Nº 96-12 au catalogue

QUELS SONT LES IMPACTS SUR LA QUALITÉ DES DONNÉES LORSQUE DES DONNÉES D'ENQUÊTE SUR LE REVENU SONT COMBINÉES À DES DONNÉES ADMINISTRATIVES?

Numéro d'enregistrement du produit 75F0002M

Novembre 1996

Sylvie Michaud, Division des méthodes d'enquêtes sociales

Michel Latouche, Division des méthodes d'enquêtes sociales

La série de documents de recherche de l'EDTR est conçue en vue de communiquer les résultats des études ainsi que les décisions importantes ayant trait à l'Enquête sur la dynamique du travail et du revenu. Ils sont offerts en français et en anglais. Pour obtenir une description sommaire des documents disponibles ou un exemplaire de ces documents, communiquez avec Philip Giles, EDTR, par la poste à Édifice Jean-Talon, 7e étage, section C6, Statistique Canada, Ottawa (Ontario), Canada, K1A 0T6; par INTERNET: GILES@STATCAN.CA; par téléphone au (613) 951-2891; ou par télécopieur au (613) 951-3253.

SOMMAIRE

En général, l'erreur de mesure pose des problèmes. Beaucoup de travaux ont été effectués à ce sujet, à la fois pour mesurer l'erreur et pour la compenser. Il a été démontré que l'erreur de mesure peut être plus problématique dans le cadre d'une enquête longitudinale, particulièrement si les données sont utilisées pour effectuer des régressions. L'Enquête sur la dynamique du travail et du revenu (EDTR) est une enquête longitudinale qui vise à déterminer les répercussions qu'ont sur le revenu les changements touchant l'activité sur le marché du travail et les caractéristiques familiales. Afin de réduire le fardeau de réponse et d'améliorer la qualité des données, on offre un choix aux répondants dans le cadre de l'enquête, à savoir répondre aux questions sur le revenu ou autoriser les responsables de l'EDTR à avoir accès à leurs dossiers administratifs. Le présent document vise à quantifier les répercussions de cette approche mixte sur l'erreur de réponse, et plus particulièrement sur les mesures du changement.

Ce document a été présenté dans le cadre du Symposium 96 de Statistique Canada intitulé *Erreurs non dues à l'échantillonnage*, qui s'est tenu à Ottawa en novembre 1996.

TABLE DES MATIÈRES

1. INTRODUCTION

L'Enquête sur la dynamique du travail et du revenu (EDTR) est une enquête longitudinale qui mesure les répercussions qu'ont sur le revenu les changements touchant l'activité sur le marché du travail ou la situation familiale. Les personnes comprises dans un panel donné demeurent dans l'échantillon pendant six ans et sont interviewées deux fois par année. En janvier, on recueille des données sur le travail, tandis qu'en mai, on recueille des données sur le revenu. L'interview sur le revenu est effectuée en mai parce que les Canadiens doivent produire leur déclaration de revenu avant à la fin d'avril et que l'on croit de façon générale qu'ils sont plus susceptibles de fournir des données précises sur leur revenu à ce moment-là. En mai 1995, dans le cadre d'un effort en vue de réduire le fardeau de réponse, on a offert aux répondants de l'EDTR de répondre aux questions sur le revenu ou d'autoriser les responsables de l'enquête à obtenir des données sur leur revenu à partir de leur déclaration de revenu. Cette offre leur est faite chaque année et, après trois ans de collecte, plus de 75 % des répondants se prévalent de la deuxième option. Toutefois, l'intégration des données d'enquête et des données fiscales ne se fait pas sans problème. Les définitions ne sont pas toujours compatibles et il existe des problèmes de couplage. D'un autre côté, il faut toutefois tenir compte des problèmes de qualité des données qui découlent généralement des enquêtes sur le revenu (par exemple la sous-déclaration de certaines sources de revenu) et de la nécessité de procéder à une imputation. Le présent document donne un aperçu des diverses sources d'erreurs que présente cette méthode et fait état de certaines des répercussions que comporte cette approche mixte. La recherche a été axée sur les microcomparaisons et on a tenté de quantifier les répercussions sur les mesures du changement.

2. PLAN D'ÉCHANTILLONNAGE DE L'EDTR

L'échantillon de l'EDTR est choisi à partir d'un plan d'échantillonnage à plusieurs degrés. Les répondants sélectionnés pour faire partie de l'échantillon de l'EDTR ont préalablement participé à l'Enquête sur la population active (EPA) pendant six mois. On les interviewe deux fois par années pendant six ans. Une première interview, qui se tient en janvier, sert à recueillir des données détaillées sur le travail. Elle permet en outre d'avoir un aperçu des changements qui touchent la composition de la famille, et des dates où ces changements se sont produits. La deuxième interview, qui se tient en mai, sert à recueillir des données détaillées sur le revenu, selon 24 catégories. Dans le cadre de l'enquête, on recueille aussi des données sur l'impôt sur le revenu versé, afin d'évaluer le revenu après impôt. Des données sur le revenu sont recueillies pour chaque membre du ménage âgé de 16 ans et plus et sont agrégées au niveau de la famille, afin de déterminer les mesures de faibles revenus.

