Division de la statistique du revenu

75F0002MIF - 00006

Pondération transversale : Combinaison de deux panels ou plus

Préparé par :
Michel Latouche
Johane Dufour
Takis Merkouris

Octobre 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-00006 au catalogue sur internet gratuitement. Pour obtenir un numéro de ce produit, les utilisateurs sont priés de se rendre à http://www.statcan.ca/cgi-bin/downpub/research_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

Pondération transversale : Combinaison de deux panels ou plus

Publication autorisée par le ministre responsable de Statistique Canada

octobre 2000

Nº 75F0002MPF - 00006 au catalogue
ISSN 0000-0000

Nº 75F0002MIF - 00006 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

TABLE DES MATIÈRES


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Résumé

Le présent document traite de méthodes et outils servant à produire des estimations transversales fondées sur la combinaison de deux panels longitudinaux pour l'Enquête sur la dynamique du travail et du revenu (EDTR). La méthode adoptée est semblable à la stratégie standard du traitement des plans d'échantillonnage à bases de sondage multiples. Il s'agit d'utiliser un estimateur combiné non standard qui accorde une importance relative aux panels en fonction de facteurs d'allocation des panels (*fap*). Ces facteurs sont déjà accessibles à l'étape de l'estimation. Au besoin, la méthode pourrait intégrer un troisième panel ou un échantillon transversal supplémentaire. L'examen de diverses stratégies a permis de choisir une combinaison de panels telle que la variance de l'estimation de niveau est minimisée. Le *nombre de personnes âgées de 15 ans ou plus* a été choisi comme variable pour le calcul des facteurs d'allocation des panels dans chaque province. Afin de simplifier le calcul des poids, il a été décidé d'obtenir les facteurs d'allocation des panels d'une source externe. Pour l'année de référence 1996, des données de l'Enquête sur la population active (EPA) et de l'EDTR ont servi à calculer ces facteurs. Dans l'ensemble, les données de 1996 indiquent que le recours à des facteurs d'allocation de panel optimal entraîne une augmentation intéressante de la précision et permet de réduire le biais éventuel dû à l'érosion. Les données suggèrent également qu'il existe entre les estimations tirées du panel 1 et du panel 2 des différences qui mériteraient une étude plus poussée.


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

1. Introduction

L'EDTR est une enquête longitudinale qui permet de produire non seulement des données longitudinales sur l'activité des individus sur le marché du travail, mais également des estimations transversales annuelles des caractéristiques du revenu des individus et des familles. Pour la première fois, l'EDTR de l'année de référence 1996 a eu recours à deux panels pour la production des estimations transversales. En théorie, chaque panel peut donner lieu à des estimations transversales. Toutefois, de meilleures estimations sont obtenues lorsque l'on combine les deux panels. De cette façon, les estimations sont obtenues d'un échantillon plus grand, ce qui permet de réduire la variance des estimations et d'utiliser des totaux de contrôle plus nombreux à l'étape du calage aux marges, ce qui réduit le biais potentiel et la variabilité encore davantage.

Une estimation combinée est une somme pondérée des estimations tirées des panels, les poids étant les facteurs d'allocation des panels (*fap*) (Merkouris, 1999). Le recours à un estimateur combiné n'est pas nouveau à Statistique Canada. L'Enquête sur la population active (Singh, Drew, Gambino et Mayda., 1990) a toujours fait appel à des facteurs d'allocation pour combiner les six groupes de renouvellement de l'enquête. L'Enquête sur les finances des consommateurs combine elle aussi les quatre groupes de renouvellement retenus de l'EPA dans son échantillon à l'aide de facteurs d'allocation. Dans le passé, ces deux enquêtes ont accordé le même poids à leurs groupes de renouvellement. En 1998, l'Enquête sur la population active a décidé d'accorder moins d'importance au groupe de renouvellement moins âgé de façon à améliorer les estimations de tendance (Singh, Kennedy, Wu et Brisebois, 1997).

Puisque plusieurs combinaisons de facteurs d'allocation permettent de produire des estimations appropriées, il faut choisir une stratégie pour le calcul des *fap* de l'EDTR. Ainsi, on pourrait accorder la même importance aux panels que pour l'Enquête sur les finances des consommateurs. Cette stratégie simple n'est pas recommandée pour deux raisons. Tout d'abord, les enquêtes longitudinales peuvent souffrir du biais d'érosion ce qui peut avoir un effet sur la qualité des estimations. Deuxièmement, puisque les échantillons de l'EDTR sont tirés des échantillons de l'EPA à différentes périodes, la fiabilité des panels risque d'être différente, surtout si ceux-ci proviennent de différents plans d'échantillonnage de l'EPA, le remaniement de celle-ci ayant lieu tous les 10 ans.

Il a été décidé de calculer les *fap* suivant l'hypothèse du pire scénario dans lequel les panels ont des impacts différents sur la qualité. Dans le cadre de cette hypothèse, le *fap* doit être tel que l'erreur quadratique moyenne de l'estimation de niveau ou de tendance annuelle est minimisée pour le plus grand nombre possible de variables d'intérêt.

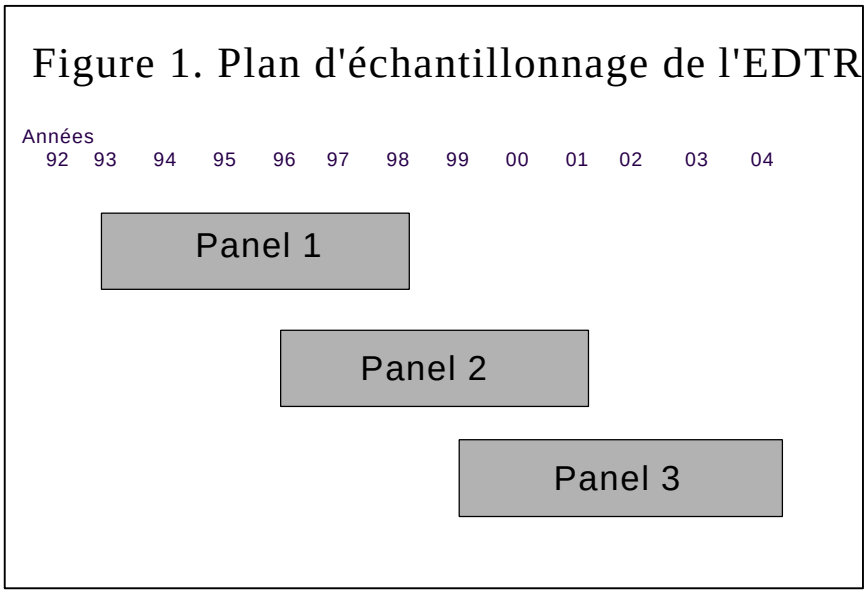
En principe, on pourrait calculer un ensemble de *fap* pour chaque variable d'intérêt. Concrètement, pour des raisons opérationnelles et par souci de simplicité, on ne calcule qu'un ensemble de *fap* dans l'espoir d'obtenir une erreur quadratique moyenne raisonnable pour toutes les variables. De cette façon, un seul ensemble de poids transversaux suffit pour toutes les variables.

L'erreur quadratique moyenne comprend la variabilité et le biais des estimations. Puisqu'il est difficile d'obtenir de bonnes estimations du biais, il a été décidé de considérer uniquement la variabilité. Nous supposons que les ajustements pour la non-réponse permettent d'éliminer le biais. Toutefois, même s'il reste à le prouver, il est possible que le panel plus âgé entraîne une part de biais. Même si le biais n'est pas considéré, par prudence nous supposons que le panel plus âgé est davantage exposé au biais que le moins âgé. La façon dont les *fap* sont calculés tient compte de cette possibilité.

