Recueil du Symposium 2016 de Statistique Canada
Croissance de l'information statistique : défis et bénéfices

Estimation de la migration interne: Enjeux reliés à l'utilisation des données fiscales

Guylaine Dubreuil et Georgina House[1]

Résumé

La migration interne constitue l'une des composantes de l'accroissement démographique estimées à Statistique Canada. Elle est estimée en comparant l'adresse des individus au début et à la fin d'une période donnée. Les principales données exploitées sont celles de la Prestation fiscale canadienne pour enfants et celles du fichier T1 sur la famille. La qualité des adresses et la couverture de sous-populations plus mobiles jouent un rôle capital dans le calcul d'estimations de bonne qualité. L'objectif de cet article est de présenter les résultats d'évaluations reliées à ces aspects, profitant de l'accès à un plus grand nombre de sources de données fiscales à Statistique Canada.

Mots Clés : Estimations démographiques, migration interne, adresses, couverture, données fiscales.

1. Introduction

Le Programme des estimations démographiques (PED) de Statistique Canada a pour objectif de produire des estimations de la population et des composantes de l'accroissement démographique. La migration interne constitue l'une de ces composantes. Celle-ci est définie par un déplacement qui entraîne un changement du lieu habituel de résidence à l'échelle interprovinciale ou intraprovinciale. Pour l'estimer, le PED exploite les données de la Prestation fiscale canadienne pour enfants (PFCE) et celles du fichier T1 sur la famille (T1FF). La qualité des adresses de ces données fiscales joue un rôle capital dans l'estimation de la migration interne. De plus, une sous-couverture de certaines sous-populations plus mobiles occasionnée par ces données pourrait entraîner un biais dans les estimations.

L'accès à un plus grand nombre de sources de données fiscales à Statistique Canada (StatCan) permet de faire davantage d'évaluations et de vérifier certaines hypothèses. Cet article présente certaines évaluations des fichiers actuellement utilisés, en confrontation ou en complément à d'autres sources. Plus précisément, un survol de la migration interne sera présenté à la section 2. Par la suite, une première évaluation comparant l'actualité de l'adresse des fichiers de la PFCE à une autre source administrative sera présentée à la section 3, suivie d'une évaluation visant à améliorer la couverture de la population soumise au risque de migrer des fichiers T1FF à la section 4. Le tout se termine par une brève conclusion.

2. Survol de la migration interne

La migration interne est estimée à plus d'une reprise au cours d'une année donnée, dite année démographique, s'étalant du 1er juillet d'une année *Y* au 30 juin de l'année *Y*+1. Des estimations provisoires sont d'abord produites environ 3 mois après la fin de l'année de référence au moyen des données de la PFCE. L'année suivante, des estimations définitives sont calculées à partir des données du T1FF. Pour ces deux séries d'estimation, la migration se fait principalement en comparant les adresses des individus au début et à la fin d'une période. À cette fin, les deux fichiers

[1]Guylaine Dubreuil, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6 (guylaine.dubreuil@canada.ca); Georgina House, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6 (georgina.house@canada.ca).

dont la mise à jour des adresses correspond le mieux possible aux périodes évaluées sont appariés. La population commune aux deux périodes correspond alors à la population soumise au risque de migrer. Sans aller dans le détail, il importe de mentionner que des étapes d'ajustements telles qu'un ajustement de couverture, sont ensuite effectuées (Statistique Canada, 2016).

2.1 Estimations provisoires de la migration interne

Les estimations provisoires de la migration interprovinciale sont produites sur une base mensuelle, trimestrielle et annuelle[2] à partir des données de la PFCE. Ces fichiers sont fournis mensuellement par l'Agence du revenu du Canada (ARC). Ils englobent les bénéficiaires de la PFCE et leurs enfants âgés de 0 à 17 ans. Les adresses sont mises à jour sur une base régulière. À titre d'exemple, les fichiers de la PFCE de juillet de l'année *Y* et juillet de l'année *Y*+1 sont utilisés pour estimer la migration interprovinciale provisoire de l'année de migration du 1er juillet de l'année *Y* au 30 juin de l'année *Y*+1. Les fichiers de la PFCE couvrent environ 95 % des enfants au Canada. Une modélisation est effectuée pour estimer la migration de la population adulte qui n'est pas couverte par la PFCE. Des facteurs d'ajustement sont appliqués pour s'assurer d'une couverture complète et pour réduire le risque potentiel de biais. Cette approche permet de produire des estimations de migration de qualité à l'intérieur d'un échéancier restreint.