La collecte de données sur le revenu n'est pas nouvelle à Statistique Canada. Dans le cadre de l'Enquête sur les finances des consommateurs (EFC), en recueille des données annuelles sur le revenu depuis les 30 dernières années, et les questions sur le revenu de l'EDTR sont identiques à celles de l'EFC. L'expérience acquise dans le cadre de cette dernière est très utile pour l'EDTR.

En général, les enquêtes sur le revenu obtiennent un taux de réponse plus faible que les autres enquêtes. Alors qu'une enquête qui ne porte pas sur le revenu, comme l'Enquête sur la population active, obtient généralement un taux de réponses de 95 %, le taux de réponse de l'Enquête sur les finances des consommateurs n'est que de 80 %. Quant à l'interview sur le revenu de l'EDTR, elle obtient un taux de réponse de 76 %. Les données ont aussi été couplées avec celles d'autres sources pour en évaluer la qualité. À partir de ces comparaisons,

on a déterminé qu'il y a sous-déclaration de certaines sources de revenu, par exemple les prestations d'assurance-chômage, les prestations d'aide sociale, ainsi que les intérêts et les dividendes[5].

Les données fiscales ont aussi été utilisées plus récemment comme source de données sur le revenu. De façon plus particulière, le fichier de données administratives longitudinales (DAL) est un fichier longitudinal fondé sur les données fiscales [7]. Un échantillon de 10 % des répondants à l'enquête a été choisi au hasard et les familles ont été rebâties, selon les renseignements fournis dans les déclarations de revenu (les conjoints et les enfants ont été créés à partir d'autres champs de la déclaration de revenu). Le fichier DAL ne correspond pas exactement aux autres sources de données administratives. On ne peut créer que des familles de recensement (c.à.d. père-mère-enfants), et on a tendance à surestimer le nombre de familles ne comptant qu'une personne. Il y a aussi sous-représentation de certains groupes d'âge (particulièrement les personnes plus âgées) et de faible revenu. Récemment, toutefois, avec l'avènement des crédits d'impôt, ou obtient à partir du système fiscal une meilleure couverture de la population de l'univers. Le problème de la reconstruction des familles subsiste toujours. Par ailleurs, la qualité des données sur le revenu provenant de sources fiscales est jugée supérieure à celle des données d'enquête.

Dans le cadre de l'EDTR, on utilise les deux approches, afin de maximiser les taux de réponse et la qualité des données. Cette façon de faire comporte toutefois des problèmes.

3. SOURCES THÉORIQUES DES ERREURS DÉCOULANT DE L'UTILISATION DE DONNÉES D'ENQUÊTE ET DE DONNÉES FISCALES

Le principe de l'utilisation de données administratives comme sources de données sur le revenu comporte un certain nombre d'enjeux. Par exemple, le moment de la production de ces données peut avoir des répercussions sur les dates cibles de l'enquête. Le couplage avec les fichiers administratifs avec ou sans identificateur unique peut aussi poser certains problèmes, et si l'on tente d'évaluer les répercussions globales d'une stratégie mixte, on doit tenir compte de ces éléments. Toutefois, dans le cas présent, la discussion se limitera aux questions de couplage des deux sources, du point de vue de la qualité des données. Une discussion générale de l'utilisation des dossiers fiscaux dans le cadre de l'EDTR figure dans l'ouvrage [1] de la bibliographie.

Afin d'assurer une mesure globale de la qualité, les enquêtes doivent permettre de calculer l'erreur quadratique moyenne, c'est-à-dire la somme de la variance d'une variable donnée et le carré du biais. Cette opération est généralement ardue parce que le biais peut être très difficile à mesurer. Le tableau 1 vise à déterminer les avantages ou inconvénients possibles de chacune des méthodes (données d'enquêtes et données fiscales) ainsi qu'à préciser sur quelle composante de l'erreur elles peuvent avoir des répercussions.

La couverture n'est touchée que par l'utilisation des données fiscales. Le champ des dossiers fiscaux s'est élargi au fil des ans, et il comprend maintenant 94 % de la population âgée de 20 ans et plus. Si la population des personnes qui ne produisent pas de déclaration de revenu diffère de celle constituée des personnes qui en produisent une, cela peut entraîner un biais dans les données. Les étudiants, par exemple, sont un groupe susceptible d'être sous-représenté chez les personnes

qui produisent une déclaration de revenu; le fait qu'ils sont généralement associés à un plus faible revenu et qu'il peut exister une différence entre les étudiants qui produisent une déclaration de revenu et ceux qui n'en produisent pas peut créer un biais.

Tableau 1. Comparaison de la collecte au moyen de l'enquête et de l'utilisation des données fiscales.

