymétrique.

Par exemple, pour le premier panel de l'EDTR, la variable la plus importante pour discriminer la réponse et la non-réponse est une variable indiquant si oui ou non une personne est propriétaire d'une maison. Pour les personnes qui sont propriétaires d'une maison, la variable la plus importante significativement pour expliquer la non-réponse est une variable indiquant si oui on non une personne est née soit aux États-Unis, au Royaume-Uni ou au Canada. Pour les personnes qui ne sont pas propriétaires d'une maison, la prochaine variable à être la plus importante significativement pour expliquer la non-réponse est une variable indiquant si oui ou non une personne est âgée de plus de 65 ans (voir figure 3.1 ci-dessous).

Figure 3.1 : Premières branches de l'arbre de décision pour le panel 1, année de référence 1997



Un autre avantage de la modélisation par segmentation est qu'elle peut être accomplie à l'aide du logiciel Knowledge Seeker; logiciel qui est très convivial, rendant ainsi la tâche de créer un arbre de décision plus facile. Cependant, modéliser par segmentation comporte certains inconvénients. Malgré qu'il est facile de créer un arbre de décision, la prochaine étape dans le processus est lente et fastidieuse puisqu'il faut coder les définitions de tous les RHG dans le programme de pondération longitudinale de l'EDTR pour pouvoir calculer le facteur d'ajustement pour la non-réponse de chaque individu. Le logiciel Knowledge Seeker ne founissant pas de fichier définissant chacun des RHG avec leur taux de réponse associé.

Un autre inconvénient de la modélisation par segmentation est directement relié à l'arbre de désicion lui-même. Lorsqu'un arbre de désicion est formé, il peut facilement devenir assez lourd : dès qu'il comporte plusieurs embranchements, il devient difficile de

l'interpréter. Les arbres de décision sont aussi assez instables; l'ajout, le retrait ou le changement d'une variable ou d'un individu peut avoir un effet important sur l'arbre lui-même. Il est alors très difficile de comparer plusieurs arbres de décision entre eux.

À chaque vague, l'EDTR crée un modèle d'ajustement pour la non-réponse pour chaque panel séparément. Chaque modèle est au départ défini au niveau national. Des embranchements au niveau provincial sont retenus dans l'arbre de décision si ces derniers sont significatifs. L'EDTR utilise un modèle au niveau national puisqu'il est plus pratique et plus facile que de créer un modèle pour chaque province. Une étude faite en 1999 a déterminé que la différence entre l'utilisation d'un modèle national ou de modèles provinciaux est négligeable. Pour plus de détails sur cette étude, se référer à Franklin (1999a).

Pour tenir compte du plan d'échantillonnage, les poids initiaux normés sont utilisés pour calculer une statistique du chi-deux pondérée. Il est à noter que l'EDTR ne fait pas de distinction entre les différents types de non-réponse (non-réponse causée par l'érosion, non-dépisté, refus, etc.). Il est également important de mentionner que certaines contraintes sont appliquées au modèle lors de la modélisation par segmentation chaque GHR doit contenir au moins trente personnes (pondéré et non pondéré) et doit avoir un taux de réponse pondéré d'au moins de 50%. Kalton et Kasprzyk (1986) suggèrent d'avoir de telles contraintes puiqu'elles permettent d'éviter une trop grande variabilité dans les poids et diminuent la perte de précision dans les estimations.

Une fois l'arbre de décision obtenu, les GHR sont formés en considérant chaque extrémité des branches de l'arbre. Les individus qui ont été utilisés pour modéliser la non-réponse (i.e. les adultes répondants ainsi que le adultes non-répondants dans les ménages non-répondants) et les adultes non-répondants qui sont dans un ménage répondant sont alors classés dans les différents GHR. Il est à noter que les adultes non-répondants dans les ménages répondants ne sont pas utilisés lors de la modélisation de la non-réponse même s'ils sont considérés comme répondants puisque dans les faits ils n'ont pas répondu à l'enquête.

L'ajustement pour la non-réponse est fait à l'intérieur de chaque GHR : on divise le poids des répondants d'un GHR par le taux de réponse pondéré observé dans le GHR (ce qui revient à multiplier le poids des répondants par un facteur d'ajustement correspondant à l'inverse du taux de réponse pondéré observé). Pour la cinquième vague du premier panel, 282 GHR ont été formés et avaient un taux moyen de réponse pondéré de 84 %; alors que 134 GHR ont été formés pour la deuxième vague du second panel, avec un taux moyen de réponse pondéré de 88 %.

On obtient les poids corrigés pour la non-réponse en multipliant le poids initial de chaque individu répondant par le facteur d'ajustement pour la non-réponse (ou en divisant par le taux de réponse pondéré observé dans le GHR final). On met le poids des non-répondants à 0 et on garde le poids initial pour les enfants et les individus non-éligibles transversalement.

À chaque vague, les poids longitudinaux ajustés pour la non-réponse sont dérivés de la façon suivante :

$$w_{ajust} = \begin{cases} 0, & \textit{pour les personnes dans les ménages non\&répondants} \\ w_{initial}, & \textit{pour les personnes non\&éligibles transversalement ou pour les enfants} \\ \dfrac{w_{initial}}{R_{GHR}}, & \textit{pour les personnes dans les ménages répondants} \end{cases}$$

où w_{ajust} = poids longitudinal ajusté pour la non-réponse
$w_{initial}$ = poids longitudinal initial (= $w_{initial,p_1}$ pour le panel 1 ou $w_{initial,p_2}$ pour le panel 2)
R_{GHR} = taux de réponse pondéré observé dans le GHR.

3.4 Ajustement pour les valeurs influentes

Une fois que les poids ont été ajustés pour compenser pour la non-réponse, on procède à l'identification des observations influentes qui ont un impact important sur les estimations ponctuelles du revenu ainsi que sur les estimations de variance. Une valeur est dite influente si elle contribue trop fortement à l'estimation transversale du total des revenus provinciaux. Un facteur d'ajustement entre 0 et 1 est calculé et multiplié aux poids des individus identifiés influents (ainsi que le poids des autres membres de leur ménage) afin de diminuer la contribution des poids de ces individus sur les estimations du revenu. Pour la cinquième vague du premier panel, aucun individu n'a été identifié comme étant influent; et pour la deuxième vague du deuxième panel, seulement deux individus du deuxième panel ont été identifiés comme étant influents.

L'identification des valeurs influentes est seulement effectuée par rapport aux estimations transversales. La même procédure de calcul d'ajustement pour les valeurs influentes pourrait s'appliquer également avec les poids longitudinaux. On obtiendrait ainsi des ajustements pour les valeurs influentes potentiellement différents pour les pondérations transversale et longitudinale. Cependant, des tests faits pour l'année de référence 1995 ont montré qu'à ce moment-là, il n'y aurait pas eu de différence quant aux individus identifiés comme influents, et très peu de différence dans les facteurs d'ajustement comme tels. Il a donc été décidé d'utiliser, pour la pondération longitudinale, les résultats obtenus lors de l'ajustement pour les valeurs influentes du côté transversal. Pour plus de détails sur l'identification des valeurs influentes et sur le calcul du facteur d'ajustement à apporter aux poids, se référer à la section 4.6.

Le poids longitudinal ajusté pour les valeurs influentes est donc défini comme étant :

$$w_{infl} = w_{ajust} \times \beta_{infl}$$

où w_{infl} = poids longitudinal après l'ajustement pour les valeurs influentes
w_{ajust} = poids longitudinal ajusté pour la non-réponse
β_{infl} = facteur d'ajustement pour les valeurs influentes ($0 < \beta_{infl} \leq 1$).

3.5 Stratification a posteriori

Le but de la stratification a posteriori (post-stratification) est de faire en sorte que la somme des poids à l'intérieur de certains sous-groupes de l'échantillon (post-strates) corresponde aux totaux de contrôle de la population connus pour ces post-strates, pour une année donnée. La stratification a posteriori est effectuée indépendamment des deux panels. Pour chaque post-strate, la somme des poids avant la stratification a posteriori est calculée. Ensuite, pour chaque post-strate, le total de contrôle de la population est divisé par cette somme de poids. Le poids longitudinal après la stratification a posteriori est alors calculé selon la formule suivante :

$$w_{ps} \text{ ' } w_{infl} \left(\frac{T_L}{\sum_{L} w_{infl}} \right.$$

où w_{ps} = poids longitudinal après la stratification a posteriori
w_{infl} = poids longitudinal ajusté pour la non-réponse et les valeurs influentes
T_L = total de contrôle pour la post-strate L.

Pour la pondération longitudinale de l'EDTR, les post-strates sont un croisement des trois variables suivantes : province, sexe et groupe d'âge. Les groupes d'âge sont présentés dans le tableau 3.3.

Tableau 3.2 : Groupes d'âge utilisés pour la stratification a posteriori

Groupes d'âge	
Panel 1 (8 groupes d'âge)	**Panel 2 (11 groupes d'âge)**
0-15	0-6
	7-15
16-19	16-18
20-24	19-24
25-34	25-34
35-44	35-44
45-54	45-54
55-64	55-59
	60-64
65+	65-69
	70+

La variable groupe d'âge a subi d'importants changements pour le traitement des données de l'année de référence 1996. Ces modifications ont été apportées aux groupes d'âge du deuxième panel pour faciliter l'intégration de l'EFC à l'EDTR (de façon transversale) et pour

suivre certains standards adoptés par un ensemble d'enquêtes importantes à Statistique Canada. Ainsi à partir de l'année de référence 1996, les nouveaux groupes d'âge devraient être utilisés. Le deuxième panel utilise donc ces groupes d'âge depuis le tout début de son traitement puisqu'il a été sélectionné en 1996. Pour le premier panel, on conserve toujours les anciens groupes d'âge pour effectuer la stratification a posteriori pour assurer une certaine stabilité et ainsi permettre la comparaison entre les différents ensembles de poids produits aux cours des vagues précédant 1996.