2.2 Estimations définitives de la migration interne

Vers la fin du mois de juin de l'année *Y*+2, le fichier T1FF de l'année *Y* est disponible pour l'estimation définitive de la migration interne. Ce fichier est créé annuellement à StatCan, à partir des données d'impôt sur les particuliers (T1) de l'année *Y*, lesquelles sont combinées à d'autres sources de données fiscales, dont les données de la PFCE, pour ajouter les enfants et constituer les familles. Le fichier résultant couvre approximativement 95 % de la population canadienne. Étant donné que les adresses, issues du fichier T1, sont mises à jour ultérieurement à l'année d'imposition *Y*, les fichiers T1FF de l'année *Y*-1 et *Y* sont utilisés pour estimer la migration interne de l'année démographique *Y* à *Y*+1. Les estimations définitives sont calculées à la fois pour la migration interprovinciale et intraprovinciale. La migration intraprovinciale est effectuée aux échelons géographiques des régions métropolitaines de recensement (RMR)[3] et des divisions de recensement (DR)[4]. Les estimations provisoires de la migration interprovinciale sont révisées en conséquence afin de profiter de la plus grande complétude des données du T1FF.

3. Actualité des adresses des fichiers de la PFCE pour les estimations provisoires

3.1 Adresse des fichiers de la PFCE

Les fichiers de la PFCE, disponibles mensuellement, sont reconnus pour refléter rapidement les mises à jour des adresses postales. Une mise à jour de la population couverte et des adresses est effectuée à l'ARC chaque mois, mais une mise à jour annuelle plus prononcée est effectuée pour le fichier du mois de juillet. Un fichier mensuel d'un mois *m* comprend les mises à jour effectuées entre le 15e jour du mois *m*-1 au 15e jour du mois *m*. L'ARC envoie le fichier à StatCan rapidement de sorte qu'il est prêt pour utilisation dès le début du mois *m*+1.

3.2 Adresse des fichiers Ident

Depuis 2013, une nouvelle source de données fiscales, également en provenance de l'ARC, a été rendue disponible à StatCan trimestriellement, soit les fichiers Ident. Ces fichiers englobent tous les déclarants fiscaux T1 qui ont fourni

[2] Les estimations provisoires mensuelles et trimestrielles ne sont estimées que pour la migration interprovinciale.
[3] Une RMR est un territoire formé d'une ou de plusieurs municipalités voisines les unes des autres qui sont situées autour d'un noyau. Une RMR doit avoir une population totale d'au moins 100 000 habitants et son noyau doit compter au moins 50 000 habitants.
[4] Une DR est un groupe de municipalités voisines les unes des autres qui sont réunies pour des besoins de planification régionale et de gestion de services communs (comme les services de police et d'ambulance). Ces groupes sont créés selon les lois en vigueur dans certaines provinces du Canada.

au moins une déclaration d'impôt depuis 1983. Pour chacun de ces déclarants, on retrouve les cinq adresses les plus récentes et les dates des changements d'adresse. Dénotons les trimestres de janvier à mars par *T*1, avril à juin par *T*2 et ainsi de suite. Le fichier Ident *T*1 de l'année *Y* est en fait rendu disponible à StatCan au début du mois d'avril de l'année *Y*+1, et chaque autre fichier trimestriel est rendu disponible trois mois plus tard. Ces fichiers se distinguent par l'ajout de nouveaux déclarants fiscaux pour l'année *Y* et la mise à jour des adresses suite aux déménagements qui ont pu survenir jusqu'à la date de création du fichier. À titre d'exemple, le fichier Ident *T*1 de l'année *Y* reflète tous les changements d'adresse perçus à l'ARC jusqu'en date du 1er avril *Y*+1. Chaque fichier Ident transmis à StatCan est rendu accessible pour utilisation le mois suivant la réception, soit après avoir traversé quelques étapes de vérification et de traitement.