La section qui suit fournit des renseignements sur le plan d'échantillonnage de l'EDTR. La section 3 traite des méthodes envisagées pour le calcul des *fap*. La section 4 aborde les calculs, y compris le choix de la variable à optimiser de même que les définitions opérationnelles. La section 5 présente le calcul pour l'année de référence 1996 et les vagues à venir. Les sections 6 et 7 traitent des répercussions sur les estimations et sur leur fiabilité. Enfin, la section 8 formule des recommandations pour les futures procédures liées aux échantillons de l'EDTR.

2. Le plan d'échantillonnage de l'Enquête sur la dynamique du travail et du revenu

L'EDTR est une enquête annuelle comportant deux panels (Lavigne et Michaud, 1998). Le premier panel est constitué de personnes habitant l'une ou l'autre des dix provinces en janvier 1993 (à l'exclusion des personnes habitant des bâtiments militaires ou demeurant dans un établissement). Un échantillon longitudinal est tiré de cette population. Ces unités longitudinales demeurent dans l'échantillon pendant six ans. Le deuxième panel est tiré de la population de 1996 et il dure également six ans. Les panels se chevauchent tel qu'indiqué à la figure 1. Par la suite, un nouveau panel sera choisi à tous les trois ans pour remplacer le plus ancien des deux panels. Un tel plan signifie qu'il est possible de produire des estimations transversales annuelles à l'aide de deux panels, exception faite pour les trois premières années de l'enquête (1993-1995).



À des fins longitudinales, chaque panel doit représenter la population dont il est tiré, tandis que pour les estimations transversales annuelles, la population cible change chaque année. L'introduction d'un nouveau panel aide à tenir compte des naissances aussi bien que de l'arrivée d'immigrants, facteur important d'un point de vue transversal.

Chaque panel de l'EDTR est un sous-échantillon de 15 000 ménages (40 000 personnes environ) tiré de l'EPA. L'échantillon de l'EPA est tiré d'une base aréolaire en conformité avec un plan d'échantillonnage aléatoire à plusieurs degrés (Singh et coll., 1990). L'EPA se fonde sur six panels, (groupe de renouvellement) avec renouvellement d'un panel tous les mois. L'unité d'échantillonnage du dernier degré est le ménage. Les ménages sélectionnés pour l'EDTR sont ceux qui terminent leur participation à l'EPA au début de la période de référence. Chaque panel est constitué de deux groupes de renouvellement de l'EPA.

Les individus compris dans l'échantillon au début d'un panel sont des *répondants longitudinaux*, par opposition aux *cohabitants* qui se joignent au ménage d'un répondant longitudinal plus tard au cours de la durée du panel (Lavallée, 1995). Les cohabitants font partie de la pondération transversale puisque nous nous intéressons aux caractéristiques des ménages. De plus, les cohabitants aident à améliorer la représentativité de l'échantillon transversal.

2.1 Étapes du processus de pondération

Tandis que la pondération longitudinale se fait séparément pour chaque panel, la pondération transversale, elle, se fonde sur la combinaison de deux panels (Lévesque et Franklin, 2000). Le processus de pondération comporte six étapes :

- Poids de base
- Ajustement pour la non-réponse à l'EDTR
- Combinaison des deux panels
- Partage des poids
- Ajustements analytiques
- Poststratification

1- Poids de base

Pour chaque panel, le poids de base est calculé en fonction des probabilités de sélection de l'EPA. Il comprend également l'ajustement pour la non-réponse à l'EPA. Le poids est alors corrigé en fonction du fait que l'EDTR utilise seulement deux des six groupes de renouvellement de l'EPA. Les unités du panel 1 subissent également une étape spéciale d'ajustement pour la non-réponse à l'interview préliminaire. Cet ajustement spécial s'applique aux strates de sélection en conformité avec le plan de l'EPA.

2- Ajustement pour la non-réponse à l'EDTR

Seuls les répondants de l'EPA sont compris dans l'échantillon de l'ETDR. Puisqu'il existe des données de l'EPA pour toutes les unités échantillonnées, il est possible de modéliser la non-réponse à l'aide des données de l'EPA. L'ajustement pour la non-réponse se fonde donc sur des groupes de réponse homogènes. Les variables utilisées pour constituer ces groupes de réponse sont déterminées à l'aide d'une modélisation de segmentation ou de régression logistique (Dufour, Gagnon, Morin, Renaud, et Särndal, 1998). Ces catégories d'ajustement permettent de neutraliser le biais éventuel causé par le fait que le processus de non-réponse varie selon les caractéristiques des individus. La modélisation et l'ajustement pour la non-réponse se font séparément pour chaque panel.

3- Combinaison des panels

Les deux panels sont regroupés de façon à former un grand échantillon, grâce aux facteurs d'ajustement des panels. Cette étape, qui constitue le thème principal du présent document, est expliquée à la section 3.

4- Partage des poids

Cette étape, qui est particulière à la pondération transversale, est nécessaire à cause de la présence de cohabitants. En effet, des cohabitants sont compris dans l'échantillon, simplement parce qu'ils se sont joints à des ménages qui englobent au moins un individu compris dans l'échantillon longitudinal. Puisque les cohabitants ne sont pas présélectionnés dans le cadre d'un plan d'échantillonnage à probabilité connue, il faut avoir recours à la méthode du partage des poids afin d'obtenir des estimations sans biais (Lavallée, 1995).

5- Ajustements analytiques

Ces ajustements de poids se font soit parce que les poids sont jugés extrêmement élevés comparativement à d'autres poids pour la même province, soit parce qu'un élément individuel pondéré a une valeur élevée dans les estimations globales du revenu. Dans le premier cas, on parle de poids extrêmes; dans le deuxième, on parle de valeurs aberrantes (Lévesque et Franklin, 2000). Les poids extrêmes sont causés par la mobilité interprovinciale; au fil des ans, certains répondants déménagent vers une autre province. L'ajustement des valeurs aberrantes se fait pour des raisons de confidentialité et aussi pour assurer la représentativité.

6- Poststratification

Enfin, les poids subissent une poststratification en fonction de comptes démographiques pour les variables croisées (province, âge, sexe) pour l'année de référence.

3. Méthodes envisagées et données utilisées

Le *fap* permet d'accorder une importance égale ou inégale aux panels afin de répondre à certaines exigences. On compte examiner et ajuster les *fap* à chaque vague, mais des considérations analytiques favorisent un recours à des *fap* stables dans la mesure du possible.

Le *fap* fait partie du calcul de la variance. Il peut également influencer le biais, mais cet aspect n'est pas traité en détail ici. Les données transversales de l'EDTR fournissent des estimations de niveau et de tendance annuelle, les deux étant influencées par le choix du *fap*. Pour la période de référence 1996, il a été décidé de calculer le *fap* en fonction des estimations de niveau, pour les raisons ci-dessous :

- Stabilité accrue dans le temps. Pour ce qui est des tendances, en 1996 seul le premier panel chevauche avec 1995, mais en 1997 les deux panels chevaucheront avec 1996. L'optimisation des estimations de tendance pourrait entraîner des différences appréciables de *fap* entre 1996 et 1997 (voir l'annexe A). Par exemple, le panel 1 recevrait beaucoup plus d'importance en 1996 et beaucoup moins en 1997, le deuxième panel participant alors lui aussi à la réduction de la variance des estimations de tendance.

- Meilleures estimations transversales : le fait que l'échantillon de l'EDTR presque tout entier chevauchera deux années consécutives deux fois sur trois et que, la troisième fois, près de 50 % de l'échantillon se chevauche entraîne une réduction du besoin d'optimisation des estimations de tendance annuelle. De plus, l'EDTR n'a pas été conçue à l'origine pour optimiser les estimations transversales.