Il est important de noter que la province utilisée pour la post-stratification longitudinale est la province habitée par l'individu au moment de la sélection de l'échantillon d'un panel donné, et que le groupe d'âge auquel un individu appartient est déterminé en fonction de son âge toujours au moment de la sélection de l'échantillon du panel concerné. De plus, comme les deux panels ont été sélectionnés à des années différentes, les poids longitudinaux après la stratification a posteriori du premier panel doivent correspondre aux totaux de contrôle pour la population canadienne au 31 décembre 1992 alors que les poids longitudinaux après la stratification a posteriori du deuxième panel doivent être identiques aux totaux de contrôle de la population canadienne du 31 décembre 1995. Ces totaux de contrôle sont obtenus à partir des projections démographiques produites par la Division de la Démographie. Ces données sont seulement disponibles au milieu de chaque mois (et sont donc représentatives de la population des dix provinces canadiennes, avec certaines exclusions, au milieu du mois). L'EDTR utilise les projections démographiques du mois de janvier.

L'ajustement moyen pour la stratification a posteriori apporté aux poids longitudinaux ajusté pour la non-réponse et les valeurs influentes est de 1,11 pour la cinquième vague du premier panel et de 1,10 pour la deuxième vague du second panel.

3.6 Ajout de bruit

Au moment de la production des poids longitudinaux (et transversaux) pour l'année référence 1997, deux fichiers de microdonnées à grande diffusion, un fichier longitudinal et un fichier transversal, devaient être produits annuellement. (Cela n'est plus le cas à partir de l'année de référence 1998.) Par mesure de protection de la confidentialité des données, les fichiers produits ne devaient pas permettre la reconstitution des ménages. Malgré les ajustements décrits dans les sections précédentes, il arrive que des personnes provenant d'un même ménage aient un poids identique, ce qui peut entraîner la reconstitution, même partielle, d'un ménage. Afin d'éviter ceci, il s'est avéré nécessaire, comme dernière étape de la pondération, d'ajouter du bruit aux poids qui apparaîtront sur les fichiers de microdonnées à large diffusion. L'ordre de grandeur du bruit ajouté aux poids peut être établi en fonction de la distribution des différences entre les poids consécutifs, une fois ceux-ci triés en ordre croissant.

L'ajout de la perturbation dans les poids post-stratifiés est effectué selon les étapes suivantes. D'abord, tous les gens en provenance d'un même ménage et avec un poids identique, qu'ils habitent encore ensemble ou non, ont été identifiés. Ensuite, ces gens ont été groupés deux par deux. Dans le cas où un nombre impair de personnes dans un ménage affichaient un poids identique, le poids d'une personne choisie au hasard n'a pas été modifié. Pour chaque paire, une valeur aléatoire e obtenue d'une distribution uniforme U(0,1) a été générée. À cette valeur, on ajoute une valeur a, déterminée de la façon suivante :

$$a \cdot \frac{\max(w_{ps}) - \min(w_{ps})}{n}$$

où n est la taille d'échantillon et $\max(w_{ps})$ et $\min(w_{ps})$ sont calculés au niveau national. La valeur $e + a$, comprise entre a et $a+1$, a ensuite été ajoutée au poids de la première personne de la paire et soustraite du poids de la deuxième. Ainsi, du bruit est ajouté aux poids et les valeurs identiques qui provenaient d'un même ménage sont éliminées. Les totaux de contrôle pour chaque post-strate sont toujours respectés puisque les individus d'un même ménage qui ont des poids identiques à l'étape de l'ajout de bruit proviennent nécessairement de la même strate, et ont eu les mêmes facteurs d'ajustement pour la non-réponse, pour les valeurs influentes et pour la postratification a posteriori.

Le poids après l'ajout de bruit, que l'on appelle poids externe, est donc obtenu de la façon suivante :

$$w_{bruit} = \begin{cases} w_{ps} \pm (e \% a), & \textit{pour les paires d'individus dont le poids est identique à l'intérieur d'un ménage} \\ w_{ps}, & \textit{dans les autres cas} \end{cases}$$

où w_{bruit} = poids longitudinal après l'ajout de bruit
w_{ps} = poids longitudinal après la stratification a posteriori
e = perturbation aléatoire
a = perturbation longitudinale.

Il est à noter que la valeur 1,25 utilisée pour les années références 1995 et 1996 a été réutilisée comme valeur de a pour l'année référence 1997, et ce pour les deux panels. Pour plus de détails, voir Franklin et Lévesque (1999).

3.7 Description des poids longitudinaux finaux produits

Deux ensembles de poids longitudinaux sont produits, appelés poids internes et poids externes. Les poids internes, produits à la fin de la section 3.5, sont utilisés pour des analyses faites à l'intérieur de Statistique Canada et n'ont pas été ajoutés de bruit. Du bruit dans les poids est seulement ajouté aux poids externes puisque ces derniers sont utilisés par les utilisateurs à l'extérieur de Statistique Canada. Ces poids sont produits à la fin de la section 3.6.

Le tableau 3.4 donne les médianes des poids longitudinaux initiaux et finaux par province pour l'année de référence 1997. Seuls les individus avec des poids finaux non nuls sont utilisés dans le calcul de ces médianes. Les poids considérés ici sont les poids internes (i.e. sans ajout de bruit aux poids). La province donnée dans le tableau est la province d'origine (i.e. la province habitée par l'individu au moment de la sélection de l'échantillon du panel dans lequel il appartient).

Tableau 3.3 : Médianes des poids longitudinaux internes (initiaux et finaux), pour la vague 5 du premier panel et la vague 2 du deuxième panel

Province	Panel 1		Panel 2	
	Poids initiaux	Poids finaux	Poids initiaux	Poids finaux
Terre Neuve	207	250	242	285
Île-du-Prince-Édouard	138	164	95	107
Nouvelle-Écosse	237	305	281	322
Nouveau-Brunswick	279	321	254	296
Québec	565	699	566	676
Ontario	563	769	558	718
Manitoba	221	309	327	409
Saskatchewan	305	370	279	347
Alberta	783	870	693	967
Colombie-Britannique	661	848	811	1117

4. Pondération transversale

La présente section donne une description détaillée des étapes nécessaires pour créer un ensemble de poids représentatif de la population des dix provinces canadiennes au 31 décembre d'une année de référence donnée. Tous les individus longitudinaux qui sont dans le champ de l'enquête transversalement et leurs cohabitants sont considérés lors de la production de cet ensemble de poids. La méthodologie présentée dans cette section est celle utilisée pour l'année de référence 1997.

Plusieurs ensembles de poids transversaux sont produits par l'EDTR. Ces ensembles sont composés de poids de type individu ou intégré. Le poids individu peut être différent pour chaque personne à l'intérieur d'un ménage donné tandis que le poids intégré est identique pour chaque individu à l'intérieur d'un même ménage. Le poids intégré est adéquat pour faire des analyses au niveau ménage. Il peut aussi être utilisé pour des analyses au niveau des individus. Quant au poids individu, il permet seulement de faire des analyses au niveau des individus. Certains des ensembles de poids sont produits exclusivement pour fins d'utilisation à l'intérieur de Statistique Canada (poids internes) alors que d'autres sont produits pour les utilisateurs à l'extérieur de Statistique Canada (poids externes). Pour des raisons de confidentialité, du bruit est ajouté à tous les poids externes. Lors de la production des poids transversaux pour l'année de référence 1997, les fichiers de microdonnées à grande diffusion qui devaient être produits ne devaient pas permettre la reconstitution des ménages (par mesure de protection de la confidentialité). En raison de cela, le poids intégré est seulement disponible à l'intérieur de Statistique Canada. Pour les analyses qui doivent être effectuées à l'extérieur de Statistique Canada, le poids individu est utilisé.

En tout, quatre ensembles de poids transversaux sont produits. Le premier ensemble de poids produit est le *poids intégré*, qui rappelons-le est un poids interne. Le deuxième ensemble de poids produit est le *poids individu usuel externe*. Le troisième ensemble de poids produit est le *poids travail interne*. Ce poids est un poids individu similaire au poids individu usuel mais pour lequel la notion de répondant considérée pour produire le poids a été quelque peu modifiée. Pour identifier un répondant, le poids travail ne considère que la réponse à l'interview sur le travail, au lieu d'utiliser la réponse aux deux interviews (interview sur le travail et interview sur le revenu). Par exemple, un ménage non-répondant à l'interview sur le travail (i.e. ménage pour lequel tous ses membres sont non-répondants à cette interview) mais répondant à l'interview sur le revenu (à cause d'une permission donnée préalablement pour utiliser leurs données fiscales) sera pondéré à la pondération individuelle usuelle mais ne le sera pas à la pondération travail. L'utilisation de ce poids, au lieu d'un poids intégré ou d'un poids individu usuel, permet alors d'avoir une meilleure estimation des caractéristiques sur le travail. Le quatrième ensemble de poids produit est le *poids travail externe*.