3.3 Concordance des adresses des fichiers Ident et des fichiers de la PFCE

La venue des fichiers trimestriels Ident à StatCan permet d'évaluer en partie l'hypothèse que les fichiers de la PFCE représentent la source de données fiscales dont la mise à jour des adresses est la plus rapide au cours d'une année. Auparavant, seules des sources de données fiscales annuelles étaient disponibles, ce qui permettait d'évaluer cette hypothèse uniquement sur une base annuelle. Puisque les fichiers Ident et les fichiers de la PFCE proviennent tous les deux de l'ARC, le concept d'adresse est le même, soit l'adresse postale, et leur mise à jour est théoriquement centralisée. Le signalement d'un changement d'adresse à l'ARC est principalement observé par un déclarant qui informe directement l'ARC de son déménagement ou au moment où une nouvelle déclaration d'impôt est reçue à l'ARC avec une adresse qui diffère de l'adresse la plus récente qui était enregistrée. Les dates de changement d'adresse sur les fichiers Ident ne représentent donc pas parfaitement la date réelle où le déménagement s'est produit, mais davantage la date à laquelle l'ARC a perçu le changement d'adresse. En effet, plus de 40 % des mises à jour des adresses observées en 2014 sur les fichiers Ident 2013 auraient eu lieu en mars, avril et mai, ce qui ne correspond pas nécessairement à la période réelle où la plupart des déménagements ont lieu, mais davantage à la période de l'année au cours de laquelle les déclarants doivent remplir leur déclaration d'impôt.

Ceci étant dit, si l'on compare chaque fichier trimestriel Ident au fichier mensuel de la PFCE couvrant la fin du dernier mois de mises à jour du fichier Ident, on s'attend à observer une excellente concordance entre les adresses de la population commune aux deux sources. L'expérience a été menée sur les quatre fichiers trimestriels Ident 2013 en comparant les codes postaux des déclarants. Suivant les processus de mise à jour de part et d'autre des deux sources tels que présentés aux sections 3.1 et 3.2, la concordance optimale des adresses du Ident *T*1 2013 devrait être observée auprès du fichier de la PFCE du mois d'avril 2014 et une déduction similaire peut être faite pour les trimestres subséquents *T*2, *T*3 et *T*4, respectivement comparables aux fichiers de la PFCE des mois de juillet 2014, octobre 2014 et janvier 2015. Les codes postaux des déclarants communs des fichiers Ident 2013 ont été comparés aux fichiers mensuels de la PFCE couvrant quelques mois avant et après le mois déduit pour la PFCE pour vérifier la période de comparabilité optimale des adresses. L'hypothèse de concordance optimale entre les codes postaux a été vérifiée et elle survient aux mois prévus, avec une concordance moyenne de 97,5 %, atteignant 99,9 % pour le fichier Ident *T*2 2013 et le fichier de la PFCE de juillet 2014.