	Enquête seulement	**Données fiscales seulement**
couverture de la population (biais)		↘ personnes qui produisent une déclaration de revenu uniquement
taux de réponse (total) (variance) (biais)	↘ nature délicate du sujet fardeau de réponse dépistage	↗ toutes les personnes qui produisent une déclaration de revenu sont des «répondants»
		↘ aucun couplage ou couplage inapproprié
erreur de réponse (biais)	↘ ↘ sous-déclaration de certaines sources de revenu (assurance-chômage, intérêts...)	↘ sous-déclaration de certaines sources de revenu (économie souterraine)
	↘ arrondissement déclaration par personne interposée	
		↘ sources de revenu non imposable

	Enquête seulement	**Données fiscales seulement**
cohérence des séries chronologiques (biais)	↘ erreur de réponse et incohérences longitudinales	
		↘ incohérences possibles quant aux définitions des catégories de revenu

↗ aspect positif ↘ aspect négatif

La non-réponse peut entraîner des problèmes, tant du point de vue de la variance que du biais. Le revenu est un sujet délicat pour certains répondants et les questions s'y rapportant tendent à obtenir un taux de réponse «plus faible» dans le cadre d'une enquête. Le fait que l'EDTR soit une enquête longitudinale a aussi des répercussions sur le taux de réponse. Au fil des ans, des personnes déménagent, et l'impossibilité de les dépister a aussi pour effet de diminuer le taux de réponses. Le niveau de différence entre les non-répondants et les répondants déterminera l'amplitude du biais. L'utilisation des données fiscales devraient compenser pour certains de ces problèmes en théorie. Lorsqu'une personne produit une déclaration de revenu, il devrait être possible de retracer son dossier fiscal, ce qui devrait augmenter le taux de réponse. Toutefois, dans le cadre de l'EDTR, on ne recueille pas de numéros d'assurance sociale, lesquels constituent un lien unique avec le dossier fiscal. D'autres champs, qui seront décrits à la section 4, sont utilisés dans le cadre d'une méthode de couplage statistique pour établir un lien entre les personnes qui appartiennent à l'échantillon de l'EDTR et les dossiers fiscaux les concernant. Certaines mesures de contrôle de la qualité sont utilisées pour améliorer la qualité du couplage, mais il existe toujours une possibilité de couplage erroné ou de non-couplage, même dans le cas d'une

personne qui produit une déclaration de revenu. Cela aussi a pour effet de diminuer le taux de réponse.

L'erreur de réponse a été étudiée dans le cas de certaines variables du revenu pour lesquelles il existait une source externe en vue de valider les résultats et d'évaluer les biais possibles. Certaines études ont fait ressortir qu'il y avait sous-déclaration de certaines sources de revenu pour les données recueillies dans le cadre d'enquêtes. Pour l'EFC, par exemple, on ne saisit qu'environ 80 % des prestations d'assurance-chômage, comparativement à 94 % grâce au système fiscal. Les revenus de placements sont aussi sujets à la sous-déclaration. Cela crée un biais dans les résultats. Par ailleurs, on pense de façon générale que les données fiscales sont sujettes à la sous-déclaration dans le cas de certaines sources de revenu qui sont liées à l'économie souterraine. Toutefois, étant donné que dans le cadre de l'EDTR on demande aux répondants de consulter leurs dossiers fiscaux pour fournir des renseignements sur leur revenu, et étant donné qu'il n'est pas évident que ces répondants déclareraient ce genre de revenus dans le cadre d'une enquête, on peut conclure que les données fiscales peuvent aussi donner lieu à un biais, ce biais n'étant toutefois pas aussi prononcé que celui des données d'enquête.

L'arrondissement des montants de revenu déclarés dans le cadre d'une enquête constitue une deuxième source d'erreur de réponse. L'arrondissement du revenu total déclaré pose un problème, que vient aggraver l'arrondissement à partir de diverses sources de revenu, le revenu total correspondant alors à la somme de ces sources.

Une troisième source d'erreur de réponse, qui touche les données fiscales, est l'absence de certaines sources de revenu dans le formulaire de déclaration de revenu. Certains revenus non imposables sont déclarés, mais ils se limitent à ceux qui doivent l'être pour le calcul des crédits d'impôt. Les revenus provenant des

gains de loterie et d'héritages ne figurent pas dans la déclaration de revenu, mais sont compris dans la collecte effectuée dans le cadre de l'enquête.

Les changements quant aux définitions et aux concepts fiscaux constituent une quatrième source d'erreur de réponse. Les séries chronologiques peuvent être affectées par les changements apportés à la réglementation touchant l'impôt sur le revenu.

Comme on peut le constater, les deux sources de données comportent des problèmes. La présente étude vise à déterminer les répercussions de la collecte mixte effectuée dans le cadre de l'EDTR sur la qualité des données. Étant donné la nature longitudinale de l'EDTR, les mesures du changement sont importantes. Les erreurs de réponse posent davantage de problèmes dans le cadre d'une enquête longitudinale que d'une enquête transversale, étant donné qu'on s'attend généralement à ce que la corrélation entre les mesures répétées soit plus grande que la corrélation entre les erreurs de réponse. Étant donné les arrondissements possibles et la sous-déclaration des erreurs, on s'attend à ce que les données de sources administratives aient un taux d'erreur de réponse inférieur à celui des données d'enquête.

