

ÉVALUATION DE LA QUALITÉ DES ESTIMATIONS DÉMOGRAPHIQUES ET DES DONNÉES FISCALES UTILISÉES DANS LE CALAGE DE DIFFÉRENTES ENQUÊTES À STATISTIQUE CANADA

Sylvie Auger et Johanne Tremblay¹

RÉSUMÉ

Différentes enquêtes de Statistique Canada utilisent pour le calage des estimations démographiques dérivées principalement du dernier recensement et de fichiers administratifs. Ces estimations sont révisées après chaque recensement. De même, certaines enquêtes utilisent des totaux selon le revenu provenant de fichiers de données fiscales. Dans le cadre du développement de la stratégie de calage harmonisée pour les statistiques du revenu, on a fait une évaluation de la qualité de ces estimations et totaux. Cet article présente les résultats de l'évaluation des estimations démographiques dérivées du Recensement de 1996 et les résultats de l'évaluation des différents fichiers administratifs disponibles pour les totaux selon le revenu.

MOTS CLÉS : Contrôles; données fiscales; estimations démographiques; qualité.

ABSTRACT

When calibrating, different surveys at Statistics Canada use demographic estimates derived mainly from the last census and some administrative data files. These estimates are revised after each census. As well, some surveys use totals based on revenue coming from tax data files. Within the framework of the development of a strategy for harmonized calibration of income statistics, we did an evaluation of the quality of these estimates and totals. This paper presents the results of the evaluation of the demographic estimates derived from the 1996 Census and the results of the evaluation of the different tax data files available for the totals based on revenue.

KEY WORDS: Controls, Demographic estimates; Quality; Tax data.

1. INTRODUCTION

Dans la plupart des enquêtes auprès des ménages effectuées par Statistique Canada, on ajuste les poids de sondage pour obtenir des estimations d'enquête, principalement celles du nombre de personnes selon l'âge et le sexe, qui concordent avec des totaux provenant de sources externes considérées plus fiables. Cet ajustement, appelé calage aux marges, est effectué selon une approche décrite dans Deville et Särndall, 1992.

Dans les enquêtes visant à produire des statistiques sur le revenu, les dépenses et la richesse des ménages, des estimations du nombre de ménages, de même que des totaux du nombre de personnes par classe de revenu sont également utilisés à l'étape du calage. La stratégie d'utiliser ces estimations et totaux a principalement été élaborée dans le cadre d'un projet d'harmonisation du calage des statistiques sur le revenu dont le but était d'améliorer la qualité des estimations sur le revenu ainsi que la cohérence entre les estimations de ces différentes enquêtes et des données administratives (Tremblay, 2005). Au cours de ce projet, on a effectué une évaluation approfondie de la qualité de ces estimations et totaux qu'on appellera les contrôles. Cet article vise à présenter certains aspects de cette évaluation.

¹ Sylvie Auger et Johanne Tremblay, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Édifice R.H. Coats, 16^{ième} étage, Ottawa (Ontario), Canada, K1A 0T6, Courriel : Sylvie.Auger@statcan.ca et Johanne.Tremblay@statcan.ca.

Quoique les stratégies de calage de ces enquêtes soient en principe assez similaires, les contrôles peuvent différer en fonction des particularités de chacune des enquêtes. Par exemple, le nombre de catégories d'âge des contrôles du nombre de personnes peut varier d'une enquête à l'autre. Afin d'alléger le contenu de cet article, les résultats seront présentés uniquement en fonction des contrôles de l'Enquête sur les dépenses des ménages (EDM), excluant les territoires parce que la stratégie de calage y est différente de celle des provinces.

Une description des différents types de contrôles démographiques est présentée dans la section 2. Par la suite, la méthodologie d'évaluation de la qualité de ces contrôles ainsi qu'un résumé des résultats sont fournis dans la section 3. Les sections 4 et 5 présentent respectivement une description des contrôles selon le revenu et une évaluation des différents fichiers administratifs disponibles pour dériver ces contrôles.

