

APERÇU DE LA STRATÉGIE DE CALAGE HARMONISÉE DES STATISTIQUES DU REVENU DE STATISTIQUE CANADA

Johanne Tremblay¹

RÉSUMÉ

Les stratégies de calage des enquêtes sur le revenu, les dépenses et la richesse ont été évaluées et harmonisées dans le but d'améliorer la comparabilité des estimations de revenu entre les enquêtes de Statistique Canada ainsi qu'avec les données fiscales et le Système de comptabilité nationale. Des contrôles selon le revenu provenant de sources administratives ainsi que de nouveaux types de contrôles démographiques ont été ajoutés à l'étape de calage de ces enquêtes. Un aperçu des études effectuées et des problèmes rencontrés lors du développement et de l'harmonisation des stratégies de calage est présenté dans cet article.

MOTS CLÉS : Calage; calage selon le revenu; contrôles démographiques; données fiscales.

ABSTRACT

The income strategies of surveys on income, expenditure and wealth have been evaluated and harmonized in order to improve the comparability of income estimates across Statistics Canada surveys as well as with tax data and National Accounts. Income controls from administrative sources and new types of demographic controls were added to the calibration step of these surveys. An overview of the studies carried out and the problems encountered during the development and the harmonization of the calibration strategies is presented in this paper.

KEY WORDS : Calibration, Demographic controls, Income Calibration, Tax data.

1. INTRODUCTION

Il existe plusieurs sources de données sur le revenu de la population canadienne. Cette information est collectée sur une base annuelle à partir de certaines enquêtes de Statistique Canada. L'Enquête sur la dynamique du travail et du revenu (EDTR) est la principale source de données sur le revenu des ménages et des familles. L'Enquête sur les dépenses des ménages (EDM) ainsi que l'Enquête périodique sur la sécurité financière (ESF) collectent également de l'information détaillée sur le revenu mais principalement dans le but d'analyser les dépenses et la richesse des ménages en fonction de leur revenu. Le Système de comptabilité nationale (SCN) produit aussi des estimations de revenu et des comptes peuvent être dérivés des données administratives de l'Agence du revenu du Canada (ARC).

Des comparaisons entre les estimations provenant de ces différentes sources, effectuées à la fin des années 1990, ont révélé des différences importantes, d'ampleur plus grandes que celles observées lors des années antérieures. Ces résultats ont soulevé l'importance d'améliorer et d'harmoniser les données de revenu des enquêtes ainsi que leur comparabilité aux données externes et sont à l'origine du projet d'harmonisation du calage des statistiques du revenu (Webber et coll., 2000).

Le calage consiste à ajuster les poids de sondage des enquêtes pour obtenir des estimations qui concordent avec certains totaux provenant de sources externes considérées plus fiables. L'emphase a donc été mise sur l'harmonisation et l'amélioration des stratégies de calage des enquêtes pour accroître la comparabilité des données sur le revenu. Une première

¹ 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 : Johanne.Tremblay@statcan.ca.

composante du projet visait à harmoniser l'utilisation des estimations démographiques dans ces enquêtes. La seconde consistait à évaluer si l'ajout d'information sur le revenu provenant de fichiers de données fiscales permettait d'améliorer la qualité des estimations des enquêtes ainsi que la comparabilité avec les autres sources de données.

Cet article présente un aperçu des problèmes rencontrés ainsi que des résultats des études effectuées sur une période d'environ cinq ans dans le but de développer une stratégie de calage harmonisée. Les problèmes de comparabilité entre les différentes sources de données, qui sont à l'origine du projet, sont décrits dans la section 2. Les sections 3 et 4 portent respectivement sur l'harmonisation des contrôles démographiques et l'ajout de contrôles selon le revenu. Les phases de mise en œuvre sont présentées à la section 5 et les conclusions par rapport à la nouvelle stratégie sont résumées dans la section 6.