De façon plus particulière, partons de l'hypothèse qu'on veuille quantifier une variable X (revenu), mais qu'on mesure $x = X + u$ en réalité, u correspondant à l'erreur de réponse. Dans une régression,
où on désire obtenir : $Y = X\beta + \epsilon$,
on mesure en réalité : $y = x\beta' + \epsilon$
où le biais de β' tend vers zéro, avec les hypothèses habituelles d'erreurs indépendantes et normales. Si on voulait analyser le changement $\Delta Y = Y_{t+1} - Y_t$ par régression, on constate dans [1] que le biais dans l'équation du changement est plus grand que le niveau, ce qui s'exprime par la formule mathématique

$$\frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2} < \frac{\sigma_{\Delta x}^2}{\sigma_{\Delta x}^2+\sigma_{\Delta u}^2}$$

4. DÉTERMINATION EMPIRIQUE DES SOURCES D'ERREURS

Pour la première année de collecte des données de l'EDTR, on n'a pas utilisé les déclarations de revenu et on a recueilli les données sur le revenu directement auprès des répondants. Afin d'assurer la qualité des données, on a procédé à une comparaison entre les données d'enquête et les données fiscales. Cela a permis de déterminer et de quantifier les sources d'erreurs.

Les données fiscales sont le résultat d'un couplage, mais étant donné que l'on ne demande pas aux répondants leur numéro d'assurance sociale (NAS), le couplage se fait au moyen d'une méthode statistique. Les dossiers ont d'abord fait l'objet d'un couplage direct à partir du nom, du code postal, de la date de naissance, du sexe et de l'état matrimonial. Grâce à cette méthode, on a couplé 50 % des dossiers. Les dossiers qui n'ont pas été couplés ont fait l'objet d'un couplage statistique (tenant compte des valeurs manquantes ou des écarts pour un ou plusieurs des éléments de couplage). Cela a donné lieu à un couplage global de 85 %. L'étude a mis l'accent sur le taux de réponse, la couverture, le couplage et les erreurs de réponse. Une attention spéciale a été accordée aux répercussions sur les tendances annuelles.

4.1 Taux de réponse et biais possibles

Le fichier de l'échantillon de l'EDTR a été couplé aux dossiers fiscaux de 1993, à partir de méthodes de couplage statistique et direct. Le tableau 2 présente la

distribution de l'échantillon, selon le taux de réponse à l'interview sur le revenu et les résultats du couplage avec les données fiscales.

Tableau 2. Taux de réponse en fonction du couplage avec les données fiscales.

	Aucun couplage avec les données fiscales	Couplage avec les données fiscales	Total
Répondants	3 605	20 651	24 256 (76 %)
Non-répondants	1 774	5 709	7 483 (24 %)
Total	5 379 (17 %)	26 360 (83 %)	31 739

Si toutes les personnes dont les réponses ont fait l'objet d'un couplage avec les dossiers fiscaux avaient accepté cette façon de faire, on aurait assisté dans les faits à une augmentation du taux de réponse. Toutefois, seulement 75 % des répondants ont approuvé l'utilisation de leurs données administratives. Par ailleurs, dans le cadre de l'EDTR, on tente aussi de recueillir des données sur le revenu des personnes qui disent ne pas remplir de déclaration de revenu. Dans l'ensemble, deux groupes de personnes pourraient avoir eu des répercussions sur les taux de réponse. D'abord, les non-répondants à l'interview sur le revenu ayant autorisé l'accès à leurs donnés fiscales et ayant fait l'objet d'un couplage auraient eu un effet positif sur le taux de réponse. Viennent ensuite, les personnes ayant autorisé l'accès à leur dossier fiscal, mais n'ayant pas fait l'objet d'un couplage, auraient eu un effet négatif. Chacun des deux groupes comportait environ 1 700 personnes. Cela signifie que le taux de réponse est demeuré le même avec la méthode mixte.

Toutefois, en raison des biais possibles découlant de différences entre les répondants ayant fait l'objet d'un couplage et ceux n'ayant pas fait l'objet d'un couplage, on a comparé ces deux groupes, à partir des données sur le revenu

recueillies au cours de la première année de l'enquête. Trois sous-groupes comportaient des différences marquantes : les célibataires âgés de 15 à 19 ans, les célibataires âgés de 20 à 24 ans et les femmes mariées âgées de 45 ans et plus. Chez ces groupes, on notait un pourcentage élevé de dossiers non couplés et, de façon générale, les revenus des personnes ayant fait l'objet d'un couplage et de celles n'ayant pas fait l'objet d'un couplage étaient différents (ces dernières avaient un revenu inférieur). Par la suite, on a comparé les groupes de personnes n'ayant pas fait l'objet d'un couplage, selon que ces personnes avaient donné leur autorisation ou non. Cinq grandes catégories ont été utilisées pour effectuer ces comparaisons : les revenus d'emploi (salaire et revenu d'emploi autonome), les revenus de placements (revenus de placement imposables, y compris intérêts et dividendes), les paiements de transfert gouvernementaux (assurance-chômage, aide sociale, prestation fiscale pour enfants, sécurité de la vieillesse, Régime de pension du Canada, indemnisation des accidents du travail et crédits pour la taxe sur les produits et services (TPS)), et le revenu total. Les comparaisons ont été effectuées pour un sous-ensemble de dossiers étiquetés répondants «acceptables». On a procédé ainsi pour écarter les effets possibles de l'imputation. Les résultats figurent au tableau 3. Une tendance similaire s'est dégagée pour toutes les catégories de revenu. Cela laisse supposer que si l'on définissait un ensemble approprié de répondants, un modèle d'imputation relativement valide pourrait être établi pour les personnes n'ayant pas fait l'objet d'un couplage, étant donné qu'il ne semble pas y avoir de différence entre celles qui ont donné leur autorisation et celles qui ne l'ont pas fait.