2. DESCRIPTION DES CONTRÔLES DÉMOGRAPHIQUES

Les contrôles démographiques du nombre de personnes ou de ménages proviennent des données du dernier recensement ajustées pour le sous-dénombrement. Ces contrôles sont ensuite mis à jour afin de représenter la population cible de l'enquête selon l'année de référence. Ces ajustements sont faits à l'aide de fichiers administratifs et de modèles. La méthodologie pour dériver les contrôles du nombre de personnes est décrite dans Bender (1992). Celle développée pour produire des contrôles au niveau des ménages et des familles économiques est présentée dans Schembari (2001).

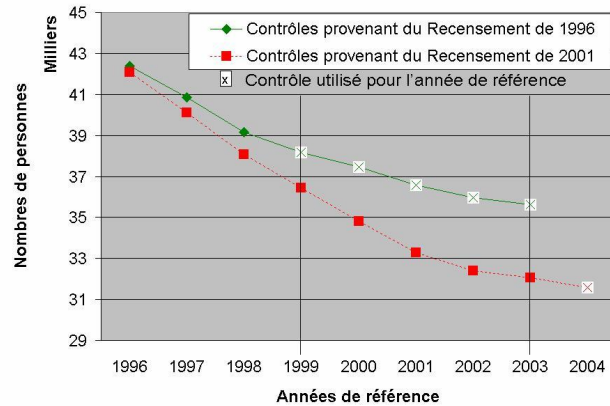
À l'EDM, on utilise des contrôles du nombre de personnes à l'échelle provinciale pour chacun des deux sexes selon neuf catégories d'âge ainsi qu'à l'échelle de 14 régions métropolitaines de recensement (RMR) selon deux catégories d'âge. L'EDM utilise aussi des contrôles du nombre de ménages à l'échelle provinciale selon trois tailles de ménage. Des contrôles du nombre de ménages monoparentaux et de ménages composés d'un couple avec des enfants ont été introduits dans la stratégie de calage de l'enquête à la fin des années 1990 dans le but de corriger un problème de représentativité des ménages monoparentaux (Arsenault et coll., 2001). Ces contrôles sont dérivés à partir des estimations démographiques des familles (Statistique Canada, 2003). Le Tableau 1 résume les quatre types de contrôles de l'EDM.

Tableau 1 : Les contrôles utilisés dans l'EDM

Géographie	Types de contrôle	Catégories	Nombres de contrôles
Province (10)	Âge * Sexe	(0-6, 7-17, 18-24, 25-34, 35-54, 55-59, 60-64, 65-69, 70 et plus) * (Homme, Femme)	180
RMR (14)	Âge	0-17, 18 et plus	28
Province (10)	Taille de ménage	1, 2, 3 et plus	30
Province (10)	Type de ménage	Couple avec enfants, Monoparental	20

Lorsque les données d'un nouveau recensement ajustées pour le sous-dénombrement sont disponibles, elles deviennent la nouvelle base pour produire les contrôles. Cette information est disponible environ deux ans après le recensement. On produit alors non seulement les contrôles démographiques pour les années ultérieures, mais également des contrôles révisés pour les années antérieures. À partir de 2004, les contrôles provenant du Recensement de 2001 ont remplacé ceux de 1996 et des contrôles révisés pour les années antérieures ont été produits. L'EDM a utilisé les contrôles provenant du Recensement de 1996 pour les années de référence 1999 à 2003 et ceux du Recensement de 2001 pour l'année de référence 2004. Le Graphique 1 illustre deux séries de contrôles et ceux utilisés par l'EDM lors de la diffusion des données.