2. MISE EN CONTEXTE

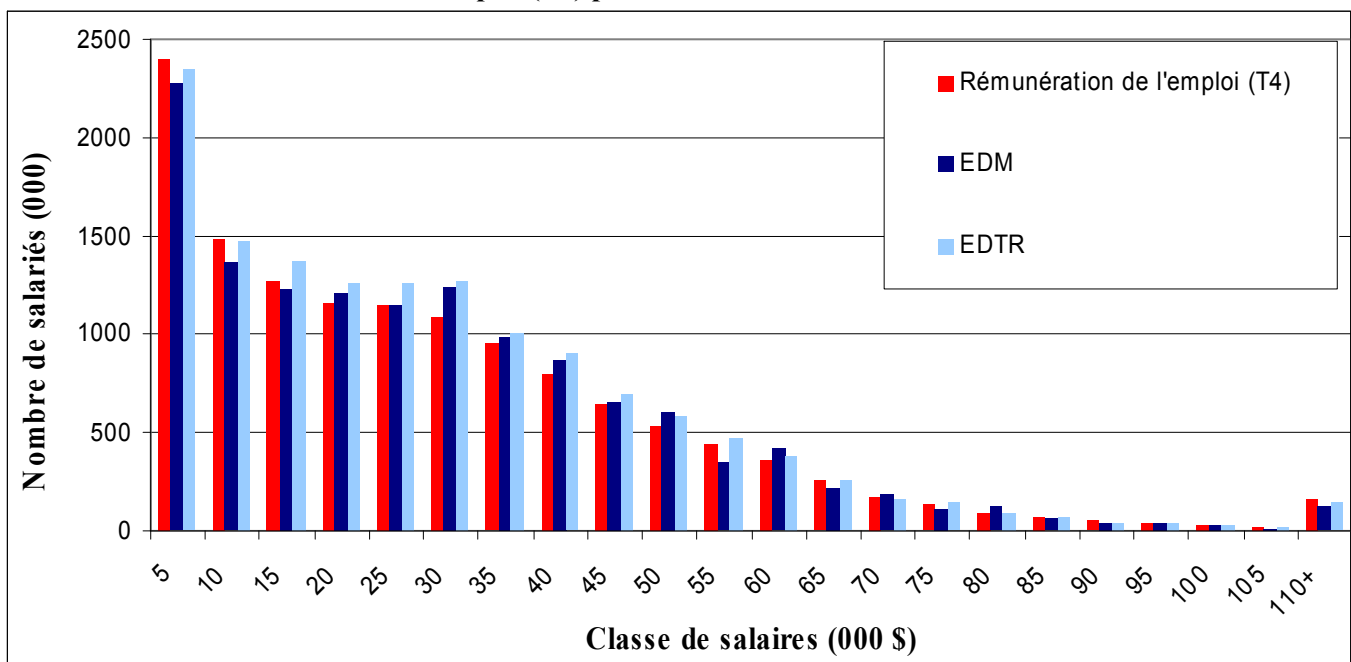
Les comparaisons effectuées à la fin des années 1990 montraient que les estimations du revenu total de la population canadienne provenant de l'EDTR étaient généralement comparables aux estimations du SCN. On constatait toutefois que l'enquête surestimait les revenus en salaires et traitements, qui représentent environ 65% du revenu total, alors qu'elle sous-estimait les revenus d'investissements et les transferts gouvernementaux. Les différences relatives entre les estimations de l'EDTR et du SCN de 1996 et 1997, pour les trois principales composantes du revenu, sont présentées dans le Tableau 1.

Tableau 1 : Différences relatives entre les estimations de l'EDTR et du SCN selon les trois principales composantes du revenu - Années 1996 et 1997

Principales composantes du revenu	Proportion du revenu total	1996	1997
Salaires et traitements	65 %	4,6 %	6,0 %
Revenu d'investissements	15 %	-44,2 %	-46,0 %
Transferts gouvernementaux	13 %	-14,2 %	-4,3 %
Revenu total		-2,3 %	0,8 %

Les estimations du total en salaires et traitements des enquêtes étaient aussi plus élevées que les totaux dérivés des données fiscales de l'ARC. De plus, en comparant la distribution des salariés par classe de salaires provenant des enquêtes à celle du fichier administratif contenant l'état de la rémunération annuelle payée par chaque employeur à chaque salarié (T4), on constatait que le nombre d'individus dans la classe moyenne était surestimé dans les enquêtes. Ce problème est illustré à partir des distributions des salariés de l'EDTR et l'EDM de 1997 présentées à la Figure 1.