La pondération transversale s'effectue selon les huit étapes suivantes : la détermination des individus éligibles à la pondération transversale (section 4.1), l'ajustement pour la non-réponse (section 4.2), l'application des facteurs d'allocation des panels (section 4.3), le partage de poids (section 4.4), l'ajustement pour la migration inter-provinciale (section 4.5), l'ajustement pour les valeurs influentes (section 4.6), la stratification a posteriori (section 4.7) et l'ajout de bruit dans les poids (section 4.8). Une description des poids transversaux finaux produits est donnée à la section 4.9.

4.1 Détermination des individus éligibles à la pondération transversale

L'échantillon transversal pour une année donnée est représentatif de la population des dix provinces canadiennes au 31 décembre de l'année de référence (à l'exception des ménages constitués exclusivement d'immigrants). Cet échantillon est composé des individus longitudinaux faisant toujours partie de la population des dix provinces canadiennes au 31 décembre de l'année de référence et des cohabitants de ces personnes. On dit qu'une personne est *éligible transversalement* si elle fait partie de cet échantillon.

Pour déterminer quels individus auront un poids transversal non-nul, il faut d'abord identifier ceux qui sont éligibles transversalement. Tous les individus longitudinaux qui ne font plus partie de la population cible au 31 décembre de l'année de référence (comme les personnes ayant déménagé en dehors des dix provinces canadiennes, les personnes décédées ou en institution) reçoivent un poids transversal nul, même s'ils sont répondants (ou considérés comme répondants). Pour les individus longitudinaux qui font partie de la population cible au 31 décembre de l'année de référence, seuls ceux qui sont dans un ménage répondant ont un poids transversal non nul. Pour la production des poids individus usuels et des poids intégrés, la notion de ménage répondant est la suivante. Un ménage répondant est un ménage dans lequel il y a au moins une personne qui a répondu à au moins une des deux interviews (interview sur le travail, interview sur le revenu). Les personnes non-répondantes dans un ménage répondant auront de l'information imputée. Pour le poids travail, un ménage répondant est un ménage dans lequel il y a au moins une personne qui a répondu à l'interview sur le travail.

4.2 Ajustement pour la non-réponse

Comme mentionné précédemment, l'échantillon transversal pour une année donné est représentatif de la population des dix provinces canadiennes au 31 décembre de l'année de référence. Cependant, comme dans toute enquête, l'EDTR est sujette à la non-réponse (on parle ici de non-réponse totale) et la représentativité de l'échantillon transversal en est alors affectée. Pour réduire l'effet du biais causé par la non-réponse, on ajuste les poids des répondants pour compenser pour les non-répondants.

Du point de vue transversal, on détient trop peu d'information sur les cohabitants non-répondants pour permettre un ajustement pour la non-réponse à ce niveau. En effet, pour plusieurs cohabitants, les seules variables disponibles sont la province, l'âge et le sexe, variables qui seront de toute façon utilisées à l'étape de la stratification a posteriori. Considérant cela, deux méthodes d'ajustement pour la non-réponse au niveau transversal ont été proposées dans le passé (Latouche, Michaud, 1997). La première solution consistait à tout d'abord faire un ajustement pour la non-réponse pour les individus longitudinaux seulement, et à ensuite utiliser la méthode du partage de poids pour attribuer un poids aux cohabitants. De cette manière, les poids des cohabitants sont indirectement ajustés pour la non-réponse. La deuxième alternative consistait à faire un ajustement pour la non-réponse au niveau des ménages en se servant des strates et des groupes de renouvellement des logements initiaux selon le plan de sondage de l'EPA. L'ajustement au niveau des ménages tend toutefois à sous-estimer légèrement les taux de réponse.

La première méthode décrite ci-haut est celle qui a été adoptée par l'enquête. Transversalement, l'étape d'ajustement pour la non-réponse ne relève donc pas de la pondération transversale mais plutôt de la pondération longitudinale. La modélisation et l'ajustement pour la non-réponse longitudinale ont été décrits à la section 3.3. Rappelons tout de même que, pour la pondération longitudinale, chaque panel est traité séparément.

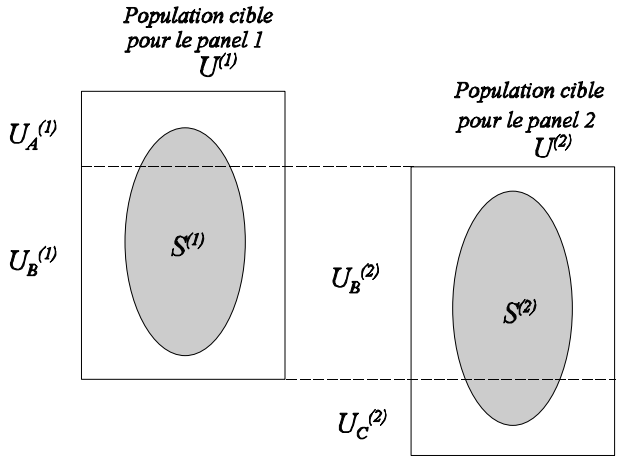
Le poids transversal ajusté pour la non-réponse longitudinale est donc égal au poids w_{ajust} dérivé à la section 3.3. Dénotons par $w_{1997,ajust}$ le poids transversal ajusté pour la non-réponse longitudinale. Ce poids est utilisé comme poids initial pour la production de tous les poids transversaux dérivés (poids individu usuel, poids intégré et poids travail).

4.3 Application des facteurs d'allocation des panels

L'année de référence 1996 était la première année pour laquelle l'EDTR avait deux panels qui se chevauchaient. En théorie les deux panels peuvent être utilisés indépendamment l'un de l'autre pour produire des estimations transversales, mais en pratique il est plus avantageux de combiner les échantillons transversaux des deux panels afin de produire les estimations transversales. La taille d'échantillon est alors doublée, ce qui réduit la variabilité des estimations et offre la possibilité d'utiliser plus de totaux de contrôle lors de l'étape de la stratification a posteriori.

Comme les deux échantillons longitudinaux sont représentatifs de la population des dix provinces canadiennes à des périodes de référence différentes (31 décembre 1992 pour le premier panel et 31 décembre 1995 pour le deuxième panel), il faut trouver un moyen permettant de combiner les deux échantillons pour que ces derniers représentent ensemble la population cible d'une année référence donnée. Le problème qui se pose dans la combinaison des échantillons du premier panel et du deuxième panel est le temps écoulé entre la sélection de ces échantillons. La figure 4.1 présente la dynamique de la population des dix provinces canadiennes entre la sélection des deux panels de l'EDTR.

Figure 4.1 : Dynamique de la population des dix provinces canadiennes entre la sélection des deux panels



En janvier 1993, on a tiré un échantillon $S^{(1)}$ de la population des dix provinces canadiennes $U^{(1)}$. (Cet échantillon étant l'échantillon longitudinal associé au premier panel). En janvier 1996, un échantillon $S^{(2)}$ a été sélectionné de la population des dix provinces canadiennes $U^{(2)}$ pour former l'échantillon longitudinal du second panel. La majorité des individus dans la population en janvier 1993 sont présents dans la population en janvier 1996 ($U_B^{(1)}=U_B^{(2)}$). Les différences entre les populations $U^{(1)}$ et $U^{(2)}$ sont causées par les personnes qui sont sorties de la population cible entre janvier 1993 et janvier 1996 (personnes décédées, institutionnalisées, personnes ayant déménagé à l'extérieur des provinces, etc.) et par les

nouveaux arrivés dans la population après janvier 1993 (immigrants, naissances). Une partie de l'échantillon du premier panel comprend des personnes de la population sortante entre janvier 1993 et janvier 1996 ($U_A^{(1)}$); les individus de cette population ne font alors plus partie de la population canadienne au moment de la sélection de l'échantillon du deuxième panel et ont ainsi aucune chance d'être choisis. De même, une partie de l'échantillon du second panel est composée de personnes qui sont dans la population des dix provinces canadiennes au moment de la sélection de cet échantillon mais qui n'étaient pas présentes dans la population lors de la sélection du premier panel ($U_C^{(2)}$); ces personnes n'ayant eu aucune chance d'être choisies à ce moment.