L'estimateur combiné transversal est décrit comme suit : soit $\hat{Y}$ l'estimation transversale pour une variable d'intérêt donnée; soit $\hat{Y}_1$ et $\hat{Y}_2$ les estimations de Horwitz-Thompson produites à l'aide du panel 1 et du panel 2 respectivement. L'estimation composite ou combinée est alors donnée par :

$$\hat{Y} = p_1\hat{Y}_1 + p_2\hat{Y}_2 + \hat{Y}_2' \qquad (1)$$

où p_1 et p_2 sont les *fap* du panel 1 et du panel 2. $\hat{Y}_1$ et $\hat{Y}_2$ sont obtenus à l'aide des répondants admissibles à faire partie des deux panels, tandis que $\hat{Y}_2'$ s'obtient à l'aide des répondants qui se joignent à la population cible après le tirage du panel 1 (naissances, immigrants, etc.). Cela correspond à une faible population qui représente annuellement 0,3 % seulement de la population totale. Le plus souvent, il n'est pas possible de déterminer exactement à quel moment les répondants se sont joints à la population cible; les données recueillies ne comportent pas ce genre d'information. Néanmoins, très peu de répondants arrivent après le tirage du panel 1. Par conséquent, on suppose que tous les répondants sont éligibles à la sélection des deux panels, et on utilise l'estimation :

$$\hat{Y} = p_1\hat{Y}_1 + p_2\hat{Y}_2 \quad . \qquad (2)$$

À noter que $\ddot{Y}$ est sans biais uniquement si $p_2 = 1 - p_1$ comme ce sera le cas. À noter également qu'à l'étape de la production, le *fap* est appliqué au niveau des microdonnées. Autrement dit, chaque poids des répondants longitudinaux est multiplié par son *fap* associé. Cela se fait après l'ajustement pour la non-réponse du panel, mais avant le partage des poids et le calage aux marges, de sorte que le *fap* est compris dans le poids d'un cohabitant.

On peut montrer que la variance est minimisée lorsque

$$p_j = \frac{n_j / deff_j}{\sum_{j=1}^{2} n_j / deff_j} \qquad (3)$$

où n_j est le nombre de répondants longitudinaux (poids longitudinal non nul) dans le panel *j* et $deff_j$ représente les effets de plan de sondage du panel *j*. Pour le moment, il n'existe que deux panels, mais la formule serait la même si l'on ajoutait un autre panel ou un échantillon transversal (*j*=3).

3.1 Données utilisées

Lorsque la présente étude a été entreprise, les données de l'EDTR étaient disponibles uniquement pour le premier panel. Par contre, les données de l'Enquête sur les finances des consommateurs (EFC) étaient disponibles pour plusieurs années. Le choix des données de l'EFC se justifiait par le fait que le plan d'échantillonnage de celle-ci est semblable à celui de l'EDTR (il s'agit dans les deux cas de sous-échantillons de l'EPA). De plus, le contenu sur le revenu est le même dans les deux cas, de sorte que l'on peut étudier plusieurs variables d'intérêt de l'EDTR. Pour les variables démographiques, des données de l'EPA ont été utilisées, puisqu'elles permettent d'avoir recours à un échantillon plus grand. L'utilisation des données de l'EPA est décrite plus en détail à la section 4.2.

Des données de l'EFC des années de référence 1993 et 1996 ont été utilisées. Ces années correspondent aux années de sélection des panels 1 et 2. À noter que le remaniement de l'EPA a eu lieu en 1994. Cela signifie que les deux échantillons relèvent d'un plan différent et correspondent donc exactement à la situation de l'EDTR. Enfin, l'utilisation de ces deux années correspond à l'écart temporel entre les deux panels de l'EDTR.

Il a fallu laisser tomber un des quatre groupes de renouvellement de l'EFC puisqu'il était très différent des autres groupes de renouvellement dans deux provinces (le groupe de renouvellement 6 pour l'année 1993 et le groupe de renouvellement 2 pour les données de 1996).

4. Définitions et considérations opérationnelles

Pour établir une définition à la fois utile et pratique du *fap*, il faut considérer plusieurs aspects. Ceux-ci se divisent en deux catégories :

- variable(s) à considérer dans le processus d'optimisation
- type de données et période de référence à utiliser.

Ces catégories sont abordées dans les sous-sections qui suivent.

4.1 Variable(s) à considérer

L'EDTR se penche sur quatre catégories de données : démographie, travail, revenu et mesure des faibles revenus. Chaque variable de ces groupes comporte ses propres effets de plan de sondage. Le tableau 1 indique la variabilité des ratios des effets de plan pour quatre variables selon les estimations de l'EFC.

Tableau 1
Ratio des effets de plan de l'EFC de 1994 à 1997
selon la variable et la province[1]

PROVINCE	NOMBRE DE PERSONNES ÂGÉES DE 15 ANS+	REVENU TOTAL	NOMBRE DE MÉNAGES DE TAILLE 2+	NOMBRE DE PERSONNES EN DEÇÀ DU SFR[2]
Terre-Neuve	1,60	2,33	2,66	1,11
Île-du- Prince-Édouard	2,19	2,51	3,84	1,05
Nouvelle-Écosse	4,45	2,75	4,58	1,53
Nouveau-Brunswick	2,97	2,58	3,37	0,76
Québec	2,92	4,70	2,90	1,37
Ontario	1,71	2,18	1,77	1,55
Manitoba	4,83	2,70	4,18	1,47
Saskatchewan	3,02	2,67	2,96	1,34
Alberta	2,95	2,30	2,44	1,74
Colombie-Britannique	2,71	2,44	2,73	1,78
Canada	2,24	2,48	2,29	1,52

[1] Le groupe de renouvellement 2 est exclu du calcul pour toutes les variables sauf le SFR puisqu'il donne des résultats très différents. Les ratios des effets de plan pour le SFR comprennent le groupe de renouvellement 2, mais ce n'est pas pour cette raison que les ratios sont plus petits que ceux des autres variables.

[2] Seuil de faible revenu

Le plus souvent, les ratios sont supérieurs à un. Le *fap* du panel 2 est donc forcément supérieur à celui du panel 1. Une optimisation pour toutes les variables entraînerait le calcul de nombreux *fap* et ensembles de poids. Cette possibilité est exclue pour des raisons analytiques et opérationnelles, vu le grand nombre de variables clés des produits de l'EDTR. Il faut donc choisir une variable unique, en calculer le *fap* et espérer que les autres variables ne subiront pas trop lourdement l'effet de ce choix. Idéalement, le *fap* est accessible à l'avance afin que les calculs du *fap* ne ralentissent pas le processus de pondération. À l'avenir, lorsque les données de deux panels seront disponibles, le *fap* pourra être calculé à l'aide de données de la vague précédente de l'EDTR. Entre-temps, il faut faire appel à des sources externes. À l'heure actuelle, les deux principales sources externes d'information pour l'EDTR sont l'EPA et l'EFC.

La variable recommandée pour le calcul du *fap* est le *nombre de personnes âgées de 15 ans ou plus*, ce qui correspond à la population cible de l'EPA. Ce choix devrait donner lieu à une estimation plus stable des *DEFF* puisqu'il correspond à un grand domaine et puisqu'il s'agit d'une variable catégorique. D'autres facteurs expliquent ce choix. Tout d'abord, il existe une corrélation à peu près égale entre cette variable et toutes les autres. Deuxièmement, cette variable est préparée directement à l'aide de l'EPA, de sorte qu'elle sera toujours accessible à l'avenir, contrairement aux données de l'EFC qui sera intégrée à l'EDTR pour l'année de référence 1998. Troisièmement, cette variable est définie pour tous les répondants de l'EDTR. Enfin, l'EPA peut fournir des estimations plus fiables des effets de plan de sondage puisqu'elle comprend jusqu'à six groupes de renouvellement dont on peut se servir, au lieu de deux pour l'EDTR et de quatre pour l'EFC. On suppose bien sûr que les véritables effets de plan ne dépendent pas de la taille de l'échantillon, ce qui est vrai lorsque l'échantillon augmente selon le nombre de groupes de renouvellement.