Figure 1 : Comparaison des distributions du nombre de salariés par classe de salaires de l'EDTR, l'EDM et du fichier de rémunération de l'emploi (T4) pour l'année 1997



Il est difficile d'identifier les causes de ce problème de représentativité, mais l'effet de la non-réponse aux enquêtes pourrait en expliquer une partie. Les taux de réponse dans les enquêtes sur le revenu et les dépenses sont généralement entre 70 % et 80 % et on y observe un taux de réponse plus faible pour les ménages avec un revenu élevé. Malgré les ajustements effectués lors de la pondération des enquêtes pour réduire l'effet de la non-réponse sur les estimations, les résultats indiquent quand même une surestimation du nombre de ménages dans la classe moyenne. L'utilisation d'information auxiliaire supplémentaire sur le revenu a donc été considérée dans le but d'améliorer la qualité des estimations des enquêtes ainsi que la cohérence entre les différentes sources.

3. L'HARMONISATION DES CONTRÔLES DÉMOGRAPHIQUES

Les enquêtes sur le revenu, les dépenses et la richesse ajustaient déjà leurs poids de sondage afin que les estimations de l'enquête correspondent à certaines estimations post-censitaires de la population. Cet ajustement est effectué à partir d'une procédure d'estimation par régression qui assure un poids égal à tous les membres d'un même ménage (Lemaître et Dufour, 1987).

Avant l'harmonisation, le nombre de groupes pour les contrôles démographiques de population variait beaucoup d'une enquête à l'autre. Il était par exemple de 18 groupes par province, soit 9 catégories d'âge croisées en fonction du genre, pour l'EDTR alors qu'il n'était que de trois groupes d'âge par province pour l'EDM. Des problèmes de représentativité par catégorie d'âge, observés même après le calage, ont justifié l'augmentation du nombre de groupes pour l'EDM (Arsenault et coll., 2001). Comme les tailles d'échantillon des enquêtes peuvent être assez différentes, le but de l'harmonisation n'était pas d'utiliser exactement le même nombre de groupes pour chacune des enquêtes. Toutefois, 22 groupes de base ont été créés et chacune des enquêtes utilisent maintenant ces groupes ou des regroupements de ces groupes.

De plus, avant l'harmonisation, des estimations de nombre de ménages étaient utilisées seulement pour le calage de l'EDM. Puisqu'il n'existait aucune estimation post-censitaire au niveau du nombre de ménages, des contrôles étaient dérivés spécifiquement pour cette enquête à partir des données de l'Enquête sur la population active et du plus récent recensement. Dans le but d'améliorer la qualité de ces contrôles et d'en étendre l'utilisation à l'ensemble des enquêtes sur le revenu, les dépenses et la richesse, une méthodologie pour dériver des estimations du nombre de ménages et du nombre de familles économiques en fonction de la taille (1, 2 et 3 personnes ou plus) a été développée (Schembari, 2001).

Bien qu'il était prévu que l'EDTR utilise les trois catégories de taille pour les ménages et pour les familles économiques, la stratégie retenue consiste à se restreindre à deux groupes, soit celui composé d'une seule personne et celui composé de deux personnes, à la fois pour les contrôles au niveau des ménages et au niveau des familles (LaRoche, 2005). Ce changement découle de problèmes de représentativité des familles monoparentales lorsque la première approche était considérée.

À l'EDM, le calage tient compte des trois catégories de taille de ménages mais le concept de familles économiques n'est pas applicable à cette enquête. On observe quand même une sous-estimation importante des ménages monoparentaux (Arsenault et coll., 2001). La sous-représentativité des adultes est beaucoup plus importante que celle des enfants dans cette enquête, donc globalement, les poids des adultes devraient être augmentés de façon plus importante que ceux des enfants. Comme tous les membres d'un ménage doivent avoir le même poids final, le poids de sondage des ménages avec un plus grand ratio du nombre d'enfants sur le nombre d'adultes est augmenté de façon beaucoup moins importante que les autres, ce qui mène à une sous représentativité des ménages monoparentaux.

4. LE CALAGE SELON LE REVENU

Les données fiscales sur le revenu n'avaient jamais été utilisées dans le calage d'enquêtes auprès des ménages de Statistique Canada. Il fallait donc, en premier lieu, identifier une source de données fiscales qui contienne de l'information sur les composantes de revenu dont les concepts étaient similaires à ceux des enquêtes. On devait également s'assurer que ce fichier ait une bonne couverture de la population cible. La qualité de l'information se trouvant sur les fichiers en fonction des besoins pour le calage et la disponibilité du fichier en fonction des délais de diffusion des enquêtes étaient aussi des éléments clés. Ces divers aspects sont traités dans cette section, suivis de quelques résultats sur l'impact de l'utilisation des données fiscales sur les estimations des enquêtes.