Graphique 1 : Contrôles du nombre d'hommes de 25 à 34 ans à Terre-Neuve



3. ÉVALUATION DE LA QUALITÉ DES CONTRÔLES DÉMOGRAPHIQUES

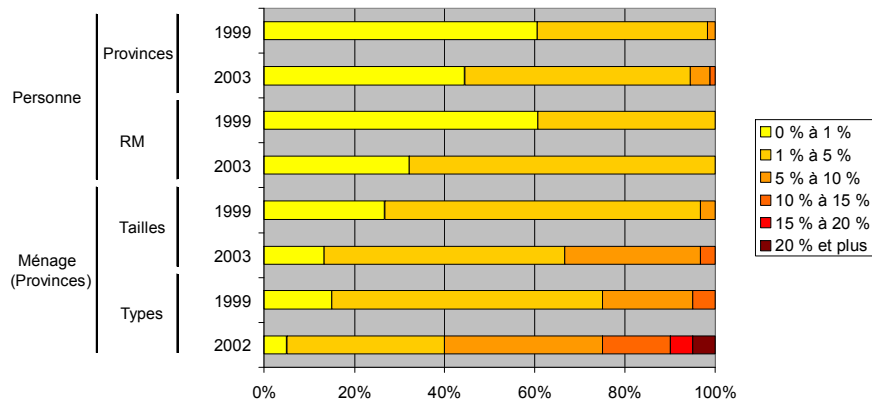
La qualité des contrôles démographiques dépend de la qualité des données du recensement et des fichiers administratifs nécessaires aux mises à jour. Elle dépend également de la validité du modèle d'estimation du nombre de personnes ou du nombre de ménages pour les années subséquentes au dernier recensement. On s'attend donc en général à ce que la qualité des contrôles se détériore à mesure que l'on s'éloigne de ce recensement.

Dans cette évaluation, on veut analyser la qualité des contrôles dérivés du Recensement de 1996. On fait donc l'hypothèse que les contrôles démographiques provenant du recensement le plus récent, le Recensement de 2001, sont plus exacts que ceux produits à partir du Recensement de 1996. Les analyses sont effectuées à partir des différences relatives en pourcentage² par rapport aux contrôles dérivés du Recensement de 2001.

3.1 Comparaison de la qualité des contrôles démographiques utilisés dans l'EDM entre 1999 et 2003

La première comparaison vise à mesurer l'ampleur des différences relatives entre les deux séries de contrôles démographiques, à mesure que l'on s'éloigne du Recensement de 1996. La période d'évaluation comprend les années de référence 1999 à 2003. Cette période correspond aux cinq années où l'EDM a utilisé, lors du calage, des contrôles dérivés à partir du Recensement de 1996. Les résultats pour les années extrêmes, 1999 et 2003, sont présentés pour chaque type de contrôle dans le Graphique 2.

Graphique 2 : Distributions cumulatives des différences relatives selon le type de contrôle pour 1999 et 2003³



On peut voir pour les contrôles du nombre de personnes à l'échelle provinciale en 1999 que les différences relatives en valeur absolue sont inférieures à 1 % pour 60 % des contrôles et inférieures à 5 % pour 98 % des cas. Quatre ans plus tard, en 2003, on observe peu de détérioration de la qualité puisque 95 % des différences demeurent inférieures à 5 %. Seulement 1 % des différences sont entre 10 % et 15 % en 2003. À l'échelle des RMR, les différences des contrôles du nombre de personnes sont moins de 5 % que ce soit en 1999 ou quatre ans plus tard en 2003. Les différences relatives des contrôles du

² $[(C_{1996} - C_{2001}) / C_{2001}] * 100$ ou C_x représente le contrôle selon le Recensement de l'année x .

³ L'année 2002 a été utilisée pour les types de ménages parce que les données de 2003 n'étaient pas disponibles.

nombre de personnes par RMR sont moins élevées que celles par province puisqu'il n'y a que deux catégories d'âge à l'échelle des RMR en comparaison aux neuf catégories d'âge croisées selon le sexe à l'échelle provinciale.

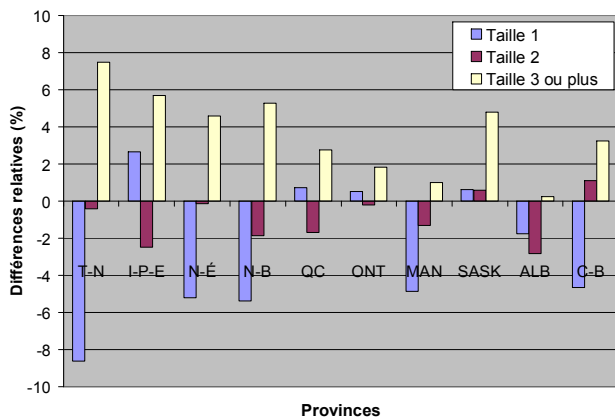
On observe des différences plus élevées pour les contrôles du nombre de ménages que pour les contrôles du nombre de personnes. Ceci s'explique par le fait qu'il existe peu de données administratives concernant les ménages et que la dynamique des ménages rend difficile l'estimation de ces contrôles à partir de modèles. Néanmoins, même en 2003, plus de 65 % des différences relatives pour les contrôles selon la taille du ménage sont inférieures à 5 % et presque toutes les autres sont inférieures à 10 %. Par contre, les différences pour les contrôles du nombre de ménages selon le type de ménage sont plus élevées. L'évaluation pour ce type de contrôle a été effectuée à partir des données de 2002 parce que les données de 2003 n'étaient pas disponibles. Déjà en 2002, on observe des différences supérieures à 5 % pour 60 % des contrôles, certaines de ces différences dépassant les 15 %.