De plus, le tableau 1 indique que les ratios des effets de plan de sondage varient d'une province à l'autre. Puisque les ratios de la taille des échantillons varient également de façon appréciable d'une province à l'autre, il est recommandé de calculer le *fap* au niveau de la province.

4.2 Type de données et période de référence à utiliser

Plusieurs définitions opérationnelles de la taille de l'échantillon et des effets de plan de sondage peuvent servir au calcul du *fap*. Si l'on examine la taille des échantillons utilisés dans l'équation (3), seuls les répondants longitudinaux au moment de l'estimation sont considérés malgré la présence de cohabitants dans les estimations transversales. Lavallée

(1994) a montré que, dans le processus de pondération, il faut combiner les deux panels avant l'intégration des cohabitants si l'on utilise la méthode du partage des poids. Par conséquent, nous n'utilisons ici que les répondants longitudinaux. De plus, l'élimination des cohabitants du calcul du *fap* aide à diminuer l'importance du panel plus âgé et donc à minimiser l'effet du biais éventuel dû à l'érosion.

Puisque les unités d'échantillonnage de l'EDTR sont les ménages, les effets de plan pour la variable d'intérêt sont calculés au niveau des ménages. De cette façon, les effets de plan sont calculés suivant l'hypothèse d'un échantillonnage aléatoire simple de ménages.

Deux types d'effets de plan de sondage sont considérés : un qui suppose un échantillon aléatoire simple par strate et un autre par province. On utilise ce dernier type parce qu'il reflète les effets de la stratification aussi bien que de l'effet de grappe. À noter que les effets de plan sont calculés à l'aide des poids après l'ajustement pour la non-réponse (sous-poids), mais avant le calage aux marges.

Afin d'obtenir des estimations comportant une variance minime, du moins pour la variable *nombre de personnes âgées de 15 ans ou plus*, il convient de calculer le *fap* à l'aide de données de l'EDTR pour les deux panels au moment de l'estimation. Cela signifie qu'il faut d'abord calculer les poids individuels de chaque panel, suivant la description de Latouche, Michaud et Renaud (1997) ou de Dufour et coll. (1998), puis calculer le *fap* à l'aide de l'équation (3), pour enfin déterminer les poids individuels. Le résultat d'une telle stratégie est que le *fap* dépend de l'échantillon. D'un point de vue opérationnel, il est plus facile de créer un *fap* indépendant de l'échantillon en ayant recours à des sources externes (voir la description à la section 4.1).

Puisque l'on a recours à l'EPA pour calculer les effets de plan de sondage, on peut se demander quelle période de référence utiliser. Les méthodologistes techniques estiment que la période de référence devrait être celle durant laquelle le panel a été choisi. Ainsi, les effets de plan de sondage du panel 1 sont calculés à l'aide de données de l'EPA de janvier 1993, les données de janvier 1996 étant utilisées pour le panel 2. De cette façon, l'ancien plan de sondage de l'EPA est utilisé pour le panel 1 et celui de nouveau plan pour le panel 2. Par contre, les méthodologistes appliqués suggèrent que l'on utilise l'effet de plan de sondage qui correspond à la période de référence pour laquelle on prépare les estimations. Cela permet non seulement de considérer l'écart temporel entre les deux panels, mais également de tenir compte de la détérioration du plan de sondage au fil des ans. Ainsi, les effets de plan de sondage du panel 2 sont estimés en fonction de données sur l'EPA de janvier 1996, comme pour la stratégie précédente. Les effets de plan de

sondage du panel 1 sont également estimés à l'aide de données de janvier 1996. Toutefois, l'ancien plan de sondage de l'EPA ne servait plus en 1997. La période la plus proche que l'on peut utiliser est septembre 1994, qui correspond à la dernière utilisation de l'ancien plan de sondage. Pour l'année de référence 1996, il a été décidé d'avoir recours à la stratégie technique (1993 pour le panel 1 et 1996 pour le panel 2).

5. *Fap et fréquence de calcul*

5.1 Année de référence 1996

Le tableau 2 présente le plan de calcul du *fap* pour la période de référence 1996. Ce plan donne lieu au *fap* présenté au tableau 3.

Tableau 2.
Plan de calcul du *fap* de 1996

Éléments	Définition opérationnelle
Variable d'intérêt	Nombre de personnes âgées de 15 ans ou plus
Source externe	Enquête sur la population active
Période de référence	Janvier 1993 (panel 1), janvier 1996 (panel 2)
Niveau de calcul	Province
Taille de l'échantillon	Répondants longitudinaux au moment de l'estimation
Type d'effets de plan de sondage	Hypothèse d'un échantillon aléatoire de ménages
Niveau des effets de plan	Strate et grappe, à l'aide de sous-poids

Tableau 3.
Fap utilisé pour la période de référence 1996
(calculé à l'aide de données de l'EPA de 1993 et de 1996)

Province	Taille panel 1	Taille panel 2	$deff_1 / deff_2$	*Fap panel 1*	*Fap* panel 2
Terre-Neuve	1698	1315	1,71	0,4302	0,5698
Île-du-Prince-Édoaurd	575	895	1,92	0,2507	0,7493
Nouvelle-Écosse	1853	2044	4,02	0,1840	0,8160
Nouveau-Brunswick	1768	1882	1,94	0,3263	0,6737
Québec	4928	5853	2,75	0,2344	0,7656
Ontario	7054	9174	2,87	0,2113	0,7887
Manitoba	1821	2113	1,36	0,3879	0,6121
Saskatchewan	1945	1863	0,92	0,3208	0,6792
Alberta	2406	2164	2,21	0,2792	0,7208
Colombie-Britannique	2264	2570	2,87	0,2420	0,7580

Si les panels comportaient tous deux la même taille et les mêmes effets de plan de sondage pour toutes les provinces, le *fap* serait de 0,5 pour les deux panels. Cette valeur serait semblable au facteur d'allocation des groupes de renouvellement pour l'ancien plan de l'EPA, c'est-à-dire 1/6

pour chaque groupe. Pour l'EDTR, on peut constater que les *fap* sont très différents de 0,5. Cet écart est causé en grande partie par les ratios des deux effets de plan de sondage. Le tableau 4 présente les *fap* que l 'on obtiendrait si les deux panels avaient les mêmes effets de plan de sondage.

Tableau 4.
Fap de la période de référence 1996
(hypothèse d'effets de plan de sondage identiques)

Province	**Taille panel 1**	**Taille panel 2**	$deff_1 / deff_2$	***Fap* panel 1**	***Fap* panel 2**
Terre-Neuve	1698	1315	1	0,5636	0,4364
Île-du-Prince-Édouard	575	895	1	0,3912	0,6088
Nouvelle-Écosse	1853	2044	1	0,4755	0,5245
Nouveau-Brunswick	1768	1882	1	0,4844	0,5156
Québec	4928	5853	1	0,4571	0,5429
Ontario	7054	9174	1	0,4347	0,5653
Manitoba	1821	2113	1	0,4629	0,5371
Saskatchewan	1945	1863	1	0,5108	0,4892
Alberta	2406	2164	1	0,5265	0,4735
Colombie-Britannique	2264	2570	1	0,4683	0,5317

5.2 Années ultérieures

Il est évident que la taille de l'échantillon longitudinal des deux panels change à chaque vague. Les ratios des effets de plan de sondage suivant la stratégie technique sont supposés identiques jusqu'à l'introduction d'un nouveau panel. Puisque les ratios des effets de plan de sondage et les ratios de la taille des échantillons sont stables au fil des ans lorsqu'ils sont calculés pour les deux mêmes panels, il est recommandé de ne pas recalculer le *fap* aux fins de la production. De cette façon, on ne causera pas d'écart pour les tendances annuelles. Il ne faut recalculer le *fap* que lorsqu'on introduit un nouveau panel. Néanmoins, la valeur d'un *fap* devrait être surveillée annuellement dans le cadre d’un programme d’assurance de la qualité.