Sources de données

La principale source de données administratives sur les différentes composantes du revenu des individus est le fichier des déclarations fiscales des particuliers (T1) de l'ARC. Ce fichier contient la plupart des composantes du revenu incluses dans les enquêtes quoiqu'il existe de légères différences entre les concepts des données fiscales et ceux des enquêtes pour certaines variables. Par contre, ce fichier ne couvre pas nécessairement l'ensemble des individus qui ont un revenu puisque la loi n'exige pas que chaque individu remplisse une déclaration fiscale. C'est le cas par exemple des individus avec un revenu inférieur à la déduction de base ou encore des époux ou épouses déclarés par leur conjoint.

L'ARC produit aussi un fichier composé des déclarations de rémunération payée par chaque employeur pour chacun de ses employés (T4). Ce fichier permet d'obtenir le revenu en salaires et traitements des salariés. D'un point de vue conceptuel, le fichier T4 devrait avoir une très bonne couverture des salariés puisque la loi exige que chaque employeur fournisse cette déclaration. Par contre, ce fichier ne contient qu'une des composantes du revenu, les salaires et traitements, et ne permet donc pas de faire des ajustements sur les autres composantes.

La composante salaires et traitements du fichier T4 a été retenue pour effectuer les premières analyses sur le calage selon le revenu. La similarité du concept avec les enquêtes ainsi que la possibilité d'obtenir des données administratives d'une source qui semblait avoir une meilleure couverture ont été les principaux facteurs qui ont mené à cette décision.

Approches de calage

Puisqu'on cherche à améliorer à la fois les distributions et les estimations agrégées, deux approches peuvent être envisagées pour tenir compte du salaire lors du calage. On peut se servir des effectifs de salariés par classe de salaires ou encore du total en salaires en fonction de ces classes. L'approche retenue a été de caler sur les effectifs puisqu'ils sont généralement moins affectés par les erreurs des fichiers administratifs et le calage sur les effectifs a l'avantage d'être plus stable.

Deux options ont été considérées pour définir les classes de salaires. La première visait à déterminer des bornes spécifiques à chaque enquête en analysant les intervalles de sous-estimation et de surestimation de la distribution de salariés de l'enquête par rapport à celle du fichier T4. La seconde consistait à fixer les bornes à certains percentiles, à calculer leurs valeurs à partir du fichier T4 et à les appliquer à chacune des enquêtes. Les études ont démontré qu'il y avait peu de différences entre les deux approches (Webber et coll., 2000). L'option des percentiles a donc été retenue puisqu'elle est opérationnellement plus simple et conforme aux objectifs d'harmonisation. Au départ le nombre de classes avait été fixé à six par province, les bornes correspondant aux 25^{ième}, 50^{ième}, 65^{ième}, 75^{ième} et 95^{ième} percentiles. Par la suite, une classe supplémentaire, définie à partir du 10^{ième} percentile, a été ajoutée à la stratégie de calage pour l'EDTR seulement. Pour la classe de salaires la plus élevée, le 95^{ième} percentile est remplacé par le 98^{ième} ou le 99^{ième} pour mieux corriger le problème de sous-représentativité des ménages à revenu élevé lorsque la taille de l'échantillon le permet.

Qualité des fichiers T4

L'utilisation du fichier T4 étant beaucoup plus restreinte que celle du fichier T1 à Statistique Canada, plusieurs problèmes ont été identifiés. On y retrouvait par exemple des numéros d'assurance social (NAS) invalides, des doubles ainsi qu'une grande proportion de codes de province de résidence manquants. De plus, certains employeurs utilisent seulement un NAS pour fournir l'information sur leurs employés causant une sous-estimation du nombre de salariés dans les classes de plus faibles revenus. Une procédure de nettoyage a donc été développée pour résoudre ces problèmes.