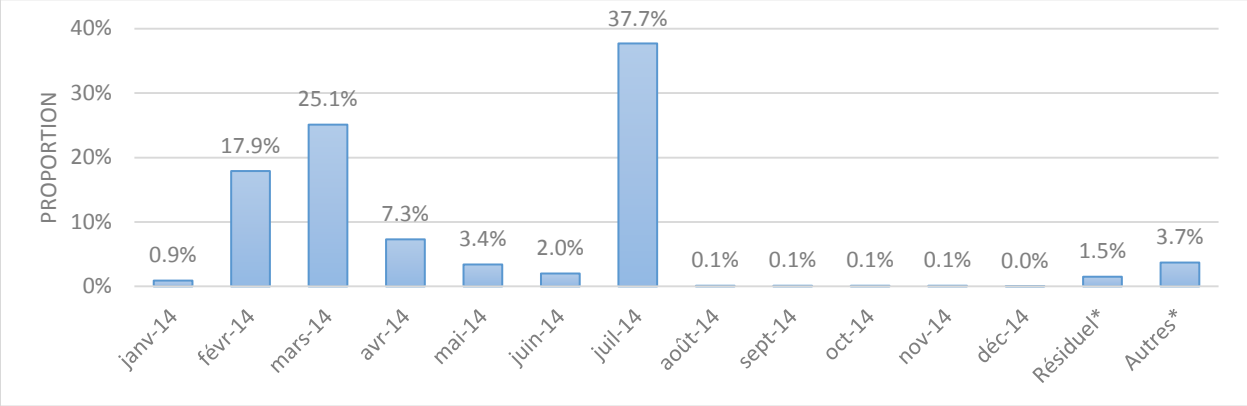
3.4 Actualité des adresses des fichiers Ident et des fichiers de la PFCE

La concordance des adresses, selon le code postal, est dans l'ensemble très bonne, mais la migration est un événement rare. Pour mieux connaître la qualité des adresses des migrants, des comparaisons d'adresses ont été faites pour la sous-population des déclarants communs ayant eu une date de changement d'adresse en 2014 sur le fichier Ident 2013. Il s'est avéré que la concordance entre les codes postaux des fichiers Ident et les fichiers de la PFCE n'était pas celle escomptée pour cette sous-population. Une évaluation plus approfondie a donc été faite en comparant le fichier de la PFCE du mois du changement d'adresse sur le fichier Ident 2013 et des mois subséquents pour vérifier à quel moment le code postal concordait sur les deux sources. L'évaluation a été faite mois par mois pour couvrir des changements d'adresse allant de janvier 2014 à décembre 2014. La figure 3.4-1 présente la distribution observée du mois où le code postal devient le même sur les fichiers mensuels de la PCFE pour les changements d'adresse de janvier 2014 identifiés sur les fichiers Ident 2013. Le premier mois, janvier, englobe les changements d'adresse qui peuvent aussi avoir été observés sur les fichiers de la PFCE avant cette période.

Le constat qui émane de cette évaluation est que la mise à jour des adresses sur les fichiers Ident et de la PFCE n'est pas synchronisée. Une portion des changements d'adresse est observée sur les fichiers de la PFCE au cours des mois

suivants la date du changement d'adresse sur le fichier Ident, mais une portion importante n'est observée que lors des mises à jour plus prononcées effectuées pour le fichier de juillet de la PFCE. Le rythme de mise à jour des adresses sur les fichiers de la PFCE semble donc dépendre du moment du changement d'adresse, bien que ces fichiers soient mensuels. L'évaluation complète au cours de l'année confirme le tout. Pour les estimations provisoires de la migration annuelle, l'impact est mineur, considérant que les fichiers de juillet *Y* à juillet *Y*+1 sont comparés et que ces fichiers profitent de la mise à jour des adresses escomptée. Cependant, pour la migration trimestrielle et mensuelle, les résultats de cette évaluation laissent croire que les estimations sont en fait décalées au cours des mois et des trimestres.

Figure 3.4-1
Distribution par mois de concordance du code postal des fichiers de la PFCE pour les changements d'adresse observés en janvier 2014 sur le fichier Ident 2013



* La catégorie « *Résiduel* » inclut les changements d'adresse qui ne sont pas encore observés sur les fichiers de la PFCE à la fin de la période évaluée et la catégorie « *Autres* » représente les cas où il y aurait incohérence du code postal avant ou après le changement observé.