Lorsqu'on introduit un nouveau panel relevant du même plan de l’EPA que le panel qui reste, la stratégie technique entraîne un *fap* qui favorise le panel plus âgé. Cette situation survient en 1999 pour la sélection du panel 3. On peut supposer que le nouveau panel comporte des effets de plan de sondage plus importants causés par une détérioration de l'homogénéité de la strate. Même si les ratios des effets de plan devraient être plus proches de l'unité que les ratios calculés à l'aide d'un plan de sondage différent, les ratios seraient inférieurs à l'unité, d'où la possibilité d'un biais dû à l'érosion plus grand. Dans le cas de la stratégie pratique, le

fap favorise le panel moins âgé et, par conséquent, c'est lui qui est recommandé. Lorsque les deux panels relèvent du même plan de sondage de l'EPA, la stratégie pratique est facile à utiliser et l'on peut éviter les retards de production en ayant recours aux effets de plan des vagues précédentes.

6. Effet sur les estimations

Tableau 5
Estimations nationales obtenues à l'aide d'un *fap* optimal et égal

Variable	*Fap* optimal	*Fap* égal	Différence relative[1]
Nombre de personnes seules	3 984 199	4 033 703	1,24
Nombre de familles de taille 2	3 407 517	3 445 625	1,12
Nombre de familles de taille 3+	4 777 024	4 742 161	-0,73
Nombre de personnes mariées	12 471 961	12 435 254	-0,29
Nombre de personnes célibataires	6 411 378	6 378 680	-0,51
Nombre de personnes séparées	648 956	727 732	**12,14**
Nombre de personnes dont l'état matrimonial est inconnu	192 204	128 209	**-33,30**
Nombre de familles habitant une région rurale	1 311 865	1 315 916	0,31
Nombre de familles habitant une région 100 000 - 499 999	1 985 941	2 043 483	**8,14**
Nombre de familles habitant une région 500 000+	5 933 449	5 893 937	**-3,57**
Total des gains (X 10^6)	421 029	425 795	1,13
Total des placements (X 10^6)	23 863	24 617	3,16
Total des transferts gouvernementaux (X 10^6)	76 467	76 196	-0,35
Total : autre revenu monétaire (X 10^6)	44 103	45 033	2,11
Total du revenu (X 10^6)	563 583	569 553	**1,06**
Revenu moyen : famille	56 955	57 290	0,59
Revenu moyen : personnes seules	24 371	24 821	**1,84**
Revenu moyen : personnes 16+	25 347	25 605	**1,02**
Pourcentage de personnes en deçà du SFR	18,60	18,01	**-3,17**

[1] Les **chiffres en caractères gras** signifient que la différence est statistiquement significative au niveau de 1 %.

Il est intéressant de comparer les estimations obtenues à l'aide d'un ensemble de *fap* optimaux et celles obtenues lorsqu'on accorde la même importance aux deux panels. En théorie, les deux ensembles produisent des estimations sans biais. De plus, si les estimations obtenues à l'aide des deux panels séparément étaient les mêmes, les estimations combinées ne dépendraient pas du *fap*. Néanmoins, puisque les panels subissent au moins une variabilité due à l'échantillonnage, les estimations pour des panels distincts sont légèrement différentes. Si la différence se trouvait à l'extérieur de l'intervalle de confiance, ce serait une indication que les panels seraient quelque peu différents et que le *fap* pourrait nettement influencer les estimations combinées. Dans un tel cas, les différences entre panels pourraient être causées par des changements de

traitement, d'erreur de réponse (biais d'accoutumance) ou de couverture de l'échantillon, y compris le biais dû à l'érosion.

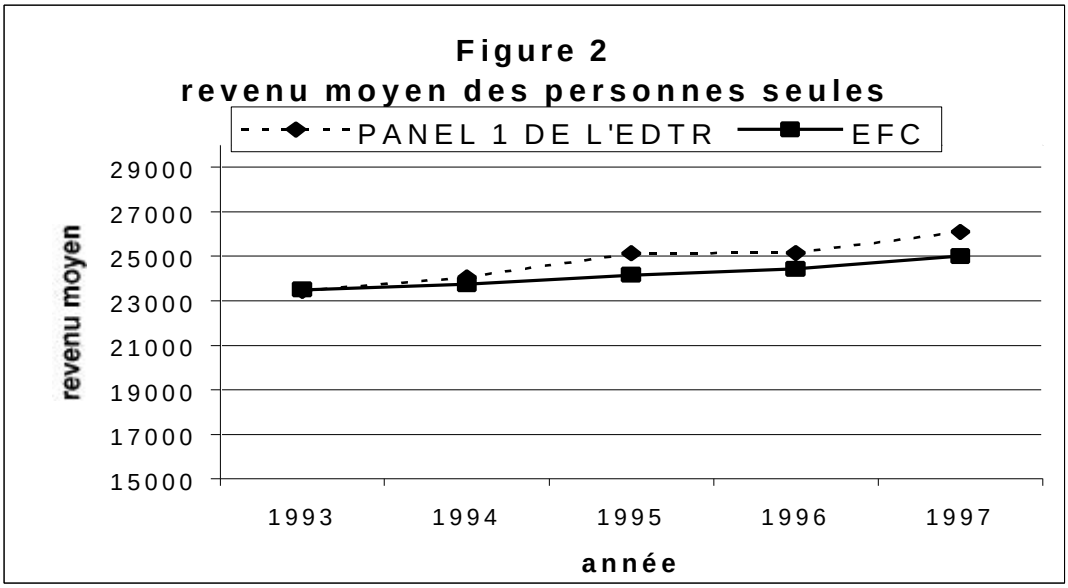
Le tableau 5 contient les estimations nationales de 1996 pour des variables clés préparées à l'aide d'un *fap* optimal (selon le tableau 3) et lorsque le *fap* est fixé à 0,5 pour toutes les provinces. Le calage utilisé pour les deux ensembles de *fap* est la simple poststratification par province, âge, sexe utilisée pour la production. Les différences relatives statistiquement significatives sont en caractères gras. Les données de 1996 suggèrent des différences dans quatre domaines : répartition de l'état matrimonial et de la taille des villes, mesure des revenus moyens et des revenus faibles. L'annexe C décrit la stratégie servant à l'approximation des variances de ces différences.

L'estimation de type *fap* optimal indique que le nombre de personnes séparées estimées pour le panel 2 est beaucoup moindre. Toutefois, elle indique également qu'il existe beaucoup plus de personnes dont l'état matrimonial n'est pas connu. Avec le temps, ces différences devraient disparaître à mesure que le fichier est épuré.

L'écart de répartition de la taille des villes est observé pour les groupes 100 000-499 999 et 500 000+; l'estimation de type *fap* égal entraîne une augmentation pour les 100 000-499 999 et une réduction pour les 500 000+. Puisque la valeur du *fap* optimal pour le panel 2 est supérieure à 0,5, c'est là une indication qu'il existe moins de personnes dans le groupe 100 000-499 999 dans le panel 2 et plus de personnes dans le groupe 500 000+. Cette différence est attribuable au remaniement de l'EPA qui a eu lieu en 1994. Essentiellement, le nouveau plan de l'EPA donne un échantillon plus grand dans les régions urbaines que dans les régions rurales comparativement à l'ancien plan. L'EFC a également été touchée par le remaniement de l'EPA.