Certaines comparaisons, indiquant des différences de 6 à 7% du nombre de salariés entre les fichiers T1 et T4, ont aussi soulevé l'importance d'effectuer une étude plus approfondie sur la qualité de ces deux fichiers en fonction des besoins des enquêtes pour vérifier si le choix du fichier T4 était vraiment le plus approprié. Les résultats de cette étude, présentés dans Auger et Tremblay (2005), confirment que l'utilisation du fichier T4 est préférable pour dériver les contrôles du nombre de salariés, les écarts entre les deux fichiers s'expliquant principalement par un problème de sous-couverture du fichier T1.

Disponibilité des fichiers T4

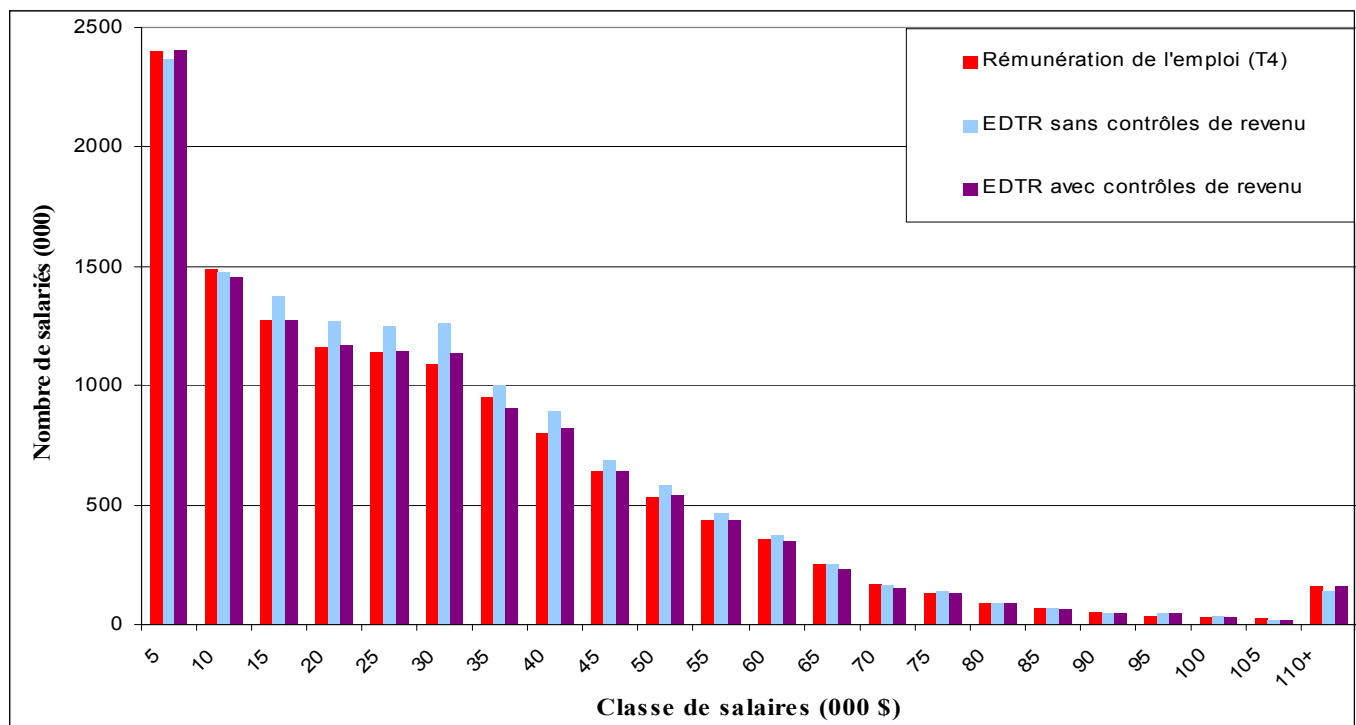
Le fichier T4 est reçu à Statistique Canada environ 13 mois après la fin de la période de référence, il n'est donc pas disponible au moment nécessaire à la pondération des enquêtes. Des projections du nombre de salariés pour chacune des classes sont donc produites à partir du fichier de l'année précédente en appliquant des facteurs d'ajustement démographique ou économique. Ce modèle utilise des données sur les salaires hebdomadaires provenant de l'Enquête sur la population active pour estimer le changement d'une année à l'autre pour certaines classes de salaires.