Tableau 3. Comparaisons du revenu total des répondants «acceptables» (à partir des données d'enquête) pour les répondants qui ont autorisé l'utilisation de leurs dossiers administratifs et pour ceux qui ne l'ont pas fait.

	«répondants acceptables» qui ont donné leur autorisation			«répondants acceptables» qui n'ont pas donné leur autorisation		
	n	moyenne	médiane	n	moyenne	médiane
célibataires âgés de 15 à 19 ans	865	2 624 $	1 500 $	727	2 458 $	1 000 $
célibataires âgés de 20 à 24 ans	606	10 987 $	8 800 $	470	9 623 $	7 188 $
femmes mariées âgées de 45 ans et plus	1 599	12 771 $	7 677 $	1 183	13 573 $	8 160 $
autres	7 857	25 657 $	20 000 $	5 509	26 667 $	21 567 $

4.2 Erreur de réponse

Des comparaisons ont été effectuées pour un certain nombre de catégories de revenu à partir des données d'enquête et des données fiscales. La référence [5] parle d'une comparaison entre des données d'enquête et des données fiscales pour les répondants «acceptables», en vue de déceler les différences possibles quant aux définitions des catégories de revenu. Leurs conclusions laissent supposer qu'il existait des différences dans le cas des revenus du travail autonome et de l'aide sociale. Il n'existait pas de différence entre le revenu moyen et le revenu médian pour les salaires et les prestations d'assurance-chômage (a.-c.). Toutefois, on notait encore une certaine sous-déclaration des prestations d'assurance-chômage. La difficulté de cette méthode a consisté à déterminer ce qui était juste. De façon plus particulière, lorsqu'il y avait des différences quant à la déclaration de revenus d'emploi autonome, il était difficile de déterminer s'il s'agissait de revenus de

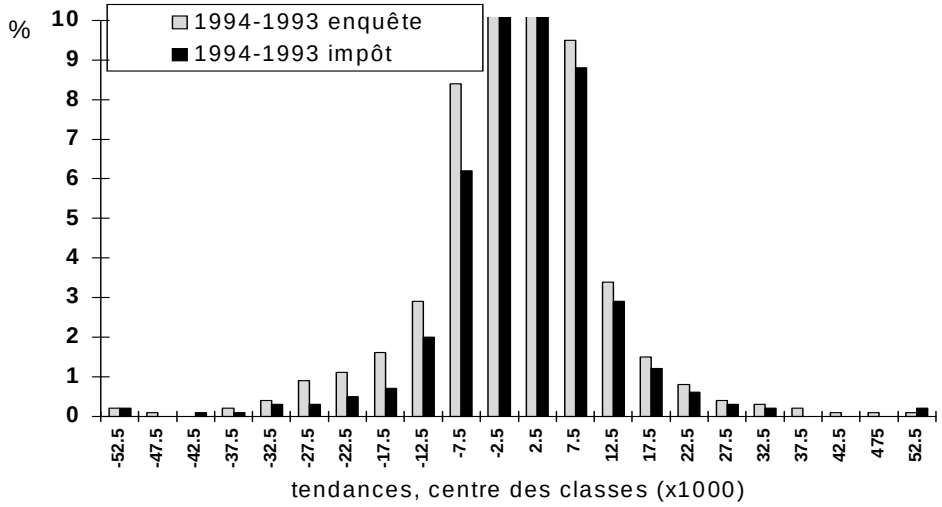
l'économie souterraine, lesquels n'auraient pas figuré dans les déclarations d'impôt, de revenus déclarés provenant d'une autre source ou encore de revenus mal déclarés. Étant donné que l'EDTR est axée principalement sur l'analyse longitudinale, il a été décidé d'étudier les différences quant aux mesures du changement et de tenter de concilier les microdifférences des dossiers comportant deux années de données. Cela permet en outre d'étudier l'erreur de réponse en tenant compte de cet aspect longitudinal.

On ne disposait de deux années de données d'enquête et de données fiscales que pour un sous-ensemble seulement des répondants de l'EDTR. Cela a limité l'étude à un sous-échantillon de 4 274 répondants. Ce sous-échantillon n'est pas très représentatif de l'échantillon global; il comporte un pourcentage légèrement plus élevé de personnes âgées de 65 ans et plus et une représentation plus faible du groupe des très jeunes, âgés de 16 à 19 ans. Les écarts n'ont pas été jugés suffisamment importants pour invalider l'étude. Dans ce sous-ensemble, 86 % des dossiers avaient été obtenus grâce au couplage direct, ce qui fait que l'on a conservé uniquement ce sous-ensemble, pour supprimer les répercussions possibles des dossiers mal couplés. Les comparaisons se sont limitées à ce sous-ensemble de 3 670 dossiers. Toutefois, on a supprimé 600 autres dossiers, dont 400 comportaient une non-réponse partielle pour la deuxième année, et la majeure partie des 200 autres, aucun revenu déclaré pour une année ou les deux.