4. Impact des changements de numéro d'identification sur les estimations définitives

4.1 Fichiers des liens entre les numéros d'identification d'un même individu

Le numéro d'identification généralement utilisé sur les fichiers de données fiscales est le numéro d'assurance sociale (NAS) pour les déclarants fiscaux et le numéro de dépendant (ND) pour les dépendants. Or, il s'avère qu'un individu peut changer de NAS au cours du temps. La raison principale survient lorsqu'un résident non permanent se voit assigner un NAS temporaire d'abord, pour ultérieurement recevoir un NAS permanent. Du côté des dépendants, un passage du ND au NAS se fait systématiquement lorsque qu'un enfant atteint un âge où il devient actif fiscalement, soit principalement entre 17 et 19 ans. Depuis peu, l'ARC fait parvenir deux types de fichiers à StatCan : le fichier NAS-NAS faisant le lien entre les NAS au cours du temps; et le fichier ND-NAS faisant le lien entre les ND et les NAS. StatCan se charge de faire des mises à jour, puis ces fichiers dérivés fiscaux sont rendus disponibles pour diverses utilisations. Ces fichiers sont historiques avec une mise à jour se produisant deux fois par année.

Rappelons que pour l'estimation définitive de la migration annuelle, la population soumise au risque de migrer est la population commune à deux fichiers T1FF consécutifs, ce qui correspond à la population qui est appariée par le NAS. Précisons que l'appariement est fait pour les déclarants fiscaux et que la migration des enfants est dérivée en fonction des déclarants auxquels ils sont associés. Sous cet angle, le fait qu'un individu puisse être identifié par plus d'un numéro d'identification peut avoir deux impacts sur les estimations définitives. D'une part, cette situation peut causer de la sur-couverture en raison de la présence de doublons. D'autre part, si des individus sont identifiés par des numéros d'identification qui diffèrent sur les fichiers T1FF consécutifs appariés, ces individus sont exclus de la population soumise au risque de migrer, entraînant une sous-couverture.

4.2 Exclusion des doublons sur les fichiers T1FF

La disponibilité des fichiers de liens permet de procéder à un nettoyage des fichiers T1FF avant qu'ils soient appariés. D'abord, les individus présents plus d'une fois sur chacun des deux TIFF à apparier sont identifiés en utilisant les mêmes versions des fichiers de liens. Le terme doublons est utilisé pour des raisons de simplicité même si certains individus sont présents plus de deux fois. En présence de doublons de NAS à NAS, on cherche à conserver un seul des enregistrements. Le choix de l'enregistrement doit être fait stratégiquement pour s'assurer qu'on optimise la possibilité de coupler l'individu entre les deux T1FF par la suite. Le choix logique est de conserver l'enregistrement avec le NAS qui a été octroyé le plus récemment. En présence de doublons de ND à NAS, on souhaite conserver l'enregistrement avec un NAS car la migration d'un individu possédant un ND n'est pas mesurée directement. Suivant ces procédures, l'exclusion des doublons a été effectuée sur les fichiers T1FF 2012 et 2013. Le tableau suivant quantifie l'exclusion des doublons par rapport à chacun de ces fichiers.

Tableau 4.2-1
Sommaire de l'exclusion des doublons des fichiers T1FF 2012 et 2013

T1FF	Population éligible initialement	Doublons NAS/NAS exclus	Doublons ND/NAS exclus	Cas plus complexes exclus	Total des exclusions
2012	33 533 211	18 449	150 106	1 694	170 249
2013	33 927 601	18 584	139 858	1 649	160 091

La majorité des doublons provient du passage du ND au NAS. Rappelons que le fichier T1FF est conçu à StatCan à partir du fichier T1 et que les enfants sont principalement ajoutés grâce à la PFCE. Sans l'utilisation des fichiers de liens entre les ND et NAS disponibles depuis peu, il n'est pas possible d'identifier simplement les individus qui sont couverts sur les deux types de fichiers sous des numéros d'identification différents, ce qui explique la présence d'un plus grand nombre de doublons pour cette catégorie. Le fichier des liens ND-NAS, de même que le fichier des liens NAS-NAS, deviennent donc des fichiers très utiles qui pourraient permettre d'exclure des doublons à même le T1FF.