Une étude a été effectuée dans le but de comparer les contrôles prédits par le modèle aux contrôles dérivés à partir du fichier T4 de l'année de référence. Les résultats montrent que l'erreur attribuable au modèle est généralement faible pour les plus petites classes de salaires mais elle devient un peu plus importante pour la classe de salaires la plus élevée. Lorsque comparée à l'erreur de l'EDM, soit l'écart entre les estimations de cette enquête et les contrôles dérivés à partir du fichier T4, on constate que l'erreur du modèle est inférieure à l'erreur de l'enquête pour les six catégories de salaires (Tremblay et coll., 2003). En particulier, si on exclut la classe de salaires la plus élevée, on observe que l'erreur du modèle est toujours inférieure à 4% alors que l'erreur de l'enquête varie de 6% à 10% selon les classes. On en conclut donc qu'il est avantageux d'effectuer le calage selon le nombre de salariés même si on doit utiliser des contrôles prédits.

Impact du calage selon le revenu

Le calage en fonction du nombre d'individus par classe de salaires permet une bonne amélioration du problème de surestimation de la classe moyenne des salariés tel qu'illustré à partir des données de l'EDTR à la Figure 2. Des résultats similaires sont observés à partir de l'EDM (Arsenault et coll., 2001).

Figure 2 : Comparaison des distributions des salariés de l'EDTR avant et après l'ajout de contrôles du nombre de salariés - Année 1997



Le calage selon le nombre de salariés permet également de réduire les écarts entre les estimations des enquêtes et le SCN pour le total du revenu en salaires et traitements de la population, tel qu'illustré dans le Tableau 2 à partir des données de l'EDTR de 1997. On observe aussi une réduction de la sous-estimation des transferts gouvernementaux. Par contre, le calage selon le nombre de salariés a peu d'impact sur la sous-estimation des revenus d'investissements. On observe finalement un léger accroissement de l'écart entre les estimations du revenu total de l'EDTR et du SCN.

Tableau 2 : Différences relatives entre les estimations de l'EDTR et du SCN avant et après l'ajout de contrôles du nombre de salariés par classe, selon les trois principales composantes du revenu - Année 1997

Principales composantes du revenu	Proportion du revenu total	Sans les contrôles du nombre de salariés	Avec les contrôles du nombre de salariés
Salaires et traitements	65 %	6,0 %	-2,0 %
Revenu d'investissements	15 %	-46,0 %	-45,0 %
Transferts gouvernementaux	13 %	-4,3 %	-1,0 %
Revenu total		0,8 %	-1,8 %

En plus de contribuer à réduire les problèmes de représentativité des enquêtes, l'ajout d'une composante de revenu a permis d'améliorer la précision des estimations de l'EDM. On a en effet observé des réductions de coefficients de variation (CV) pour environ 75% des estimations provinciales des principales composantes de revenu et de dépenses (Arsenault et coll., 2005). Pour l'EDTR, les gains en précision sont beaucoup plus mitigés. Des comparaisons effectuées sur un très grand nombre d'estimations de l'enquête, considérant des composantes détaillées et plusieurs domaines, montrent une proportion similaire de réduction et d'augmentation des CV (Latouche, 2005).

5. LA MISE EN ŒUVRE

Les différentes études qui ont mené au développement des stratégies de calage actuelles des enquêtes sur le revenu, les dépenses et la richesse ont été réalisées sur une période d'environ six ans et la mise en œuvre a été effectuée en deux phases. Il fallait en effet considérer que des modifications à la stratégie de calage peuvent avoir un impact sur la comparabilité historique des estimations d'une enquête. Les changements méthodologiques importants ne peuvent donc être effectués que lors des révisions historiques des enquêtes et la nouvelle stratégie doit alors être appliquée rétroactivement. Ces révisions sont effectuées à tous les cinq ans lorsque de nouvelles séries d'estimations démographiques deviennent disponibles suite à un recensement.

Une première phase de mise en œuvre a eu lieu lors de la révision historique qui a suivi le Recensement de 1996. Les trois enquêtes ont alors modifié leur stratégie par rapport aux contrôles démographiques afin d'adopter l'approche décrite à la section 3. L'EDM a été la première enquête à effectuer sa révision historique à peu près au même moment où les problèmes de sous-représentativité des ménages monoparentaux ont été détectés. Pour pallier ce problème, des contrôles du nombre de ménages monoparentaux et du nombre de couples avec enfants ont été ajoutés à cette enquête.