Afin de développer une méthode qui permet de combiner l'information des échantillons transversaux des deux panels, il suffit de considérer les deux panels au moment de l'introduction du second panel. La population que l'on doit considérer est $U^{(2)}$, celle effective lors de l'introduction du deuxième panel. Les échantillons longitudinaux à considérer sont les échantillons originaux choisis, l'échantillon du premier panel ayant été réduit à cause des individus qui ont quitté la population et des non-répondants. La méthode développée par Merkouris (1999) est celle que l'EDTR utilise pour jumeler ses échantillons qui se chevauchent (dans le cas présent, les échantillons des panels 1 et 2 sont visés). Soit Y, un paramètre de la population. Pour estimer Y, il s'agit d'utiliser l'estimateur combiné suivant (qui donne une importance relative à chacun des panels selon un certain facteur p) :

$$\hat{Y} = p\hat{Y}_1 + (1-p)\hat{Y}_2$$

où $\hat{Y}$ est l'estimation transversale pour une variable d'intérêt, $\hat{Y}_i$ est l'estimation produite pour cette même variable d'intérêt pour le panel i, et p est le facteur d'allocation des panels. Cet estimateur peut être vu comme une somme pondérée des estimateurs des deux panels où les poids sont les facteurs d'allocation des panels. La valeur optimale du facteur d'allocation des panels a été déterminée de façon à ce que la variance de $\hat{Y}$ soit minimale. Le facteur d'allocation des panels est défini comme suit :

$$p = \left(\frac{n_1}{d_1}\right)\left(\frac{n_1}{d_1} + \frac{n_2}{d_2}\right)^{-1} = \frac{n_1}{n_1 + n_2\left(\frac{d_1}{d_2}\right)} \qquad (4.1)$$

où d_i désigne l'effet de plan d'échantillonnage pour le panel i et n_i représente le nombre de personnes longitudinales de 16 ans et plus du panel i répondantes au temps t (t étant l'année de référence traitée, ici on parle de l'année de référence 1997). L'effet de plan d'échantillonnage est défini comme étant le rapport entre la variance de l'estimation sous-pondérée (i.e. sans étalonnage pour respecter les totaux de contrôle de population) et la variance qui aurait résulté d'un échantillon aléatoire simple de même taille. Pour l'EDTR, le nombre de personnes longitudinales de 16 ans et plus répondantes au temps t correspond au nombre de personnes longitudinales de 16 ans et plus éligibles transversalement qui sont dans un ménage répondant (puisque toutes les personnes dans les ménages répondants sont considérées comme répondantes, voir section 3.2). Même si les cohabitants sont utilisés pour les estimations transversales, seuls les répondants longitudinaux au temps t sont considérés dans le calcul des facteurs d'allocation des panels. Merkouris (1999) et Lavallée (1994) suggèrent de combiner les deux panels avant d'intégrer les cohabitants par la méthode du partage de poids. Omettre les cohabitants dans le calcul

des facteurs *p* contribue à réduire l'importance du panel le plus âgé, et d'ainsi minimiser les impacts potentiels dus au biais d'attrition.

La détermination d'un facteur *p* optimal nécessite donc l'estimation d'effets de plan d'échantillonnage et de tailles d'échantillons longitudinaux. Cependant aucun effet de plan n'était disponible pour le deuxième panel au moment du traitement des données. L'EDTR étant un sous-échantillon de l'EPA, il a été décidé d'utiliser les données de l'EPA pour obtenir les effets de plan d'échantillonnage nécessaires pour le calcul des facteurs *p*. De plus, comme c'est le rapport entre deux effets de plan qui est d'intérêt, l'EPA est tout aussi adéquat pour obtenir les effets de plan désirés. Pour déterminer les effets de plan nécessaires, les six groupes de renouvellement de l'EPA sont utilisés. Ces effets de plan sont calculés pour la caractéristique *nombre de personnes de 15 ans et plus* à l'échelle provinciale (ce qui impliquera le calcul de dix facteurs d'allocation des panels, un par province). Pour la combinaison des premier et deuxième panels, les fichiers de l'EPA de janvier 1993 et de janvier 1996 ont été utilisés pour calculer les effets de plan. Quant aux tailles d'échantillons voulues, elles sont calculées à partir des échantillons longitudinaux des deux panels pour l'année de référence traitée. Comme mentionné précédemment, les tailles d'échantillons désirées sont le nombre de personnes longitudinales de 16 ans et plus répondantes au temps t. On les obtient en considérant toutes les personnes dans les ménages répondants qui ont 16 ans et plus et qui sont éligibles transversalement.

Connaissant les effets de plan et les tailles d'échantillons, il est facile de calculer les facteurs d'allocation des panels d'après l'équation 4.1. Ces facteurs doivent être calculés pour chaque année traitée puisque la composition des échantillons longitudinaux change à chaque année. Pour des raisons pratiques, il a été décidé de calculer les effets de plan à chaque introduction d'un nouveau panel, et d'utiliser ces effets de plan dans le calcul des facteurs d'allocation des panels pour les trois années de chevauchement (Latouche et coll., 2000). Le tableau 4.1 qui suit donne les facteurs d'allocation des panels calculés pour l'année de référence 1997.

Tableau 4.1 : Facteurs d'allocation des panels pour l'année de référence 1997

Province	d_1/d_2	n_1	n_2	p (Panel 1)	1 & p (Panel 2)
Terre-Neuve	1,71	1693	1306	0,4312	0,5688
Ile-du-Prince-Édouard	1,92	568	883	0,25095	0,74905
Nouvelle-Écosse	4,02	1788	1986	0,18298	0,81702
Nouveau-Brunswick	1,94	1752	1819	0,33176	0,66824
Québec	2,75	4908	5788	0,23568	0,76432
Ontario	2,87	6893	8901	0,21249	0,78751
Manitoba	1,36	1776	2111	0,38219	0,61781
Saskatchewan	2,21	1915	1872	0,31642	0,68358
Alberta	2,87	2378	2113	0,28168	0,71832
Colombie-Britannique	2,76	2238	2509	0,24425	0,75575

d_i = effet de plan d'échantillonnage pour le panel i

n_i = nombre de personnes longitudinales de 16 ans et plus du panel i répondantes pour l'année de référence 1997

p, 1 - p = facteurs d'allocation des panels

D'après ce tableau, on constate qu'une plus grande importance est accordée au deuxième panel dans des proportions de 3 pour 1. Lors du dernier remaniement post-censitaire de l'EPA (qui s'est terminé à la fin de 1994), le plan de sondage de l'EPA s'est amélioré; ce qui fait que le rapport d_1 / d_2 est élevé, donnant ainsi une plus grande importance au second panel. Lorsque les deuxième et troisième panels de l'EDTR seront combinés, le rapport des effets de plan seront plus près de 1 puisque les deux panels auront été sélectionnés à partir du même plan de sondage de l'EPA. Les facteurs d'allocation des panels seront donc plus près de 0,5, accordant ainsi une importance similaire aux deux panels.

Les facteurs *p* sont appliqués à toutes les personnes longitudinales du premier panel qui sont dans un ménage répondant au temps t (dans ce cas-ci, au 31 décembre 1997). Les facteurs *1-p* devraient être appliqués à toutes les personnes longitudinales du deuxième panel qui sont dans un ménage répondant au temps t, à l'exception des personnes en dehors de la population cible au moment de l'introduction du premier panel (en janvier 1993) et présents dans la population cible lors de l'introduction du second panel (en janvier 1996). Pour des raisons pratiques, seuls les nouveaux-nés et les immigrants absents lors de l'introduction du premier panel et présents lors de l'introduction du second panel sont exclus lors de l'application des facteurs *1-p*. Les personnes exclues conserveront donc leur poids ajusté pour la non-réponse longitudinale. Ainsi le poids transversal ajusté pour la combinaison des deux panels est le suivant :

$$w_{1997,p}' \begin{cases} p\, w_{1997,ajust}, & \text{pour les personnes du premier panel} \\ (1-p)\, w_{1997,ajust}, & \text{pour les personnes du second panel non exclues pour l'application des facteurs d'allocation des panels} \\ w_{1997,ajust}, & \text{pour les personnes du second panel exclues pour l'application des facteurs d'allocation des panels} \end{cases}$$

où $w_{1997,p}$ = poids transversal ajusté pour la combinaison des deux panels
$w_{1997,ajust}$ = poids transversal ajusté pour la non-réponse longitudinale
p = facteur d'allocation des panels.

Ce poids est identique pour les poids transversaux individu usuel et intégré. Pour le poids transversal travail, le poids ajusté pour la combinaison des panels est calculé de la même façon mais où la notion de ménage répondant a été changée (section 4.1). Il est à noter que les facteurs d'allocation des panels sont appliqués aux poids après l'étape d'ajustement pour la non-réponse mais avant l'étape du partage de poids, pour que ces facteurs soient aussi inclus dans le poids des cohabitants.

4.4 Partage des poids

L'étape du partage des poids est essentielle et unique à la pondération transversale. Elle est engendrée par la nature longitudinale de l'EDTR. Dans une enquête longitudinale, l'échantillon évolue à travers le temps. De nouvelles personnes viennent habiter avec les individus déjà sélectionnés, des enfants naissent dans certains ménages échantillonnés. Bref, d'autres personnes qui n'ont pas été échantillonnées initialement font désormais partie de l'échantillon. Puisqu'on ne connaît pas la probabilité de sélection des personnes qui se sont jointes à l'enquête (les cohabitants), il faut trouver une façon de leur assigner un poids. Cela est fait en utilisant la méthode du partage des poids. Les prochains paragraphes

expliquent brièvement la méthodologie du partage des poids. Pour plus de détails sur la méthode du partage des poids, l'article de Lavallée (1995) est recommandé.

À partir de l'étape du partage de poids, on ne fait plus de distinction entre les deux panels et on traite tous les individus comme faisant partie d'un même échantillon, tout en conservant les notions d'individu longitudinal et de cohabitant. On pourrait penser qu'en choisissant d'effectuer le partage de poids au niveau du ménage, traiter les deux échantillons transversaux de façon individuelle ou de façon combinée reviendrait au même. Or il faut penser qu'en présence d'une enquête à panels multiples se chevauchant, il est possible d'observer des ménages pour lesquels des personnes issues du premier panel cohabitent avec des personnes issues du second panel. Lors de l'introduction du deuxième panel, tout ménage sélectionné pour faire parti de l'échantillon longitudinal comportant au moins une personne longitudinale du premier panel a été exclue de l'échantillon longitudinal du second panel afin de ne pas enquêter le ménage pour une autre période de six ans. Cependant, étant donné la nature longitudinale et dynamique de l'enquête, il se peut qu'au cours des six années, des ménages se séparent et de nouveaux ménages se forment, rendant ainsi possible le fait d'observer un ménage composé de personnes provenant des deux panels. En considérant les échantillons transversaux séparément, deux ensembles de poids distincts seraient produits pour ces ménages; ce qui causerait toutes sortes de problèmes. Il est alors préférable de considérer les deux panels ensemble.