4.3 Conversion des numéros d'identification des fichiers T1FF

En théorie, chaque individu n'apparaît qu'une seule fois sur chacun des fichiers T1FF à cette étape. Néanmoins, certains individus peuvent apparaître sous des numéros d'identification différents sur les fichiers T1FF à apparier. On peut faire l'hypothèse que cette situation surviendra tout particulièrement pour les dépendants qui deviennent nouvellement actifs fiscalement entre les deux années couvertes par les fichiers T1FF. Pour optimiser l'appariement, les fichiers de liens ND-NAS et NAS-NAS sont une fois de plus utilisés. Il s'agit ici d'identifier chaque individu pouvant apparaître sous plus d'un numéro d'identification au cours du temps et de s'assurer que sur un fichier T1FF donné, le numéro d'identification sous lequel il apparaît est le plus récent dans le cas de liens NAS-NAS ou qu'il apparaît sous son NAS dans le cas de liens ND-NAS. Si ce n'est pas le cas, une conversion du numéro d'identification est effectuée. Le même exercice est répété sur le fichier T1FF de l'année suivante en utilisant exactement les mêmes versions des fichiers de liens. À cela doivent s'ajouter quelques ajustements de variables permettant qu'un individu qui aurait été traité comme un dépendant puisse être traité comme un déclarant fiscal lors de l'estimation de la migration. La conversion des numéros d'identification des fichiers T1FF 2012 et 2013 ainsi appliquée a permis d'obtenir 382 101 liens additionnels dont plus de 86 % sont des cas d'individus passant d'un ND à un NAS.

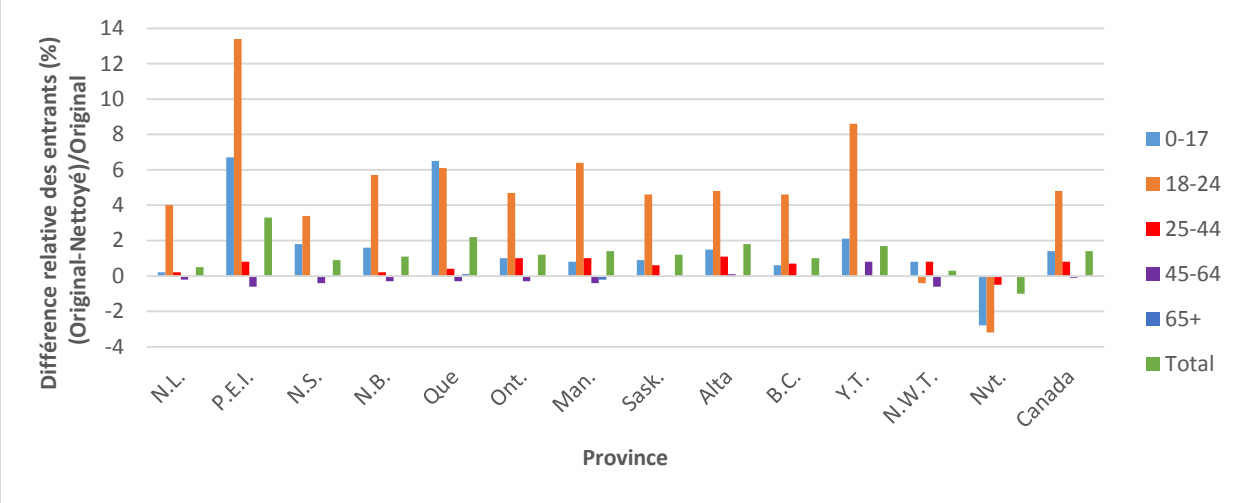
4.4 Impact du nettoyage des fichiers T1FF sur l'estimation de la migration interne