Pour ce qui est de l'agrégat du revenu total, les estimations du revenu total moyen des personnes seules et 16+ indiquent que le premier panel fournit des estimations de revenu plus élevées que le deuxième panel, surtout pour les personnes seules. Ce phénomène combiné à la taille des villes permet d'expliquer pourquoi il existe une différence entre les estimations de la mesure des revenus faibles relevant de la stratégie des fap optimal et égal. C'est aux spécialistes qu'il revient de déterminer si ces différences sont importantes. Il se peut que les différences relèvent d'un problème de couverture; avec le temps, un panel tend à comporter moins de personnes seules et plus pauvres. Toutefois, si l'on considère le panel 1 seulement, le revenu total moyen de la série des personnes seules ne diffère pas de celui obtenu à l'aide des données de l'EFC, comme l'indique la figure 2. À noter qu'aucune des estimations annuelles préparées à l'aide du premier panel de l'EDTR et de l'EFC n'est

significativement différente au niveau de 1 %. Il va falloir poursuivre les recherches à l'aide d'estimations plus précises de la variance si l'on veut expliquer les différences entre le premier panel et le deuxième pour ce qui est des données sur le revenu.



On trouvera à l'annexe B un graphique de répartition des revenus. Les deux répartitions sont très semblables.

7. Effet sur la variance

Il est utile de connaître la perte de précision attribuable à l'exigence que le *fap* soit toujours égal à 0,5. On trouvera ci-dessous une analyse simple et utile de la sensibilité, fondée sur la formulation de l'annexe A.

Puisque la variance minimale et le *fap* optimal pour le premier panel sont :

$$V_{\min}(\hat{Y}) = \frac{V(\hat{Y}_1)V(\hat{Y}_2)}{V(\hat{Y}_1)+V(\hat{Y}_2)}$$

$$p_{opt,1} = \frac{V(\hat{Y}_2)}{V(\hat{Y}_1)+V(\hat{Y}_2)}$$

et puisque la variance de $\hat{Y}$ pour un *fap* fixé à 0,5 est

$$V_{0.5}(\hat{Y}) = \frac{V(\hat{Y}_2)}{4p_{opt}}$$

la perte de précision attribuable à l'utilisation d'un *fap* de 0,5 est

$$\frac{V_{0.5}(\hat{Y})}{V_{opt}(\hat{Y})} = \frac{1}{4\,p_{opt}\,(1-p_{opt})}.$$

Le tableau 6 indique, pour chaque province, la perte de précision attribuable à l'exigence que le *fap* provincial de 1996 soit égal à 0,5. Considérons l'Ontario, par exemple. Le *fap* optimal est de 0,2113. Le fait de fixer le *fap* à 0,5 entraîne une inflation de la variance de 19 % et du c.v. de 9 %.

Tableau 6. Perte de précision attribuable à l'exigence que le *fap* soit égal à 0,5

Province	FAP optimal	Perte de variance (%)	Perte de c.v. (%)
Terre-Neuve	0,4302	426,32	129,42
Île-du-Prince-Édouard	0,2507	177,78	66,67
Nouvelle-Écosse	0,1840	96,08	40,03
New Brunswick	0,3263	56,25	25,00
Québec	0,2344	33,33	15,47
Ontario	0,2113	19,05	9,11
Manitoba	0,3879	9,89	4,83
Alberta	0,3208	4,17	2,06
Colombie-Britannique	0,2792	1,01	0,50

On peut constater que l'utilisation d'un *fap* optimal entraîne un accroissement intéressant de la précision.

8. Recommandations

Il est très difficile de déterminer les meilleures pratiques pour le calcul du *fap* dans le cas de l'EDTR. Il faut considérer un nombre de variables si élevé qu'il est impossible d'établir une stratégie qui soit optimale pour toutes les variables. Compte tenu des travaux réalisés à l'aide des données de 1996, il semble que les estimations de niveau ne changent pas beaucoup lorsqu'elles sont calculées à l'aide d'un *fap* optimal ou égal. Toutefois, certaines estimations provinciales sont touchées davantage surtout pour de petits domaines.

Pour ce qui est de la variabilité, le recours à un *fap* optimal entraîne un certain accroissement de la précision. Pour des raisons de stabilité et de biais, il est préférable de calculer le *fap* de façon à minimiser les estimations de niveau plutôt que les estimations de tendance. Cet accroissement de la précision ne semble peut-être pas très important, mais l'utilisation d'un *fap* optimal entraîne une réduction du biais éventuel dû à l'érosion parce qu'une plus grande importance est accordée au panel moins âgé. Il est donc recommandé de continuer d'utiliser un *fap* optimal.

Il a été mentionné que, pris séparément, les deux panels peuvent donner des estimations différentes. Les estimations combinées pourraient donc varier considérablement pour différentes valeurs de *fap*. Par conséquent, il est suggéré de ne réexaminer le *fap* que lors de l'introduction d'un nouveau panel afin d'éviter des complications analytiques. Pour ce qui est de la cause de ces écarts, il est important de vérifier certaines hypothèses afin d'expliquer ces différences et de continuer de surveiller la couverture des panels au fil des ans.

9. Bibliographie

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 American Statistical Association SRMS, pp. 552-557.

Latouche, M., Michaud, S. et Renaud M. (1997). Concerns Pertaining to Weighting of Longitudinal Surveys. American Statistical Association Proceedings of the Section on Government Statistics and Section on Social Statistics. Pp. 111-119.

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

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, volume 21, n° 1, juin 1995 : 27-35.

Lavigne, M. et Michaud, S. (1998). Aspects généraux de l'enquête sur la dynamique du travail et du revenu. Document de travail de l'EDTR, Statistique Canada, n° 98-05 au catalogue.

Lévesque, I. et Franklin, S. (2000). Pondération longitudinale et transversale de l'Enquête sur la dynamique du travail et du revenu, année de référence: 1997. Statistique Canada, document de travail n° 00004 au catalogue.

Merkouris, T. (1999). Cross-sectional Estimation in Multiple-panel Household Surveys. Statistique Canada, document de travail n° HSMD-99-004E.

Singh, A., Kennedy, B., Wu, S. et Brisebois, F. (1997). Composite Estimation for the Canadian Labour Force Survey. Proceedings of the Survey Research Methods Section, American Statistical Association, 300-305.

Singh, M.P., Drew, J.D., Gambino, J.G. et Mayda, F. (1990). Méthodologie de l'enquête sur la population active du Canada 1984-1990. Statistique Canada, n° 71-526 au catalogue.

ANNEXE A
ESTIMATION DE LA VARIANCE

A1- Optimisation du *fap* pour les estimations de niveau

Si l'on suppose un chevauchement complet des panels, il est possible de décrire l'estimateur combiné transversal comme suit : soit $\hat{Y}$, l'estimation transversale pour une variable d'intérêt donnée, et $\hat{Y}_1$ et $\hat{Y}_2$ les estimations résultant du premier panel et du second panel respectivement. L'estimation composite ou combinée est alors donnée par :

$$\hat{Y} = p_1\hat{Y}_1 + p_2\hat{Y}_2$$

où p_1 et p_2 sont les *fap* des premier et second panels. À noter que $\hat{Y}$ est sans biais uniquement si $p_2 = 1 - p_1$, comme ce sera le cas.