Pour effectuer le partage des poids, il faut d'abord déterminer à quel niveau dans la hiérarchie du plan de sondage on veut l'effectuer : au niveau de la province, de la strate, de l'unité primaire d'échantillonnage ou de l'unité finale d'échantillonnage. L'EDTR a choisi d'effectuer le partage des poids au niveau de l'unité finale d'échantillonnage, le ménage. Une fois cela décidé, il s'agit de déterminer le statut de chaque personne dans l'échantillon transversal, selon que la personne est un individu longitudinal, un cohabitant initialement présent ou un cohabitant initialement absent.

Un cohabitant est initialement présent s'il faisait partie de la population cible au moment de la sélection du deuxième panel (en janvier 1996). À l'inverse, un cohabitant est initialement absent s'il ne faisait pas partie de la population cible au moment de la sélection du deuxième panel (en janvier 1996). Étant donné qu'on traite ici le chevauchement de deux panels, les notions de cohabitant initialement absent et initialement présent doivent tenir compte du fait qu'on ne traite plus qu'un seul panel et qu'on doit traiter les deux panels ensemble. On doit alors considérer les deux panels au moment de la sélection du deuxième panel (en janvier 1996). Il est important de mentionner que tous les cohabitants de l'enquête ont un poids longitudinal de zéro étant donné qu'ils ne faisaient pas partie de l'échantillon longitudinal. Par contre, ils peuvent avoir un poids transversal différent de zéro qui leur est assigné par la méthode du partage des poids.

À l'étape du partage de poids, deux types de poids sont produits pour chaque individu : le poids individu (qui peut être différent pour chaque personne à l'intérieur d'un ménage donné), et le poids intégré (poids identique pour tous les membres d'un même ménage).

Dans le cas du partage des poids pour le poids individu, il faut procéder selon les étapes suivantes. D'abord, il faut identifier les ménages auxquels des cohabitants se sont greffés depuis la sélection de l'échantillon longitudinal. Pour les ménages composés exclusivement de personnes longitudinales, il n'est pas nécessaire d'appliquer la méthode du partage des poids. Par conséquent, pour tous les individus de ces ménages, le poids transversal partagé est identique au poids transversal ajusté après l'application des facteurs d'allocation des panels.

Ensuite, pour les ménages où des cohabitants sont arrivés, il s'agit de faire la somme des poids transversaux ajustés pour la non-réponse $w_{1997,p}$ de tous les individus longitudinaux à l'intérieur du ménage. Puis, il faut établir le nombre d'individus longitudinaux et le nombre de cohabitants initialement présents dans chaque ménage.

Pour les ménages où au moins un des cohabitants est initialement présent, le poids transversal partagé de tous les individus d'un ménage donné est calculé comme étant la somme des poids transversaux $w_{1997,p}$ de tous les individus longitudinaux du ménage divisé par le nombre total d'individus longitudinaux et de cohabitants initialement présents dans le ménage.

Pour les ménages où tous les cohabitants sont initialement absents, le poids transversal partagé des individus longitudinaux et des cohabitants est calculé différemment. Pour les individus longitudinaux faisant partie de ces ménages, le poids transversal partagé est égal au poids transversal ajusté après l'application des facteurs d'allocation des panels. Pour les cohabitants, le poids transversal partagé est égal à la somme des poids $w_{1997,p}$ de tous les individus longitudinaux du ménage divisé par le nombre total d'individus longitudinaux. De cette façon, un poids moyen ne sera pas assigné aux individus longitudinaux et l'ajustement pour la non-réponse sera préservé. Seuls les individus initialement absents recevront le poids moyen des individus longitudinaux du ménage.

Le poids transversal individu partagé (poids individu usuel ou poids travail) est alors défini comme étant :

$$w_{1997,part}' \begin{cases} w_{1997,p}, & \textit{pour les personnes dans les mén. composés exclusivement de personnes longitudinales} \\ & \textit{et pour les personnes longitudinales dans les ménages où tous les cohabitants sont IA} \\ \dfrac{\sum_h w_{1997,p}}{n_{L,h} + n_{IP,h}}, & \textit{pour les personnes dans les ménages ayant au moins un cohabitant IP} \\ & \textit{et pour les cohabitants dans les ménages où tous les cohabitants sont IA} \end{cases}$$

où $w_{1997,part}$ = poids transversal partagé
$w_{1997,p}$ = poids transversal ajusté pour la combinaison des deux panels
IP = initialement présent
IA = initialement absent
$n_{L,h}$ = nombre de personnes longitudinales dans le ménage *h*
$n_{IP,h}$ = nombre de cohabitants initialement présents dans le ménage *h*.

Dans le cas du partage des poids pour le poids intégré, le processus est quelque peu différent. Tout individu qui aura un poids transversal individu non nul (poids individu usuel seulement) aura tout de même un poids transversal intégré non nul. Ceci est rendu nécessaire par le fait que le poids avant stratification a posteriori doit être identique pour tous les gens d'un ménage afin de pouvoir produire un poids intégré (Lemaître, Dufour, 1987). La différence entre les poids intégrés et les poids individu provient du fait qu'il faut effectuer le partage sur tous les ménages, peu importe que des cohabitants s'y trouvent ou non. Par conséquent, pour tous les ménages, le poids partagé est calculé de la même façon pour tous les individus lorsque le poids produit est intégré. Le poids transversal partagé intégré est donc défini comme étant :

$$w_{1997,part} \quad , \quad \frac{\sum_h w_{1997,p}}{n_{L,h} \,\%\, n_{IP,h}}$$

où $w_{1997,part}$ = poids transversal partagé
$w_{1997,p}$ = poids transversal ajusté après la combinaison des deux panels
$n_{L,h}$ = nombre de personnes longitudinales dans le ménage *h*
$n_{IP,h}$ = nombre de cohabitants initialement présents dans le ménage *h*.

Pour l'année de référence 1997, un total de 80 834 personnes font partie de l'échantillon transversal. Ce nombre comprend les personnes répondantes et non-répondantes (pour les ménages non-répondants, on utilise la dernière composition connue de ces ménages). Il est sous-représentatif du nombre réel de personnes dans l'échantillon transversal puisque pour les ménages non-répondants, on a aucune information sur la composition actuelle des ménages. Parmi les 80 834 personnes, 79 218 ont des poids transversaux individu usuel et intégré non nuls et 76 421 ont un poids transversal travail non nul.

4.5 Ajustement pour la migration inter-provinciale

L'étape de l'ajustement pour la migration inter-provinciale est une autre étape directement engendrée par la nature longitudinale de l'EDTR. À travers les années, il arrive que certaines personnes échantillonnées déménagent d'une province à une autre. Étant donné que la probabilité de sélection peut varier beaucoup selon la géographie, un grand écart est observé, à l'échelle canadienne, entre les poids les plus grands et les poids les plus petits. Par conséquent, si une personne avec une probabilité de sélection assez faible (donc un poids assez grand) déménage dans un endroit où la probabilité de sélection des gens est beaucoup plus forte (donc des poids beaucoup plus petits), il y a potentiellement matière à problème. En effet, du point de vue analytique, les résultats d'analyse peuvent se trouver faussés à cause de certains cas de migration inter-provinciale.

Pour expliquer le problème, prenons un exemple simple. Supposons qu'un chirurgien cardiaque déménage de Toronto à Charlottetown. Cette situation causerait plusieurs problèmes. D'abord, son poids qui s'établirait aux environs de 400 serait beaucoup plus grand que tous les poids que l'on retrouve à l'Île-du-Prince-Édouard et qui sont en moyenne aux environs de 65 À lui seul, il représenterait tout à coup environ 3.33% de la population de la province, soit environ 5000 personnes sur 150 000, ce qui est totalement inadmissible. Ensuite, du point de vue analytique, il viendrait artificiellement augmenter le revenu total des gens à cause de son poids de 5000 et de son salaire.

Plusieurs options sont envisageables pour atténuer le problème de poids aberrants causés par la migration inter-provinciale. Ces poids extrêmes pourraient être remplacés par le poids moyen, la médiane ou un des quantiles de la distribution des poids des personnes de la province où les gens déménagent, ou toute autre possibilité basée sur cette distribution. Ces alternatives sont arbitraires et pourraient toutes être utilisées. Il s'agit en fait de trouver un compromis entre le fait d'attribuer à la personne migrante un poids similaire à la moyenne des poids des personnes de sa province de destination et le fait de ne pas trop modifier son poids d'origine (basé sur la probabilité de sélection). L'objectif ici est tout d'abord d'identifier les personnes qui ont migré dans une autre province dont leur poids contribue trop fortement à l'estimation de leur province de destination, et ensuite d'ajuster ce poids.

L'EDTR a adoptée une méthode utilisant le 95^{e} percentile de la distribution des poids. Cette méthode a l'avantage d'être simple, rapide et de ne pas nécessiter d'intervention manuelle (ou presque). Il s'agit tout d'abord d'identifier toutes les personnes qui ont déménagé dans une province particulière dont leur poids est plus grand que le maximum des poids des personnes originaires de cette province. Si c'est le cas, le poids de ces personnes est ramené au 95^{e} percentile de la distribution des poids des personnes originaires de la province. Ensuite un nombre aléatoire entre 0.00 et 10.00 est soustrait de ce percentile afin de conserver la "presque" unicité des poids (il est possible d'avoir des poids individus égaux mais cela est très rare).