On a accordé une attention particulière aux mesures du changement quant au revenu total entre les deux années, à partir à la fois des données d'enquête et des données fiscales. Pour chaque personne, un changement de revenu total ou micro-tendance a été calculé pour les données d'enquête et les données fiscales. L'histogramme 1 montre la distribution des changements. L'axe des ordonnées a été limité à 10 % pour avoir un meilleur aperçu des queues de la distribution (l'axe aurait dû atteindre près de 30 %). Le fait d'utiliser les données fiscales pour

calculer le changement moyen a eu pour résultat d'augmenter le revenu de 498 $, tandis que le recours aux données d'enquête a eu pour effet de le diminuer de 3 $

HISTOGRAMME 1.
Distribution des micros-tendances annuelles pour le revenu total, par source d'information

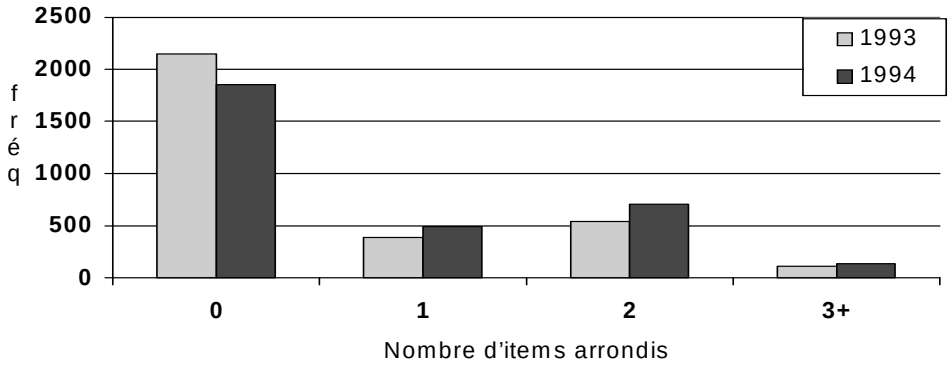


(l'écart était significatif au niveau de 1 %). Il semble en outre y avoir plus de variabilité dans les mesures du changement à partir de l'enquête.

Puis, on a comparé les modèles de déclaration. Un examen initial des données a fait ressortir deux comportements différents, selon qu'une personne fournissait des montants approximatifs (ce que l'on a décelé en examinant les sources de revenu ayant fait l'objet d'un arondissement) ou des montants exacts. L'étude a donc mis l'accent sur le niveau d'arrondissement. L'étude s'est limitée à l'arrondissement de dix catégories de revenu, les autres catégories de revenu n'étant pas déclarées de façon similaire dans les données d'enquête et les données fiscales. On a déterminé que les montants arrondis étaient ceux dont les deux derniers chiffres étaient de zéro dans les données d'enquête, mais pas dans les données fiscales. L'histogramme 2 montre la distribution des répondants selon le nombre d'éléments de revenu arrondis dans les données d'enquête pour les années de référence 1993

et 1994. Il arrive fréquemment que les montants de revenu soient arrondis. En effet, seulement 1 530 dossiers, ce qui représente 47,5 % des personnes, ne comportaient pas de montants arrondis de revenu pour les deux années. Il semble en outre que le niveau d'arrondissement ait augmenté pour la deuxième année de

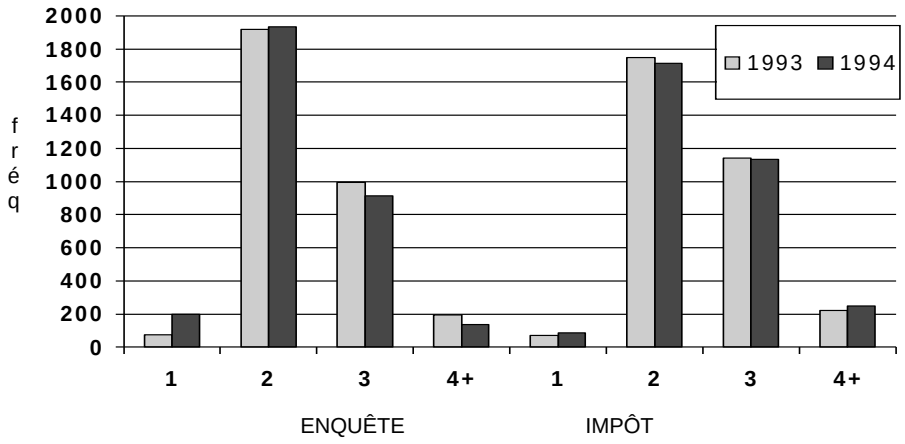
HISTOGRAMME 2.
Distribution du nombre d'items arrondis par individu par an
basée sur 10 items



collecte.