3.2 Évaluation de la qualité en fonction des catégories des différents types de contrôles

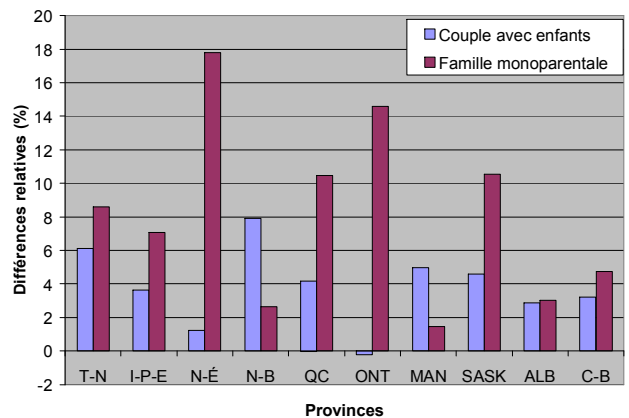
La qualité des contrôles dépend aussi du niveau de détail requis. Par exemple, il sera généralement plus difficile de bien estimer les contrôles au niveau des personnes dans le temps si certaines des catégories en fonction de l'âge, du sexe et de la géographie sont de petite taille. Pour analyser la qualité des contrôles démographiques dérivés à partir du Recensement de 1996 en fonction des catégories, on a choisi de se restreindre aux différences relatives de l'année de référence 2001. Cette année est considérée la plus pertinente parce que les contrôles de 2001 dérivés à partir du Recensement de 2001 sont moins sujets aux erreurs engendrées par les ajustements post-censitaires. La période post-censitaire n'est que de quelques mois.

Le Graphique 3 présente les différences relatives entre les contrôles du nombre de ménages selon la taille dérivés à partir du Recensement de 1996 et ceux provenant du Recensement de 2001. On observe que les différences relatives les plus élevées se trouvent au niveau des contrôles du nombre de ménages de taille 1 et de taille 3 ou plus. De plus, on note que les contrôles de taille 3 ou plus selon le Recensement de 1996 sont toujours supérieurs à ceux du Recensement de 2001. Pour les contrôles du nombre de ménages par type de ménage, on observe à partir du Graphique 4 que les différences sont très grandes pour les contrôles du nombre de ménages monoparentaux, dépassant les 10 % pour près de la moitié des provinces incluant les deux plus grandes provinces, le Québec et l'Ontario.

Graphique 3 : Différences relatives des contrôles selon la taille de ménage et la province



Graphique 4 : Différences relatives des contrôles selon le type de ménage et la province



Des comparaisons similaires ont été faites pour les contrôles du nombre de personnes à l'échelle provinciale selon les groupes d'âge et de sexe. Elles indiquent que les différences relatives pour les contrôles des catégories 0 à 6 ans, 25 à 34 ans et 70 ans ou plus sont les plus élevées, quoiqu'elles restent inférieures à 10 %. De plus, comme pour les contrôles du nombre de ménages, les différences les plus élevées sont souvent observées dans les petites provinces.

4. DESCRIPTION DES CONTRÔLES SELON LE REVENU

Les contrôles selon le revenu sont dérivés à partir du fichier administratif contenant l'état de la rémunération annuelle payée par chaque employeur à chaque salarié (T4) produit sur une base annuelle par l'Agence du revenu du Canada (ARC). Si un salarié a plus d'un employeur, il aura donc autant d'enregistrements sur ce fichier qu'il a d'employeurs.

Lors de la production de ces contrôles, on procède à quelques vérifications et imputations de base du fichier T4. La somme des salaires et traitements est effectuée pour chaque salarié. Les salariés avec des salaires et traitements inférieurs à un certain seuil sont exclus car ces montants semblent être souvent omis par les ménages lors des enquêtes. En 1997, ce seuil était de 1 500 \$ et il est ajusté annuellement au coût de la vie. Les bornes des classes de chaque province sont identifiées à partir de percentiles et on calcule le nombre de salariés pour chacune des classes.