Les ajustements pour la migration inter-provinciale sont effectués indépendamment pour les poids individus et intégrés et sont redérivés à chaque année de traitement. Pour le poids intégré, l'ajustement est fait au niveau ménage afin de préserver l'égalité des poids à l'intérieur d'un ménage. Le poids transversal ajusté pour la migration inter-provinciale est définit comme étant :

$$w_{1997,mig} \text{ ' } w_{1997,part}(a_{mig}$$

où $w_{1997,mig}$ = poids transversal ajusté pour la migration inter-provinciale
$w_{1997,part}$ = poids transversal partagé
a_{mig} = facteur d'ajustement pour la migration inter-provinciale.

Pour l'année de référence 1997, un ajustement de migration inter-provinciale s'est avéré nécessaire pour 36 personnes pour le poids transversal individu usuel, pour 15 ménages (43 personnes) pour le poids transversal intégré et pour 34 personnes pour le poids transversal travail.

4.6 Ajustement pour les valeurs influentes

Comme toute enquête essayant d'estimer la distribution des revenus des gens, l'EDTR est vraisemblablement confrontée au problème des observations influentes (ces distributions ayant de très longues queues). Les observations influentes ont un impact important sur les estimations provinciales du total et de la moyenne du revenu ainsi que sur leurs estimations de variance. L'impact devient encore plus accentué sur les estimations de plus petits domaines comme les estimations selon le groupe d'âge, le sexe ou la taille des familles. Pour diminuer l'effet de ces observations sur les estimations, il faut d'abord trouver un moyen d'identifier ces observations, puis de développer une méthode pour ajuster ces observations de telle sorte qu'elle permette de réduire l'influence de celles-ci sur les estimations.

Il est à noter que les étapes d'identification des observations influentes et de calcul des facteurs d'ajustement sont faites sur les poids individus seulement (poids individus usuels ou poids travail), mais que les facteurs d'ajustement dérivés sont appliqués aussi bien sur les poids individus que sur les poids intégrés.

Une méthode de détection et de traitement des valeurs influentes a été mise sur pied par l'EFC (Tremblay, 1998). Comme on savait que l'EFC serait intégrée à l'EDTR à la collecte de 1999, la méthodologie développée tient compte du fait que l'enquête qui remplacerait l'EFC est de nature longitudinale. Cette approche a donc été adoptée par l'EDTR pour

traiter le problème des revenus influents. La méthode consiste d'abord à identifier les observations influentes, c'est-à-dire celles qui contribuent trop fortement à l'estimation pondérée du total des revenus individuels de sa province. Il s'agit en général d'observations qui se trouvent dans une tranche de revenu élevé. Le revenu d'un individu (on parle ici du revenu total moins les gains en capital) est considéré comme influent si la contribution de son revenu pondéré par rapport à l'estimation pondérée provinciale dépasse un certain seuil. Ce seuil correspond à la borne inférieure de l'intervalle calculée à partir de la méthode des quartiles. Il se définit comme étant MED + k (Q3 - MED), où MED et Q3 sont respectivement la médiane et le troisième quartile de la contribution pondérée des individus, et k est un paramètre fixe pour chaque province qui a été déterminé après avoir examiné les observations ayant les pourcentages de contribution les plus élevés à l'estimation provinciale. (La valeur de k est de 25 pour toutes les provinces, exceptés pour le Québec et l'Ontario où k est égal à 75.)

Une fois que les observations influentes sont identifiées, il faut maintenant déterminer les facteurs d'ajustement applicables aux poids associés à ces observations. Pour ce faire, on évalue si l'échantillon surreprésente les revenus supérieurs à 99 999$ par rapport à la distribution des revenus de la population des déclarants de données fiscales (distribution obtenue du fichier T1 de Revenu Canada). Si c'est le cas, le poids est ajusté de sorte que la distribution de l'échantillon pondéré coïncide avec celle des données fiscales sauf dans le cas où cette différence est faible (ce qui vise à maintenir au minimum les ajustements de poids). L'ajustement de ce poids se fait en calculant un certain facteur d'ajustement, ce facteur sera compris entre 0 et 1. Plus précisément, voici comment le calcul du facteur d'ajustement est effectué pour chaque individu ayant un revenu influent. Notons que ces calculs sont faits au niveau provincial.

> "*Si une province contient plus d'un individu influent, le facteur d'ajustement est d'abord calculé pour l'individu influent avec le revenu (non pondéré) le plus élevé. Supposons que cet individu est dans la classe de revenu c. L'ajustement fera en sorte que l'estimation du nombre d'individus qui se trouve dans la classe c ou dans toute autre classe de revenu supérieure, soit égale au nombre d'individus dans ces mêmes classes sur le fichier de données fiscales, sujet toutefois à certaines contraintes. D'abord, uniquement les poids des individus influents devront être modifiés dans cet ajustement. De plus, si l'ajustement entraîne une hausse du poids, c'est-à-dire que le facteur d'ajustement est supérieur à 1, l'ajustement ne sera pas appliqué et le poids initial sera conservé. Si le fichier de données fiscales indique qu'il n'y a aucun individu dans les classes de revenu, l'ajustement ramènera le poids à 1. De la même façon, on ramènera le poids à 1 si le nombre d'individus estimés dans ces classes à partir des individus non influents seulement, dépasse déjà le nombre d'individus selon le fichier de données fiscales.*" (Tremblay, 1998).

Le calcul du facteur d'ajustement est dépendant d'un individu à l'autre. Lorsque le facteur d'ajustement a été calculé pour l'individu influent avec le plus haut revenu, c'est le poids ajusté de cet individu qui sera considéré dans le calcul du facteur d'ajustement de l'individu suivant.

Bien que le facteur d'ajustement soit calculé au niveau de l'individu, il devra être appliqué au poids de le personne identifiée comme influente ainsi qu'aux poids de toutes les autres personnes habitant le même ménage qu'elle et ce, pour le reste de la durée du panel même si cette personne n'est plus considérée influente dans les vagues subséquentes à celle où elle a été identifiée. L'aspect longitudinal de l'enquête entraîne cette procédure afin de préserver la cohérence à travers les années, même dans le contexte de la pondération transversale. Pour les ménages où aucun individu n'a été traité pour la valeur extrême de son revenu, le facteur d'ajustement est de 1 pour tous les membres de ces ménages.

Afin de détecter les valeurs influentes, la méthode utilise le poids obtenu après la stratification a posteriori (dont il est question à la section suivante) puisque cette dernière repose sur la distribution des revenus de la population. Il faut donc faire une stratification a posteriori temporaire avant d'effectuer l'étape de l'ajustement pour les valeurs influentes, puis utiliser les poids temporaires post-stratifiés pour identifier et traiter les observations influentes. Une fois les facteurs d'ajustement calculés, la stratification a posteriori doit être refaite en appliquant ces facteurs aux poids que l'on avait avant cette stratification a posteriori temporaire (donc aux poids transversaux ajustés pour la migration inter-provinciale $w_{1997,mig}$).

Le poids transversal après ajustement pour les valeurs influentes est donc calculé comme suit :

$$w_{1997,infl} \quad ' \quad w_{1997,mig} (ß_{infl}$$

où $w_{1997,infl}$ = poids transversal ajusté pour les valeurs influentes
$w_{1997,mig}$ = poids transversal ajusté pour la migration inter-provinciale
$ß_{infl}$ = facteur d'ajustement pour les valeurs influentes.

Pour l'année de référence 1997, deux personnes ont été identifiées comme ayant des revenus influents. De plus, les poids de sept autres personnes identifiées comme étant influentes lors des vagues précédentes doivent à nouveau être ajustés. En incluant les individus qui habitent avec toutes ces personnes, l'ajustement pour les valeurs influentes a été appliqué à un total de 29 personnes pour l'année de référence 1997. Les facteurs d'ajustement apportés aux poids de ces personnes variaient de 0,39 à 0,93, la plupart des facteurs étant autour de 0,50.

4.7 Stratification a posteriori

Rappelons que le but de la stratification a posteriori (ou post-stratification) est de faire en sorte que la somme des poids à l'intérieur de certains sous-groupes de l'échantillon (les post-strates) corresponde aux totaux de contrôle de la population connus pour ces post-strates, pour une année donnée.

Pour les poids transversaux individus, la stratification a posteriori est accomplie de la même manière que celle effectuée par la pondération longitudinale (section 3.5). Le poids après la stratification a posteriori est égal au poids avant la stratification a posteriori, multiplié par le total de contrôle de la population de la post-strate à laquelle l'individu appartient, divisé par la somme des poids avant la stratification a posteriori de cette post-strate.

Pour produire le poids intégré, la méthodologie est un peu plus complexe. Tout en respectant les totaux de contrôle pour chaque post-strate, la stratification a posteriori doit aussi s'assurer que tous les individus d'un même ménage ont le même poids final. Pour plus d'informations sur la façon de procéder, l'article de Lemaître et Dufour (1987) se veut la référence en la matière. Notons par $w_{1997,ps}$, le poids transversal après la stratification a posteriori.