Lors de cette première phase, les contrôles selon le salaire ont été utilisés seulement dans la stratégie de calage de l'EDM. En effet, pour l'EDTR, de grandes différences engendrées par l'utilisation de ces contrôles ont été observées sur les estimations de certaines variables importantes, notamment le nombre de personnes ou de familles sous le seuil de faible revenu. Des analyses supplémentaires ainsi que des consultations avec des experts ont donc été jugées nécessaires afin de mieux comprendre l'impact de cette nouvelle approche.

Les analyses effectuées par la suite sur l'EDTR ont permis d'élaborer une stratégie de calage incluant les contrôles selon le salaire (LaRoche, 2005). La mise en œuvre de cette stratégie a eu lieu en 2005 lors de la révision historique de l'EDTR qui a suivi le Recensement de 2001. Pour l'EDM, des analyses plus approfondies concernant l'impact des problèmes de représentativité des ménages monoparentaux (Lessard, 2005) ainsi que la qualité de ces estimations démographiques (Auger et Tremblay, 2005) ont mené à la décision d'exclure ces contrôles de la stratégie de calage. La mise en œuvre de cette approche a eu lieu dans le cadre de l'EDM 2004 et elle a aussi été appliquée à la révision historique suivant le Recensement 2001. Ces données d'enquêtes seront diffusées au cours des prochains mois.

6. CONCLUSION

L'ajout de contrôles selon le revenu dans le calage des enquêtes sur le revenu, les dépenses et la richesse a permis de réduire le problème de sur-représentativité des ménages de la classe moyenne observé dans ces enquêtes. De nombreuses études ont été nécessaires au cours du développement de la stratégie pour s'assurer de la qualité des données administratives utilisées ainsi que pour résoudre les problèmes de disponibilité des données fiscales par rapport aux besoins des enquêtes.

Le calage selon le nombre de salariés permet d'améliorer considérablement la comparabilité entre les enquêtes et avec les sources externes pour le revenu provenant des salaires et traitements. Tel qu'espéré, on observe aussi des gains pour d'autres composantes de revenu, notamment les transferts gouvernementaux. Toutefois cette approche a peu d'impact sur la sous-estimation des revenus d'investissements et pourrait accroître les écarts pour le revenu total. Ces résultats s'expliquent probablement par le fait que l'option retenue pour le calage ne permet pas de réduire suffisamment le problème de sous-représentativité des ménages à revenu élevé qu'on observe généralement dans ces enquêtes.

L'harmonisation des différentes séries d'estimations démographiques dans le calage de ces enquêtes est aussi une composante importante à la comparabilité. Toutefois, les analyses d'impact montrent que l'utilisation combinée d'estimations démographiques à plusieurs niveaux (personne, famille économique, ménage) engendre des problèmes de représentativité pour certains groupes de ménages, par exemple une sous-estimation des familles ou des ménages monoparentaux. D'autres exemples sont présentés dans Latouche (2005).

Dans le but de comprendre et de pallier les effets indésirables du calage sur la représentativité de certains groupes de ménages, des études de simulation sur les effets de la pondération intégrée, qui assure un poids égal à tous les membres d'un même ménage, de même que des évaluations d'approches de calage moins restrictives ont été amorcées. Les résultats obtenus jusqu'à maintenant n'ont toutefois pas permis d'identifier une approche dont les impacts seraient moins importants.

REMERCIEMENTS

L'auteur tient à remercier Sylvain Perron, Christian Nadeau, Michel Latouche et Sylvie Auger pour leurs précieux commentaires.

RÉFÉRENCES

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

Auger, S., et Tremblay, J. (2005). *Évaluation de la qualité des estimations démographiques et des données fiscales utilisées dans le calage de différentes enquêtes à Statistique Canada*. Recueil de la Section des méthodes d'enquêtes, Société Statistique du Canada.

Latouche, M. (2005). *Leçons à tirer de l'utilisation de données administratives au stade de la pondération*. Recueil de la Section des méthodes d'enquêtes, Société Statistique du Canada.

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

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

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.

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

Webber, M., Latouche, M. et Rancourt É. (2000), *Harmonized Calibration of Income Statistics*. Document interne de Statistique Canada