La variance de $\hat{Y}$ est donnée par :

$$V(\hat{Y}) = {p_1}^2 V(\hat{Y}_1) + {p_2}^2 V(\hat{Y}_2)$$

Il est possible de montrer que la variance est minimisée si

$$p_1 = \frac{V(\hat{Y}_2)}{V(\hat{Y}_1) + V(\hat{Y}_2)}$$

ce qui donne la variance minimale

$$V_{\min}(\hat{Y}) = \frac{V(\hat{Y}_1)V(\hat{Y}_2)}{V(\hat{Y}_1) + V(\hat{Y}_2)}$$

Si l'on suppose que la variance de la population (S^2) est identique pour les deux panels et que la correction de la population finie (*1-f*) est minime, on peut exprimer les *fap* comme suit :

$$p_1 = \frac{n_1}{n_1 + n_2 \, deff_1 / deff_2}$$

$$p_2 = 1 - p_1$$

où $deff_1$ et $deff_2$ sont les effets de plan des premier et second panels respectivement, et n_1 et n_2 sont le nombre de répondants longitudinaux (poids longitudinal non nul) des premier et second panels au moment de l'estimation. On voit bien que le calcul optimal du *fap* dépend uniquement de la taille de l'échantillon longitudinal et du ratio des effets de plan du panel. Enfin, si l'échantillon est constitué du panel *K* au lieu du panel 2, le *fap* est :

$$p_j = \frac{n_j / deff_j}{\sum_{j=1}^{K} n_j / deff_j} \qquad (1)$$

Cette formule s'applique toujours lorsqu'un panel est remplacé par un échantillon transversal typique.

A2. Calcul de la variance

Les estimations de la variance de l'EDTR sont obtenues de la méthode du Jackknife. Nous voulions savoir s'il fallait calculer de nouvelles valeurs *fap* pour chaque itération du Jackknife. Puisque la méthode du Jackknife dépend de la taille de l'échantillon, il n'est pas nécessaire de recalculer les *fap*. Par contre, si l'on veut simuler exactement le processus de pondération à chaque itération (en impliquant un changement de n_1 et de n_2), il convient de recalculer les *fap*. Pour découvrir quel serait l'effet sur les valeurs *fap*, nous avons exécuté une simulation afin d'observer la gamme possible des *fap* provinciaux. Le calcul du *fap* après l'élimination de la plus petite ou de la plus grande grappe dans chaque province a permis d'y arriver. Les résultats sont indiqués dans le tableau A1. Concrètement, le changement des valeurs *fap* est tellement faible que l'on peut choisir de ne pas recalculer le *fap* et d'utiliser le *fap* de production pour toutes les itérations du Jackknife. Cette façon de procéder simplifie également l'application du Jackknife.

Tableau A1
Plus petite et plus grande valeurs *fap*
pour la simulation du Jackknife

Province	Valeur *fap* véritable[1]	Panel 1		Panel 2	
		Valeur *fap* min.	Valeur *fap* max.	Valeur *fap* min.	Valeur *fap* max.
Terre-Neuve	0,37	0,35	0,37	0,38	0,40
Île-du-Prince-Édouard	0,32	0,28	0,31	0,33	0,36
Nouvelle-Écosse	0,19	0,18	0,19	0,19	0,21
Nouveau-Brunswick	0,34	0,32	0,33	0,34	0,35
Québec	0,25	0,25	0,25	0,25	0,26
Ontario	0,25	0,25	0,25	0,25	0,26
Manitoba	0,42	0,40	0,42	0,42	0,44
Saskatchewan	0,30	0,29	0,30	0,31	0,31
Alberta	0,26	0,25	0,26	0,26	0,26
Colombie-Britannique	0,25	0,25	0,25	0,26	0,26

[1] Cette simulation a été menée à l'aide de données provisoires de l'EDTR; c'est pourquoi les valeurs *fap* diffèrent quelque peu des valeurs de production indiquées au tableau 4.

A3- Optimisation du *fap* pour les estimations de tendance annuelle

L'estimateur de tendance annuelle se définit comme suit : soit $\hat{Y}_t = p_1\hat{Y}_{t,1} + p_2\hat{Y}_{t,2}$, l'estimation transversale pour une variable d'intérêt donnée pour la période de référence *t*, où $\hat{Y}_{t,1}$ et $\hat{Y}_{t,2}$ sont les estimations du premier et second panels respectivement au moment *t*.

Soit $\hat{Y}_t = p_1\hat{Y}_{t,1} + p_2\hat{Y}_{t,2}$ et $\hat{Y}_{t+1} = p_1\hat{Y}_{t+1,1} + p_2\hat{Y}_{t+1,2}$, les estimations du total aux moments *t* et *t+1* respectivement. Nous supposons ici que les *fap* sont identiques aux temps *t* et *t+1*. La tendance annuelle entre les temps *t* et *t+1* est donnée par :

$$
\begin{aligned}
{}_{t+1}\hat{D}_t &= \hat{Y}_{t+1} - \hat{Y}_t \\
&= p_1(\hat{Y}_{t+1,1} - \hat{Y}_{t,1}) + p_2(\hat{Y}_{t+1,2} - \hat{Y}_{t,2})
\end{aligned}
$$

La variance de ${}_{t+1}\hat{D}_t$ est donnée par :

$$
\begin{aligned}
V({}_{t+1}\hat{D}_t) = {} & p_1{}^2(V(\hat{Y}_{t+1,1}) + V(\hat{Y}_{t,1}) - 2COV(\hat{Y}_{t+1,1}, \hat{Y}_{t,1})) \\
& + p_2{}^2(V(\hat{Y}_{t+1,2}) + V(\hat{Y}_{t,2}) - 2COV(\hat{Y}_{t+1,2}, \hat{Y}_{t,2}))
\end{aligned}
$$

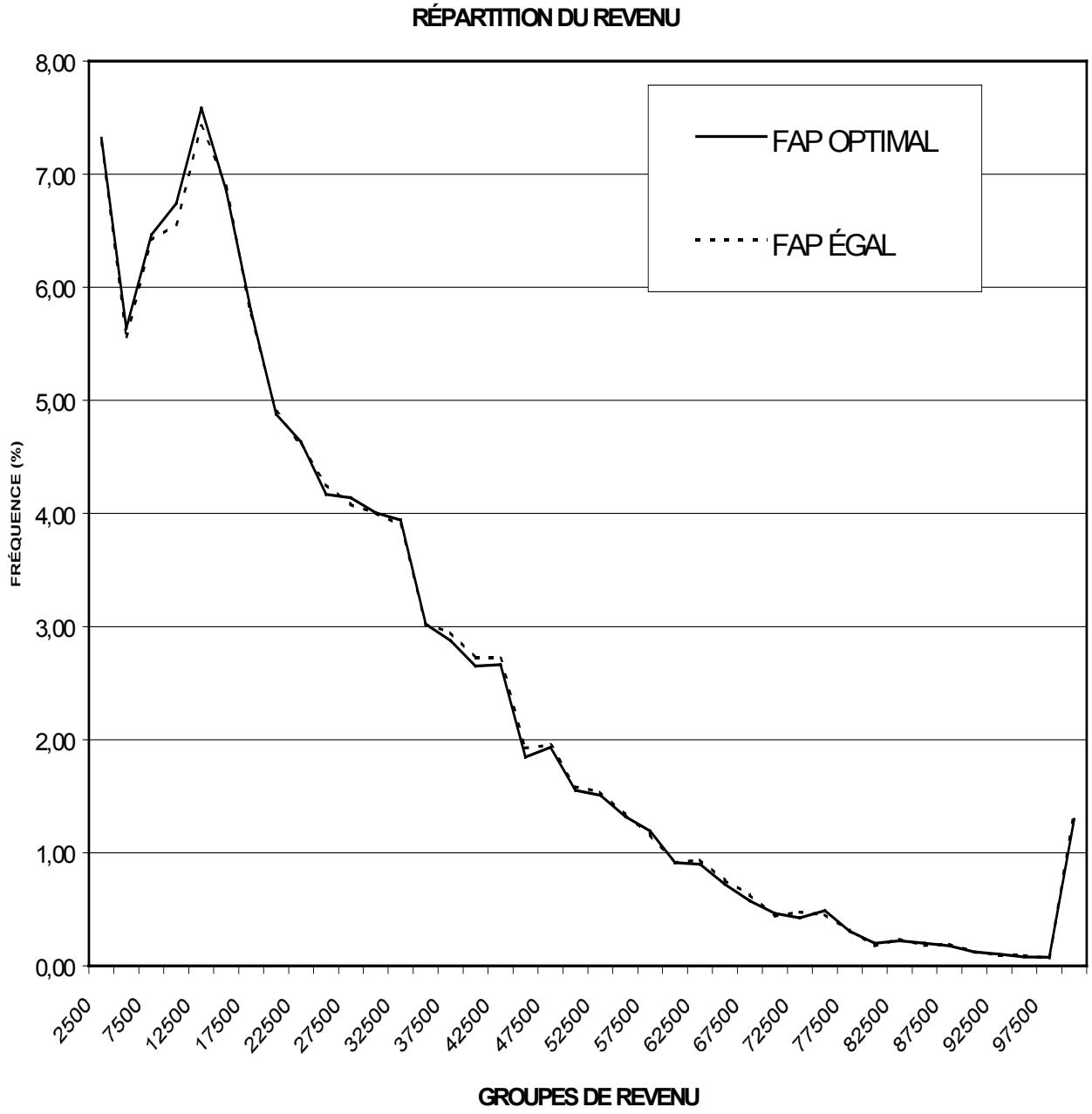
On peut montrer que la variance est minimisée si