Comme pour la pondération longitudinale, les post-strates pour la pondération transversale sont un croisement des trois variables suivantes : province, sexe et groupe d'âge. La province fait référence ici à la province de résidence au 31 décembre de l'année de

référence. De même, le groupe d'âge auquel un individu appartient est déterminé en fonction de son âge au 31 décembre de l'année de référence. Les valeurs possibles pour le groupe d'âge sont celles présentées précédemment dans le tableau 3.2.

Pour l'année de référence 1997, les poids transversaux individus usuels après l'ajustement pour les valeurs influentes ont été en moyenne multipliés par 1,13 pour permettre le redressement de l'échantillon de l'EDTR. Les poids transversaux intégrés ont aussi été multipliés en moyenne par 1,13 lors de la stratification a posteriori. Pour la pondération travail, un ajustement moyen de 1,18 a été apporté aux poids transversaux travail pour effectuer le redressement de l'échantillon.

4.8 Ajout de bruit

Comme pour le poids longitudinal, une étape d'ajout de bruit a été ajoutée pour des raisons de confidentialité. Il est important de mentionner que dans le cas de la pondération transversale, l'ajout de bruit s'applique seulement au poids individu et non au poids intégré. En effet, la présence d'un poids intégré sur le fichier de microdonnées à grande diffusion serait contraire à l'objectif d'éviter la reconstitution des ménages à partir de ce fichier. Par conséquent, seul le poids individu sera perturbé et inclut sur le fichier de microdonnées à grande diffusion.

L'ajout de bruit dans les poids transversaux individus est effectué selon les mêmes étapes que celles de la pondération longitudinale (voir section 3.6). La perturbation aléatoire e, comprise entre 0 et 1, et la perturbation a (perturbation transversale) introduites dans les poids transversaux individus sont déterminées de la même façon que pour les poids longitudinaux. Le poids transversal individu après l'ajout de bruit est obtenu de la façon suivante :

$$w_{bruit} = \begin{cases} w_{ps} \pm (e\%a), & \textit{pour les paires d'individus dont le poids est identique à l'intérieur d'un ménage} \\ w_{ps}, & \textit{dans les autres cas} \end{cases}$$

où w_{bruit} = poids transversal après l'ajout de bruit
w_{ps} = poids transversal après la stratification a posteriori
e = perturbation aléatoire
a = perturbation transversale.

Pour l'année référence 1997, le facteur a des années références 1995 et 1996, qui était égal à 1, a été réutilisé pour les poids individu usuel. Pour le poids travail, un facteur a de 1,25 a été utilisé.

4.9 Description des poids transversaux finaux produits

La pondération transversale produit quatre ensembles de poids transversaux différents, un poids intégré et trois différents poids individus : le poids individu usuel externe, le poids travail interne et le poids travail externe. Le poids intégré et le poids travail interne sont produits à la fin de l'étape de la stratification a posteriori (section 4.7), alors que le poids individu usuel externe et le poids travail externe sont produits à la fin de l'étape de l'ajout

de bruit dans les poids (section 4.8). Le tableau 4.3 donne les médianes des poids transversaux initiaux de certains poids transversaux finaux, pour l'années de référence 1997. Seuls les individus avec des poids finaux non nuls sont utilisés dans le calcul de ces médianes. La province donnée dans le tableau fait référence à la province de résidence au 31 décembre 1997.

Tableau 4.2 : Médianes des poids transversaux initiaux et finaux, pour l'année de référence 1997

Province	Poids initiaux	Poids finaux		
		Poids intégré	Poids individu usuel externe	Poids travail interne
Terre Neuve	133	127	129	133
Île-du-Prince-Édouard	71	71	75	77
Nouvelle-Écosse	105	113	113	117
Nouveau-Brunswick	150	150	155	159
Québec	270	309	309	318
Ontario	253	301	302	314
Manitoba	148	167	171	180
Saskatchewan	161	163	165	170
Alberta	261	303	302	316
Colombie-Britannique	305	421	411	426

5. Changements à venir et développements futurs

Afin d'améliorer la qualité et la fiabilité des estimations, les étapes de pondérations longitudinale et transversale de l'EDTR sont régulièrement révisées. Ces révisions sont effectuées en tenant compte de toute modification ayant trait aux objectifs, au traitement et à la diffusion.

Un des changements à venir concerne la stratégie de diffusion des microdonnées de l'EDTR. Avant l'intégration de l'EFC à l'EDTR (comme c'est le cas pour l'année de référence 1997), la stratégie de diffusion consistait à produire annuellement un ensemble de fichiers transversaux et longitudinaux de microdonnées à grande diffusion pour lesquels la reconstitution des ménages au sein de ces fichiers n'étaient pas possible. En raison de l'intégration de l'EFC à l'EDTR, il est maintenant devenu une priorité de diffuser des fichiers transversaux qui satisfont les besoins des utilisateurs de l'EFC. Pour ce faire, le contenu des fichiers de microdonnées transversaux à grande diffusion doit être redéfini et doit inclure des identificateurs de ménage et de famille pour permettre la reconstitution des ménages et des familles.

La diffusion de tels fichiers transversaux remets en question la diffusion de fichiers de microdonnées longitudinales à grande diffusion. On pense que le risque de divulgation associé aux fichiers longitudinaux permettant la reconstitution des ménages et des familles serait trop élevé, et qu'il serait probablement difficile de créer des fichiers longitudinaux pour lesquels le risque d'appariement avec les fichiers transversaux serait bas. Pour des raisons prioritaires, il a été décidé de se concentrer sur la stratégie de diffusion des fichiers de microdonnées transversales et de considérer d'autres possibilités pour rendre les données longitudinales disponibles aux utilisateurs à l'extérieur de Statistique Canada. Parmi ces possibilités, il y a le télé-accès des données et le centre d'accès aux données de recherche. Pour plus de renseignements sur la stratégie de contrôle de la divulgation des microdonnées, se référer à l'article de Nadeau, Gagnon et Latouche (1999).
Comme la diffusion d'un fichier de microdonnées longitudinales à grande diffusion est dorénavant compromise, le production du poids longitudinal externe ne semble plus nécessaire. De plus, avec la nouvelle stratégie de diffusion permettant la reconstitution des ménages et des familles à partir des fichiers transversaux, le poids transversal individu externe n'est plus utile. Pour l'instant, ces deux poids sont toujours produits mais l'abolition de ces derniers sera discutée dans un proche avenir.

Un autre changement qui sera apporté à l'enquête est l'implantation de l'imputation des données sur le travail principalement pour les personnes non-répondantes à l'interview sur le travail mais répondante à l'interview sur le revenu. Avant l'intégration de l'EFC à l'EDTR, les fichiers diffusés pouvaient comporter des données manquantes pour les caractéristiques du travail. Avec l'intégration de l'EFC à l'EDTR, on veut maintenant fournir des fichiers de microdonnées complets, sans aucune donnée manquante. Considérant cela, l'utilité du poids travail est remise en question. La non-réponse à l'interview sur le travail étant maintenant compensée en imputant les données, on ne verrait plus l'utilité de produire un poids travail qui compense pour la non-réponse en ne considérant dans la pondération que les répondants à l'interview sur le travail. Le poids travail est toujours produit, mais une fois que toutes les procédures d'imputation des données sur le travail seront en place, l'élimination de ce poids sera considérée plus sérieusement.

Un changement au niveau du redressement de l'échantillon transversal est aussi prévu. L'EDTR révise présentement sa stratégie de redressement de l'échantillon dans le but d'améliorer les estimations et d'harmoniser les concepts entre les différentes enquêtes-ménages. Jusqu'à maintenant, l'EDTR utilise des totaux de contrôle au niveau de la province, du groupe d'âge et du sexe pour effectuer le redressement de son échantillon

transversal. Des nouveaux ensembles de totaux de contrôle sont envisagés : le nombre de familles économiques de taille 1, 2 et 3 et plus par province, et le niveau de revenu (bas, moyen, élevé) par province.

De tels ajouts remettent en question la définition des ensembles de totaux de contrôle utilisés jusqu'à maintenant, plus particulièrement au niveau des catégories de groupe d'âge. L'ajout de ces nouveaux ensembles de totaux entraîne aussi la modification de la méthode utilisée pour effectuer le redressement de l'échantillon : la stratification a posteriori devra être remplacée par un calage sur marges. L'équipe de méthodologie de l'EDTR révise présentement sa stratégie pour permettre d'intégrer facilement ces nouveaux totaux à la méthode déjà en place pour effectuer le redressement de l'échantillon transversal. L'impact d'un tel changement sur les poids sera aussi étudié en profondeur.

Au niveau de la pondération longitudinale, une des nouveautés à venir est la production d'un poids qui combinerait les deux panels de l'EDTR. Le défit ici sera de définir la population longitudinale pour plusieurs panels. Actuellement, la pondération longitudinale considère les deux panels de façon indépendante, produisant des poids pour chacun des panels indépendamment l'un de l'autre. Le calcul d'un poids longitudinal qui utiliserait les répondants des deux panels permettrait d'accroître la précision des estimations lors d'analyses longitudinales (puisque la taille d'échantillon serait alors doublée). Pour produire un tel poids, on s'inspirera de la méthode utilisée pour combiner les échantillons transversaux de l'EDTR.