Le nettoyage des fichiers T1FF 2012 et 2013 a potentiellement un impact sur l'estimation définitive de la migration interne. Bien que divers ajustements soient appliqués lors de l'estimation de la migration, dont des ajustements pour la sous-couverture, il est préférable d'avoir la meilleure couverture possible. Une étude de couverture de la population soumise au risque de migrer résultant du couplage des fichiers T1FF 2012 et 2013 à chaque étape de nettoyage a donc été faite. La couverture est mesurée en comparant l'effectif soumis au risque de migrer aux estimations démographiques selon différents groupes d'âge. Dans l'ensemble, le taux de couverture global, tout âge confondu, passe de 90,9 % à 91,5 %. L'impact principal est observé auprès des groupes d'âge de 0 à 17 ans et de 18 à 24 ans, soient les âges associés au passage du ND au NAS. La couverture de la population âgée de 18 à 24 ans, nettement inférieure à celle des autres groupes d'âge, passe de 77 % à 75 % après l'exclusion des doublons, pour remonter à près

de 80 % après la conversion des numéros d'identification. Même s'il est modéré, ce gain de couverture est appréciable puisque la population âgée de 18 à 24 ans est plus mobile que l'ensemble de la population.

Enfin, une évaluation directe de l'impact du nettoyage des fichiers T1FF sur les estimations de migration s'impose. Pour ce faire, le processus complet d'estimation définitive de la migration interne a été appliqué en utilisant les fichiers finaux nettoyés des T1FF 2012 et 2013. L'impact a été mesuré par rapport aux estimations originales obtenues en production. La figure suivante résume l'impact sur l'estimation interprovinciale définitive par groupe d'âge pour les entrants interprovinciaux.

Figure 4.4-2
Estimation des entrants interprovinciaux à partir des fichiers originaux et nettoyés T1FF 2012 et 2013 par groupe d'âge



Une fois de plus, l'effet est plus prononcé pour la population âgée de 18 à 24 ans et, dans une moindre mesure, pour celle âgée de 0 à 17 ans. Les différences varient d'une province à l'autre et elles sont perceptibles au niveau national, indiquant une croissance du nombre d'entrants interprovinciaux de près de 5 % pour la population âgée de 18 à 24 ans, ce qui se traduit par une croissance de 1,4 % des entrants pour la population totale. Cela tend à démontrer que l'exclusion de sous-populations plus mobiles peut causer un biais dans les estimations de migration. Les individus qui passent d'un ND à un NAS entre deux années consécutives sont souvent plus mobiles, car ce passage peut être relié à un déménagement pour les études ou pour entrer sur le marché du travail. Ces individus sont actuellement exclus de la population soumise au risque de migrer au moment où ils sont susceptibles d'être migrants. Ils seront inclus au cours des années suivantes, lorsqu'ils apparaîtront sur deux fichiers T1FF consécutifs sous un même numéro d'identification. Cependant, la mobilité qu'ils ont peut-être eue lorsqu'ils étaient exclus ne sera jamais perçue, ce qui entraîne une certaine sous-estimation de la migration et justifie la conversion des ND au NAS. Les étapes de nettoyage des fichiers T1FF seront par conséquent intégrées au processus de production des estimations définitives de la migration interne dès la prochaine révision historique.

5. Conclusion

La plus grande disponibilité des fichiers administratifs est un atout certain pour de nombreux programmes à StatCan. Dans le cas du PED, les deux évaluations présentées dans cet article témoignent de la pertinence d'analyser les nouvelles sources à notre disponibilité. De nombreuses questions émergent alors pour bien comprendre ces nouvelles sources. Elles aident en même temps à vérifier des hypothèses ayant trait à des sources déjà utilisées ou elles peuvent permettent de les améliorer. Une bonne communication avec les agences qui fournissent les fichiers administratifs, par exemple l'ARC, est un élément essentiel. Il importe également de trouver des moyens efficaces de se tenir informé des nouvelles sources administratives disponibles à StatCan et des évaluations qui y sont faites. Des évaluations faites par un groupe d'utilisateurs peuvent potentiellement être profitables à un autre groupe. Cela étant dit, avec la venue

des nouvelles sources administratives, la méthodologie utilisée dans le cadre du PED, notamment la composante de migration, est un travail continu d'évaluation, d'adaptation et de développement.

Bibliographie

Statistique Canada (2016), *Méthodes d'estimation de la population et des familles à Statistique Canada, Division de la démographie, No 91-528-X au catalogue, 103p.*