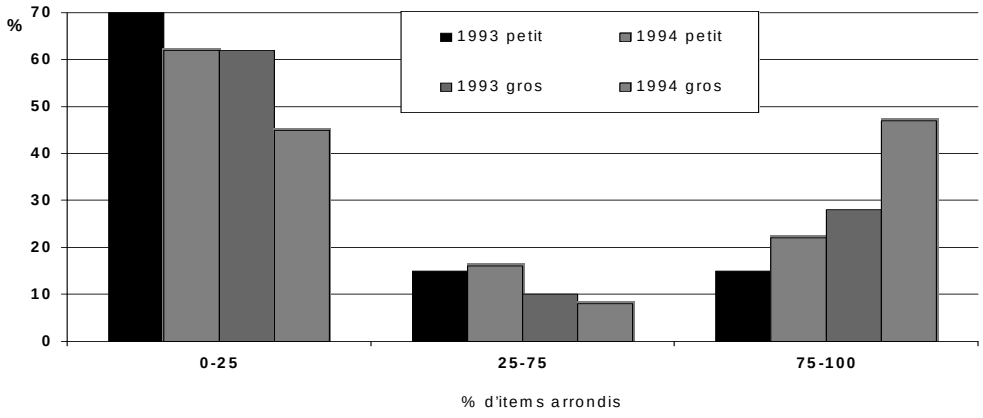
L'histogramme 3 montre la distribution des répondants selon le nombre d'éléments non nuls déclarés dans le cadre de l'enquête. Il est intéressant de noter que le nombre moyen d'éléments déclarés dans le cadre de l'enquête a diminué

HISTOGRAMME 3.
Distribution des items non-nuls par individu
par année et par source



légèrement pour la deuxième année d'enquête, par rapport à la première année, mais cela ne ressort pas des données fiscales déclarées pour ces mêmes éléments. Il était intéressant de déterminer si l'arrondissement a été différent pour les divers groupes de répondants. On a comparé le comportement en matière d'arrondissement au fil des ans pour divers groupes de revenu. Comme le montre l'histogramme 4, le revenu semble avoir des répercussions sur l'arrondissement, c'est-à-dire que les personnes qui ont un revenu élevé ont tendance à l'arrondir

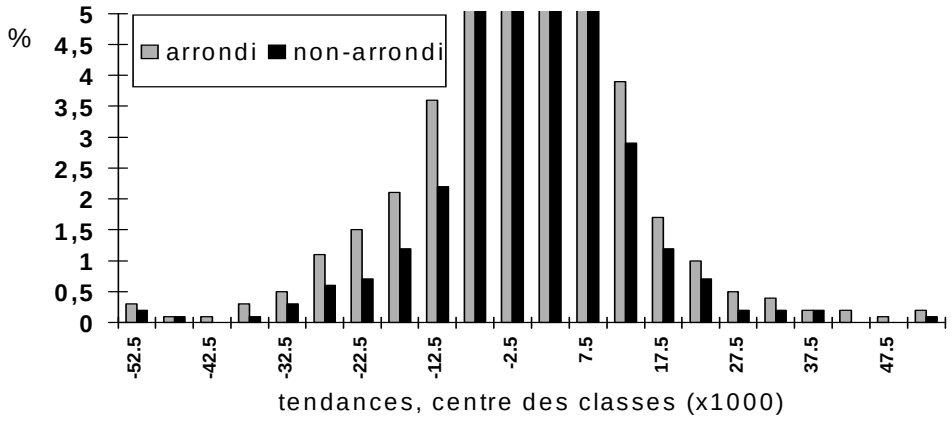
HISTOGRAMME 4.
Distribution des individus par proportion d'items arrondis
années de référence 1993 et 1994
pour les répondants avec un petit et un gros revenu total



davantage que celles qui ont un revenu faible. Les personnes âgées, quant à elles, semblent moins susceptibles d'arrondir les montants qu'elles fournissent.

Encore une fois, on a comparé le revenu total. Les mesures du changement ou des microtendances ont fait l'objet de comparaisons à partir des données d'enquête, après division des répondants en deux groupes : les répondants ayant arrondi leur revenu déclaré et les répondants ne l'ayant pas arrondi. Les résultats figurent à l'histogramme 5. Un répondant a été mis dans le groupe ayant fourni des revenus arrondis s'il avait arrondi au moins un des éléments déclarés. Il est intéressant de

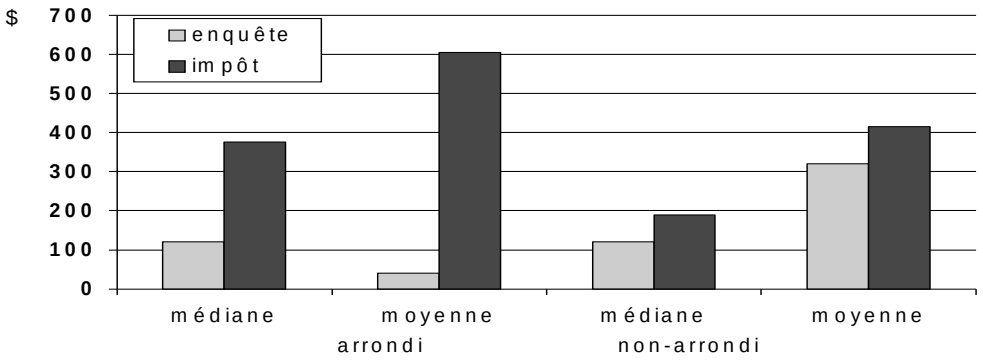
HISTOGRAMME 5.
Distribution des micros différences (1994-1993)
pour les groupes "arrondi" et "non-arrondi"



noter qu'une proportion importante de la variabilité des mesures du changement observée à l'histogramme 1 a trait au groupe de répondants qui ont arrondi leur revenu pour au moins une source.