6. Conclusion

L'objectif de l'EDTR de fournir des données longitudinales et transversales pour plusieurs caractéristiques reliées à la dynamique du travail et du revenu, qui sont représentatives de la population des dix provinces canadiennes, est un grand défi. Les pondérations longitudinale et transversale expliquées en détails dans ce document permettent de produire des poids longitudinaux et transversaux nécessaires à ces fins. Pour produire ces poids, plusieurs étapes sont requises : la détermination des poids initiaux, l'ajustement pour la non-réponse, le partage des poids (seulement pour les poids transversaux), les ajustements pour la migration inter-provinciale (seulement pour les poids transversaux) et les valeurs influentes, la stratification a posteriori et l'ajout de bruit dans les poids. Au fil des ans, ces étapes ont été améliorées et optimisées mais avec l'évolution des stratégies de traitement et de diffusion, de nouvelles mises au point sont nécessaires pour entre autre préserver la qualité et la fiabilité des données. Par exemple, l'ajout de nouvelles post-strates pour le redressement de l'échantillon transversal, la production d'un poids longitudinal qui combine les deux panels et le changement dans la stratégie de diffusion qui remettra en question l'utilité de certains poids longitudinaux et transversaux.

Remerciements

Les auteurs tiennent tout d'abord à remercier Christian Nadeau pour son support et ses précieux conseils tout au long de la rédaction de ce document. Ils tiennent également à remercier Michel Latouche, Guy Laflamme et Martin St-Pierre pour leurs commentaires qui ont permis d'améliorer la qualité de ce document.

Bibliographie

Cotton, C., Bishop, K., Giles, P., Hewer, P. et Saint-Pierre, Y. (1999). *Comparaison des résultats de l'Enquête sur la dynamique du travail et du revenu (EDTR) et de l'Enquête sur les finances des consommateurs (EFC) 1993-1997 : mise à jour*, publication de Statistique Canada NE75F0002MIF - 99007.

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 Survey Research Methods Section, American Statistical Association.

Durning, A. (1994). *Weighting of the SLID Preliminary File*, document interne, Statistique Canada.

Franklin, S. (1999a). *Modelling SLID's Nonresponse at the Provincial Level*, document interne, Statistique Canada.

Franklin, S. (1999b). *Study of the Impact of Modifying SLID's Nonresponse Model*, document interne, Statistique Canada.

Franklin, S. et Lévesque, I. (1999). *L'ajout de bruit aux poids post-stratifiés*, document interne, Statistique Canada.

Gagnon, F. (1997). *Pondération longitudinale - Deuxième vague de l'EDTR*, document interne, Statistique Canada.

Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B. et Lindeyer, J. (1998). *Méthodologie de l'enquête sur la population active du Canada*, Statistique Canada, Catalogue 71-526-XPB.

Grondin, C. (1996). *Pondération longitudinale - Première vague de l'EDTR*, document interne, Statistique Canada.

Kalton, G. et Kasprzyk, D. (1986). *Le traitement des données d'enquête manquantes*, Techniques d'enquête, vol. 12, no 1, pp. 1-17.

Knowledge SEEKER IV, Version 4.2.2, Copyright 1996, Angoss International Limited.

Latouche, M., Dufour, J. et Merkouris, T. (2000). *SLID Cross-Sectional Weighting: Combining Two or More Panels*, document interne à paraître, Statistique Canada.

Latouche, M. et Michaud, S. (1997). *Concerns pertaining to weighting of longitudinal surveys*, Proceedings of the Section on Government Statistics and Section on Social Statistics, American Statistical Association.

Lavallée, P. (1994). *Ajout du second panel à l'EDTR: sélection et pondération*, document interne, Statistique Canada.

Lavallée, P. (1995). *Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage de poids*, Techniques d'enquête, vol. 21, no 1, pp. 27-35.

Lavigne, M. et Michaud, S. (1998). *Aspects généraux de l'enquête sur la dynamique du travail et du revenu*, document de travail sur la dynamique du revenu et du travail, Statistique Canada, 75F0002M No. 98-05.

Lemaître, G. et Dufour, J. (1987). *Une méthode intégrée de pondération des personnes et des familles*, Techniques d'enquête, vol. 13, no 2, pp. 211-220.

Merkouris, T. (1999). *Cross-sectional Estimation in Multiple-Panel Household Surveys*, document de travail de la direction de méthodologie, Statistique Canada, HSMD-99-004E.

Nadeau, C., Gagnon, É. et Latouche, M. (1999). *Disclosure Control Strategy for the Release of Microdata in the Canadian Survey of Labour and Income Dynamics*, Proceedings of the Survey Research Methods Section, American Statistical Association.

Renaud, M. (1997). *Pondération transversale - Deuxième vague de l'EDTR*, document interne, Statistique Canada.

Singh, M.P., Drew, J.D., Gambino, J.G. et Mayda, F. (1990). *Méthodologie de l'enquête sur la population active du Canada*, Statistique Canada, Catalogue 71-526.

Statistique Canada (1997). *Répartition du revenu au Canada selon la taille du revenu*, publication de Statistique Canada NE13-207-XPB, pp. 52-59.

Tambay, J. L., Schiopu-Kratina, I., Mayda, J., Stukel, D. et Nadon, S. (1998). *Traitement de la non-réponse du cycle deux de l'enquête sur la santé de la population*, Techniques d'enquête, vol. 24, no 2, pp. 159-169.

Tremblay, J. (1998). *Détection des observations influentes pour l'enquête sur les finances des consommateurs (EFC) et l'enquête sur la dynamique du travail et du revenu (EDTR)*, document interne, Statistique Canada.

Annexe A : Liste des variables utilisées dans le modèle d'ajustement pour la non-réponse

Les variables dichotomiques suivantes ont été utilisées pour modéliser la non-réponse pour la cinquième vague du premier panel.

Aborig	- autochtone
Canada	- né au Canada
US	- né aux États-Unis
UK	- né au Royaume-Uni
Europe	- né en Europe
Autpays	- né ailleurs (pas au Canada, ni aux États-Unis, ni au Royaume-Uni, ni en Europe)
EOBR	- origine ethnique britannique
EOCAN	- origine ethnique canadienne
EOEUR	- origine ethnique européenne
EOFR	- origine ethnique française
IMSTAT	- statut d'immigration
VISMIN	- minorité visible
OWNER	- propriétaire
RURAL	- habite dans une région rurale (basé sur le secteur de dénombrement)
REMOTE	- habite dans une région éloignée (basé sur le secteur de dénombrement)
URBAIN	- habite dans une région urbaine (basé sur le secteur de dénombrement)
EMPLOYE	- personne occupée
NEMPLOYE	- chômeur
NLF	- inactif
MAR_UL	- marié ou vivant en union libre
SEP_DIV	- séparé ou divorcé
VEUF	- veuf ou veuve
CELIB	- célibataire, n'a jamais été marié
PROX	- l'interview préliminaire a été menée via une personne interposée
REV0	- indique si le revenu du ménage était manquant à l'interview préliminaire. Cette variable a été ajoutée pour l'année de référence 1997 puisque le taux de non-réponse pour cette variable était élevé : 29 %. Pour les années de référence précédentes, les données manquantes étaient incluses dans la même catégorie que REV1 (revenu du ménage <=10 000$).
REV1	- revenu du ménage <=10 000$
REV2	- revenu du ménage > 10 000$ et <=25 000$
REV3	- revenu du ménage > 25 000$ et <=50 000$
REV4	- revenu du ménage > 50 000$
PRESENF	- présence d'enfant(s) au moment de l'interview préliminaire
P10 – P59	- une variable pour chaque province
HOMME	- sexe
AGE_1	- age >=16 et <19
AGE_2	- age >=19 et <25
AGE_3	- age >=25 et <35
AGE_4	- age >=35 et <45
AGE_5	- age >=45 et <55
AGE_6	- age >=55 et <65
AGE_7	- age >=65
Educ1 – Educ9	- 9 niveaux de scolarité

Les variables dichotomiques suivantes ont été utilisées pour modéliser la non-réponse pour la deuxième vague du second panel.

Educ1	- 0-8 années de scolarité
Educ2	- 9-13 années de scolarité (études secondaires)
Educ3	- études post-secondaires excluant le collège communautaire, le CÉGEP et l'université
Educ4	- collège communautaire, CÉGEP ou université (incluant doctorat)
RENTER	- 1=propriétaire, 0=locataire
EMPLOYE	- personnes occupées
NON_EMPL	- chômeurs
NLF	- inactifs
MARIE	- marié ou vivant en union libre
SEP_DIV	- séparé ou divorcé
VEUF	- veuf ou veuve
CELIB	- célibataire, n'a jamais été marié
P10 – P59	- une variable par province
SEXE	- sexe
AGE_1	- age >=16 et <19
AGE_2	- age >=19 et <25
AGE_3	- age >=25 et <35
AGE_4	- age >=35 et <45
AGE_5	- age >=45 et <55
AGE_6	- age >=55 et <65
AGE_7	- age >=65
HHSZ1-HHSZ5	- nombre de personnes dans le ménage (HHSZ5=>5)
FAMTYPE1	- personne non apparentée
FAMTYPE2	- famille époux-épouse ou famille monoparentale sans jeune enfant
FAMTYPE3	- famille époux-épouse ou famille monoparentale avec des enfants âgés de 0 à 17 ans
FAMTYPE4	- autre genre de famille
STUDENTS	- 1=étudiant (à temps plein ou à temps partiel), 0=n'est pas un étudiant
CWORK1	- catégorie de travailleur = employé du secteur public
CWORK2	- catégorie de travailleur = employé du secteur privé
CWORK3	- est un travailleur indépendant du secteur privé avec ou sans employés ou est un travailleur familial non rémunéré