$$p_1 = \frac{V(\hat{Y}_{t+1,2}) + V(\hat{Y}_{t,2}) - 2COV(\hat{Y}_{t+1,2}, \hat{Y}_{t,2})}{\sum_{j=1}^{2} V(\hat{Y}_{t+1,j}) + V(\hat{Y}_{t,j}) - 2COV(\hat{Y}_{t+1,j}, \hat{Y}_{t,j})}$$

où $COV(\hat{Y}_{t+1,j}, \hat{Y}_{t,j})$ est la covariance du panel *j* entre les deux années; cette covariance est toujours positive et devrait être grande. À noter que, lorsqu'un panel est utilisé pour la première fois, il ne comporte pas de covariance. Dans une telle situation, seul le panel plus âgé joue un rôle dans la covariance. Par conséquent, le panel plus âgé comporte un *fap* plus grand et l'estimation est davantage exposée au biais dû à l'érosion. Une année plus tard, les deux panels comportent une covariance et le *fap* doit être calculé de nouveau, ce qui risque de provoquer une instabilité des estimations et d'entraîner des problèmes analytiques.

ANNEXE B

ESTIMATIONS NATIONALES ET PROVINCIALES
RÉSULTANT DE L'UTILISATION DE *FAP* OPTIMAL ET ÉGAL[1]



[1] L'hypothèse de l'égalité des estimations préparées à l'aide des deux méthodes n'a pas été vérifiée sauf pour les estimations présentées au tableau 5.

ANNEXE C
TEST D'HYPOTHÈSE SUR LES ESTIMATIONS

Pour déterminer quelles différences entre les estimations de *fap* égal et de production étaient statistiquement significatives, nous avons utilisé la stratégie ci-dessous afin d'obtenir une approximation de la variance de ces différences.

Soit

$$\hat{t}_p = p_1\hat{Y}_1 + p_2\hat{Y}_2$$
$$\hat{t}_q = q_1\hat{Y}_1 + q_2\hat{Y}_2$$

les estimations relevant de deux ensembles de *fap*, p et q respectivement. Soit $\hat{Y}_i$, l'estimation du panel *i.* Ainsi, *p* pourrait représenter l'ensemble de *fap* égaux (0,50) et *q* l'ensemble de *fap* de production. Considérons la différence

$$\begin{aligned} D_{pq} &= t_p - t_q \\ &= (\hat{Y}_1 - \hat{Y}_2)(p_1 - q_1) \end{aligned}$$

Si nous supposons que $V(\hat{Y}_1) \approx V(\hat{Y}_2)$, il est possible de montrer que

$$V(D_{pq}) = 2V(\hat{Y}_2)(p_1 - q_1)^2$$

Pour la production de 1996, en établissant la moyenne pour la province nous avons $p_1 - q_1 \approx 0{,}22$ de sorte que

$$V(D_{pq}) = 0{,}0968V(\hat{Y}_2) = 0{,}0968cv^2(\hat{Y}_2)\hat{Y}_2^2$$

où $cv(\hat{Y}_2)$ est le coefficient de variation de l'estimation du second panel. Enfin, on détermine que la différence est statistiquement significative au niveau de 1% si

$$z = \frac{|D|}{cv(\hat{Y}_2)\hat{Y}_2\sqrt{0{,}0968}} \geq 2{,}57$$

À noter que cette vérification correspond à une vérification de l'hypothèse nulle : $\hat{Y}_1 = \hat{Y}_2$. Le tableau C1 indique, pour certaines variables clés, le c.v., la valeur z et la probabilité, dans le cadre de l'hypothèse nulle, d'observation d'une telle valeur z.

Tableau C1
Test d'hypothèse pour l'égalité des estimations de *fap* optimal et égal

Variable	*Fap* optimal	*Fap* égal	Différence relative	Coefficient de variation (%)	z	Probabilité de z plus grand	Source du c.v.
Nombre de personnes seules	3984199	4033703	1,24	2,3	1,74	0,0413	Jackknife EDTR
Nombre de familles de taille 2	3407517	3445625	2,24	2,5	1,44	0,0752	intrapolation de l'EDTR
Nombre de familles de taille 3+	4777024	4742161	-2,63	2	1,17	0,1204	intrapolation de l'EDTR
Nombre de personnes mariées	12471961	12435254	-0,29	1	0,95	0,1721	tableau brut de l'EDTR
Nombre de personnes célibataires	6411378	6378680	-0,51	1,8	0,91	0,1812	tableau brut de l'EDTR
Nombre de personnes séparées	648956	727732	12,14	7	5,57	0,0000	tableau brut de l'EDTR
Nombre de personnes dont l'état matrimonial est inconnu	192204	128209	-33,3	11,1	9,64	0,0000	tableau brut de l'EDTR
Nombre de personnes habitant une région rurale	3434513	3430380	-0,12	2,7	0,14	0,4430	tableau brut de l'EDTR
Nombre de personnes habitant une région 100K-499 999K	4759717	4921432	3,4	2	5,46	0,0000	tableau brut de l'EDTR
Nombre de personnes habitant une région 500K+	14151545	13939624	-1,5	0,7	6,88	0,0000	tableau brut de l'EDTR
Total des gains (en millions $)	421029	425795	1,13	1,94	1,88	0,0304	fonction généralisée de l'EDTR
Total des placements (en millions $)	23863	24617	3,16	12,33	0,82	0,2051	fonction généralisée de l'EDTR
Total des transferts gouvernementaux (en millions $)	76467	76196	-0,35	3,07	0,37	0,3553	fonction généralisée de l'EDTR
Total des autres revenus monétaires (en millions $)	44103	45033	2,11	3,56	1,90	0,0285	fonction généralisée de l'EDTR
Total du revenu (en millions $)	563583	569553	1,06	1	3,40	0,0003	Jackknife EDTR
Revenu moyen : famille	56955	57290	0,59	1,21	1,56	0,0591	EFC=0,69. EDTR=EFC/0,57
Revenu moyen : personnes seules	24371	24821	1,84	2,21	2,69	0,0036	EFC=1,26. EDTR=1,26/0,57
Revenu moyen : personnes 16+	25347	25605	1,02	1	3,27	0,0005	Jackknife EDTR. EFC=0,57
Pourcentage de personnes en deçà du SFR	18,6	18,01	-3,17	3,1	3,29	0,0005	Jackknife EDTR