L'EDM, lors de l'étape du calage, utilise les contrôles du nombre de salariés à l'échelle provinciale selon six classes, définies à partir des 25, 50, 65, 75 et 95^{ième} percentiles. Puisque le fichier T4 disponible au moment du calage est celui de l'année précédant l'année de référence de l'EDM, on utilise un modèle afin de dériver le nombre de salariés par classe pour l'année de référence. Les principaux résultats d'une analyse sur la qualité de ce modèle sont présentés dans Tremblay et coll., 2003.

5. ÉVALUATION DE LA QUALITÉ DES CONTRÔLES SELON LE REVENU

Comme on l'a déjà mentionné, les contrôles selon le revenu sont produits à partir des fichiers T4. Cependant, une autre source de données pourrait être utilisée, les fichiers administratifs contenant les déclarations de revenus et de prestations des particuliers (T1) produits sur une base annuelle par l'ARC.

Dans le cadre du projet d'harmonisation du calage des statistiques sur le revenu, les fichiers T4 ont été préférés aux T1. On supposait une meilleure couverture des salariés sur le fichier T4 car la loi exige que tous les employeurs complètent pour chacun de leur salarié un formulaire « État de la rémunération payée » tandis que ce n'est pas tous les particuliers qui sont tenus de compléter la « Déclaration de revenus et de prestations ». Lors de la mise en oeuvre, des écarts plus importants que prévus du nombre de salariés entre les fichiers T1 et les fichiers T4 ont été observés surtout à partir de l'année de référence 1998. Ces écarts ont soulevé l'importance d'évaluer la qualité de ces deux fichiers en fonction des besoins pour le calage des enquêtes afin de vérifier si le fichier T4 demeurerait le meilleur choix.

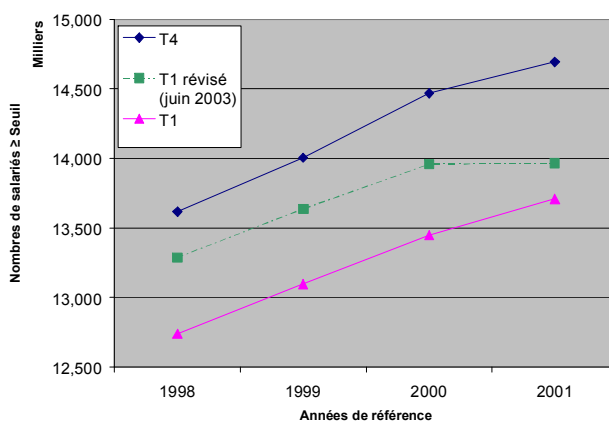
5.1 Comparaison du nombre de salariés entre les fichiers T1 et T4

Les versions finales des fichiers administratifs T1 et T4 que reçoit Statistique Canada de l'ARC sont disponibles environ un an après la fin de l'année de référence. La différence du nombre de salariés entre ces deux fichiers est approximativement d'un million à chaque année pour les années 1998 à 2001 telle qu'illustrée dans le Graphique 5. Cet écart équivaut à environ 6 à 7 % de salariés de moins sur les fichiers T1.

En juin 2003, Statistique Canada a reçu de l'ARC une série de fichiers T1 révisés pour les années 1998 à 2001 qui ont permis de faire une analyse de la couverture des fichiers T1 utilisés par SC. Les fichiers T1 continuent d'être mis à jour même après qu'ils soient remis à SC, par exemple des déclarations reçues plus tard ou des modifications apportées à certaines déclarations, d'où l'existence des fichiers révisés. On doit tenir compte que pour chacun des fichiers révisés produits en juin 2003, la période de temps séparant la version révisée de la version finale varie. Par exemple, le fichier révisé de 2001 a été produit seulement cinq mois après la version finale, produite au début de l'année 2003. Par contre, le fichier révisé de 1998 a été produit trois ans et demi après la production de la version finale. On s'attend donc à ce que le fichier révisé de 1998, qui est celui ayant la période de révision la plus longue, soit le plus complet.