Pour confirmer cette hypothèse, les mesures du changement ont été comparées entre les données d'enquête et les données fiscales pour le groupe de répondants qui n'avaient pas arrondi les montants de revenus déclarés. L'histogramme 6

HISTOGRAMME 6.
Mesures de centralité des micros-tendances
pour les groupes arrondi et non-arrondi



indique la microtendance moyenne et médiane séparément pour le groupe aux montants «arrondis» et celui aux montants «non arrondis». Dans le cas des personnes qui n'ont pas arrondi leur revenu, les différences n'étaient pas si importantes. Comme on pouvait s'y attendre, les écarts les plus grands ont été enregistrés pour le groupe des personnes ayant arrondi leur revenu.

5. CONCLUSIONS

Nous n'avons fait ici qu'un survol de la question de l'erreur de réponse. Lorsque certaines des différences les plus importantes ont été examinées par des spécialistes, elles ont été attribuées à une erreur de réponse dans le cadre de l'enquête pour approximativement 80 % des cas. Environ 10 % des différences ont été attribuées à une «erreur» dans les données fiscales (un élément non imposable était absent pour une des deux années ou il semblait y avoir une erreur dans le champ de la déclaration de revenu). Enfin, les autres 10 % de différences n'ont pas pu être expliqués.

On est arrivé à certaines conclusions intéressantes, par exemple : environ 30 % des gens étudiés ont fourni exactement les mêmes montants (au dollar près) dans le cadre de l'enquête et dans les dossiers fiscaux, pour au moins une année de données. Pour le reste des répondants, une erreur de réponse a été notée dans le cadre de l'enquête ou dans les dossiers fiscaux, ou encore dans les deux. Les sources de données comportent toutes deux leurs limites; celles des données fiscales ont trait aux personnes qui ne remplissent pas de déclaration de revenu et à la sous-déclaration des revenus non imposables, tandis que les données d'enquête semblent donner lieu à des erreurs de réponse. L'erreur de réponse relative aux données d'enquête semble aussi augmenter avec le temps. À partir des résultats observés, et même si les données fiscales sont sujettes à erreur, le recours à ces dernières dans le cadre de cette approche mixte entraînera probablement une

amélioration de la qualité des données sur le revenu, plus particulièrement du fait de la nature longitudinale de l'enquête. Il subsiste des questions à examiner. En effet, les conclusions générales ne semblent pas s'appliquer aussi bien aux travailleurs autonomes. Ce groupe devrait donc faire l'objet d'une analyse plus poussée. De même, l'étude devrait être plus précise du point de vue de la déclaration des revenus selon la source. Les répercussions des revenus non imposables sur la mesure du changement devraient aussi être évaluées de façon plus détaillée. Enfin, un certain nombre des techniques ont été proposées pour tenir compte de l'erreur de réponse dans les ouvrages [3] et [6] de la bibliographie. Ces techniques devraient être appliquées et mises à l'essai pour vérifier si elles peuvent contribuer à améliorer la qualité des mesures du revenu.

Les auteurs voudraient souligner la contribution de Chantal Grondin, Martin Renaud, Carole Janelle et Elaine Fournier pour la préparation de l'étude.

BIBLIOGRAPHIE

[1] Bound, J., Brown, C., Duncan, G. et Willard, R. (1991), «Measurement Error in Cross-Sectional and Longitudinal Labor Market Surveys: Validation Study Evidence», Panel Data and Labor Market Studies, J. Hartog, G. Ridder et J. Theeuwes (éd.), Elsevier Science Publishers

[2] Dibbs, R., Poulin, S., Webber, M. (1994), «Utilisation des données fiscales dans l'Enquête sur la dynamique du travail et du revenu : rapport sommaire», Série des documents de recherche de l'EDTR, n° 94-11 au catalogue.

[3] Fuller, W. (1987), Measurement error models, Wiley

[4] Groves, R. (1989), Survey Errors and Survey Costs, Wiley

[5] Michaud, S., Dolson, D., Renaud, M., (1995), «Combinaison des données administratives et des données d'enquête en vue d'alléger le fardeau des répondants dans les enquêtes longitudinales», Série des documents de recherche de l'EDTR, n° 95-19 au catalogue.

[6] Plewis, I. (1985), Analysing change, Measurement and Explanation using longitudinal data, Wiley.

[7] Statistique Canada (1995), Un aperçu de DAL, données administratives longitudinales. Rapports de DAL, Réf #94-20-01F, mai 1995.Préparé par la, Division des données régionales et administratives et la division des méthodes d'enquêtes sociales (1995). Statistique Canada.