Le Graphique 5 nous indique que les fichiers T1 révisés pour les années 1998 à 2000 contiennent pour chaque année environ 500 000 salariés de plus (soit environ 4 % de plus) que les fichiers T1 utilisés par SC. L'écart en 2001 est plus petit, environ 250 000 salariés, mais la période de révision n'est que de cinq mois pour le fichier révisé de 2001. De plus, on constate que l'augmentation du nombre de salariés des fichiers révisés de 1998 à 2000 est de la même ampleur, bien que la période de révision du fichier de 1998 soit plus longue que celle du fichier révisé de 2000. En résumé, on observe un gain considérable en couverture sur les fichiers révisés, cependant, lorsque la période de révision dépasse un an et demi, le gain devient négligeable.

Graphique 5 : Comparaison du nombre de salariés entre les fichiers T1, T1 révisés et T4



Toujours à partir du Graphique 5, on observe que les écarts entre les fichiers T1 et les fichiers T1 révisés sont plus élevés que ceux entre les fichiers T1 révisés et les fichiers T4⁴. Les fichiers T4 sont donc plus près des fichiers T1 révisés qui eux sont considérés de meilleure qualité que les fichiers T1.

L'écart entre le fichier T4 et le fichier T1 de 1998 est de 6,4 % alors qu'il n'est que de 2,4 % avec le fichier T1 révisé. Une analyse plus approfondie indique que l'écart avec le fichier révisé est en grande partie attribuable aux salariés gagnant moins de 7 000 \$. Ces salariés font probablement partie des gens qui ne sont pas tenus de compléter une déclaration de revenu en raison d'un revenu total plus petit que le minimum à déclarer. Lorsqu'on exclut ces salariés, l'écart entre le fichier T1 révisé et le fichier T4 diminue à 1,2 %.

5.2 Appariement des fichiers T1 et T4 de 2001

Une autre façon d'évaluer la qualité des fichiers T1 et T4 est d'analyser les résultats d'un appariement entre ces fichiers. Le besoin d'une telle évaluation a justifié la permission de faire cet appariement. Les fichiers de 2001 ont été appariés en utilisant le numéro d'assurance social (NAS) comme variable d'appariement. On fait l'hypothèse que les personnes sur le fichier T1 n'ayant aucun revenu en terme de salaires et traitements et qui ne sont pas présentes sur le fichier T4 ne sont pas des salariés. Après avoir exclu ces personnes du fichier T1, le nombre de NAS qui devraient être appariés au fichier T4 est de 15,2 millions. Le fichier T4 contient 15,9 millions de NAS.

Au Tableau 2, on peut voir que 14 millions de NAS sont présents sur les deux fichiers avec des salaires et traitements égaux, à l'intérieur d'une marge de 100 \$. Donc, environ 90 % de ces fichiers ont les mêmes informations. De plus, 600 000 NAS sont aussi présents sur les deux fichiers, mais avec des salaires et traitements différents. Le tiers de ces NAS sont des salariés qui ont plus d'un employeur sur le fichier T4 et dont la somme des salaires et traitements pour un sous-ensemble de leurs employeurs égale le montant présent sur le fichier T1. Il est raisonnable de croire pour ces cas que des salaires et traitements sont manquants sur le fichier T1.

Tableau 2 : Résultats de l'appariement entre les fichiers T1 et T4 de 2001

T1	Statut	T4
Nombre de NAS		Nombre de NAS
	T4 seulement	1 300 000
14 000 000	Appariés et égaux	14 000 000
600 000	Appariés mais non égaux	600 000
600 000	T1 seulement	
15 200 000	Total	15 900 000

Le Tableau 2 indique également que 1,3 million de NAS sur le fichier T4 sont absents du fichier T1. Près de 45 % de ces cas ont des salaires et traitements de 7 000\$ ou moins. Ce sont probablement des personnes qui ne sont pas tenus par la loi de compléter la « Déclaration de revenus et de prestations » du fichier T1. Toujours parmi ces 1,3 million de NAS, 250 000

⁴ Des versions révisées existent également pour les fichiers T4, cependant ces fichiers montrent peu de changements par rapport aux versions finales reçues par SC.

(20 %) se trouvent sur le fichier T1 révisé. Il s'agit probablement des particuliers qui remettent leur déclaration de revenus trop tard pour que l'ARC puisse les inclure sur la version finale remise à Statistique Canada.

Finalement, on retrouve 600 000 NAS sur le fichier T1 qui ne sont pas sur le fichier T4. Cependant, 500 000 de ces personnes ont des salaires et traitements de 1 \$. On suppose que ce sont des cas en suspens pour l'ARC.

6. CONCLUSION

Divers aspects de la qualité des contrôles démographiques et du nombre de personnes par classe de revenu utilisés à l'étape du calage dans les enquêtes sur le revenu, les dépenses et la richesse des ménages ont été évalués. Les résultats ont été présentés en fonction des contrôles définis pour l'EDM.

On a évalué la qualité des contrôles démographiques dérivés à partir du Recensement de 1996 en les comparant à ceux produits à partir du Recensement de 2001. Les différences observées pour les contrôles au niveau des personnes sont presque toujours inférieures à 5 % et ce même plusieurs années après le Recensement de 1996. La qualité des contrôles au niveau des ménages selon la taille est un peu moindre lorsqu'on s'éloigne de ce recensement mais pour la majorité de ces contrôles, on observe aussi des différences inférieures à 5 %. Par contre, la qualité des contrôles au niveau du type de ménage, plus particulièrement les contrôles du nombre de ménages monoparentaux, est définitivement moins bonne. On observe des différences supérieures à 10 % dans quatre des dix provinces incluant les deux plus grandes provinces, le Québec et l'Ontario.

On a évalué également la qualité des deux fichiers qu'on peut utiliser pour dériver les contrôles selon le revenu soit le fichier fiscal de la rémunération annuelle payée par les employeurs à chaque salarié (T4) et le fichier de déclarations de revenus et de prestations des particuliers (T1). Dans environ 90 % des cas, on y retrouve les mêmes salariés et des valeurs de salaires presque égales. Quant aux écarts entre ces deux fichiers, ils s'expliquent principalement par un problème de sous-couverture important du fichier T1. Ce problème est attribuable surtout aux individus ayant un salaire inférieur à 7 000 \$ et aux particuliers qui remettent leur déclaration de revenus trop tard pour que l'ARC puisse les inclure sur la version finale remise à Statistique Canada. On en conclut que le fichier T4 est plus approprié pour dériver ces contrôles.

Cette évaluation de la qualité des contrôles a été un des éléments importants considérés lors des dernières modifications apportées aux stratégies de calage de l'EDM (Lessard, 2005) et de l'Enquête sur la dynamique du travail et du revenu (LaRoche, 2005).

REMERCIEMENTS

Les auteurs tiennent à remercier Sylvie LaRoche, Michel Latouche, Mylène Lavigne et Jenny Lynch pour leurs précieux commentaires.

RÉFÉRENCES

Arsenault, S., Gaudet, J., Nadeau, C., Tremblay, J. (2001). *Introduction of a New Calibration Strategy for the Survey of Household Spending*, 2001 Proceedings - American Statistical Association.

Bender, R. (1992). *Population Estimation Programme of Labour and Household Surveys Analysis Division : A Methodological Documentation*, document interne de Statistique Canada.

Deville, J.C., Särndall, C.E. (1992). *Calibration Estimators in Survey Sampling*, Journal of the American Statistical Association, Volume 87, pages 376-382.

LaRoche, S. (2005). *Stratégie de calage de l'Enquête canadienne sur la dynamique du travail et du revenu*, Recueil de la Section des méthodes d'enquête, Société Statistique du Canada.

Lessard, S. (2005). *Révision de la stratégie de calage de l'Enquête sur les dépenses des ménages*, document interne de Statistique Canada.

Schembari, P. (2001). *Estimations des ménages privés et des entités économiques – Rapport technique*, document interne de Statistique Canada.

Statistique Canada (2003). *Méthodes d'estimation de la population et des familles à Statistique Canada*, No. 91-528-XIF au catalogue.

Tremblay, J. (2005). *Aperçu de la stratégie de calage harmonisée des statistiques du revenu de Statistique Canada*, Recueil de la Section des méthodes d'enquête, Société Statistique du Canada.

Tremblay, J., Nadeau, C., Auger, S., LaRoche, S., Latouche, M. (2003). *Developments on the Harmonised Calibration of Income Statistics Project*, document interne de Statistique Canada.