Proceedings of Statistics Canada Symposium 2016
Growth in Statistical Information: Challenges and Benefits

Estimating internal migration: Issues related to using tax data

Guylaine Dubreuil and Georgina House[1]

Abstract

Internal migration is one of the components of population growth estimated at Statistics Canada. It is estimated by comparing individuals' addresses at the beginning and end of a given period. The Canada Child Tax Benefit and T1 Family File are the primary data sources used. Address quality and coverage of more mobile subpopulations are crucial to producing high-quality estimates. The purpose of this article is to present the results of evaluations of these elements using access to more tax data sources at Statistics Canada.

Key words: Population estimates, internal migration, addresses, coverage, tax data.

1. Introduction

The purpose of Statistics Canada's Population Estimates Program (PEP) is to produce estimates of the population and of the components of population growth. Internal migration, one of those components, is defined as an interprovincial or intraprovincial movement that leads to a change in usual place of residence. The PEP produces these estimates using data from the Canada Child Tax Benefit (CCTB) and the T1 Family File (T1FF). Therefore, the quality of addresses in the tax data plays a crucial role in estimating internal migration. Furthermore, undercoverage of certain more mobile subpopulations in these data may bias the estimates.

Access to more sources of tax data at Statistics Canada (StatCan) makes it possible to perform more evaluations and to test certain assumptions. This article presents evaluations of the files currently in use either by comparing or supplementing other sources. An overview of internal migration is presented in Section 2. Section 3 describes an initial evaluation that compares the currency of addresses in the CCTB files to another administrative source. Then, a description of an evaluation to improve the coverage of the population at risk of migrating in the T1FF files is given in Section 4. The article finishes with a brief conclusion.

2. Overview of internal migration

Internal migration is estimated a number of times in a given year, referred to as the demographic year, which runs from July 1 of year Y to June 30 of year Y+1. Preliminary estimates are first produced about three months after the end of the reference year using CCTB data. The following year, final estimates are derived from T1FF data. For both series of estimates, migration is primarily determined by comparing the addresses of individuals at the beginning and end of a period. As a result, the two files with address updates that correspond as closely as possible to the periods evaluated are linked. The population found in both periods corresponds to the population at risk of migrating. Without going into detail, it is important to mention that adjustments, such as coverage adjustment, are then applied (Statistics Canada 2016).

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

2.1 Preliminary internal migration estimates

Preliminary interprovincial migration estimates are produced monthly, quarterly and annually[2] using CCTB data. These files, which are provided monthly by the Canada Revenue Agency (CRA), cover CCTB recipients and their children aged 0 to 17 years. Addresses are updated regularly. For example, CCTB files from July of year *Y* and July of year *Y*+1 are used to estimate preliminary interprovincial migration from July 1 of year *Y* to June 30 of year *Y*+1. CCTB files cover approximately 95% of children in Canada. Modelling is used to estimate migration of the adult population not covered by the CCTB. Adjustment factors are applied to ensure complete coverage and reduce the potential risk of bias. This approach produces quality migration estimates in a short timeline.

2.2 Final internal migration estimates

Toward the end of June in year *Y*+2, the T1FF file for year *Y* is available to produce final internal migration estimates. This file is generated annually by StatCan from individual tax data (T1) from year *Y*, which are combined with other tax data sources, including CCTB data, to add children and construct families. The resulting file covers approximately 95% of the Canadian population. Given that addresses from the T1 file are updated after the *Y* taxation year, the T1FF files from year *Y*-1 and *Y* are used to estimate internal migration of the *Y* to *Y*+1 demographic year. Final estimates are derived for both interprovincial and intraprovincial migration. Intraprovincial migration is determined at the census metropolitan area (CMA)[3] and census division (CD) geographic levels.[4] Preliminary estimates of interprovincial migration are subsequently revised benefitting from the greater completeness of the T1FF data.

3. Timeliness of addresses from CCTB files for preliminary estimates

3.1 Address from CCTB files

CCTB files, available monthly, are recognized for quickly reflecting mailing address updates. The CRA updates the population covered and addresses each month with a more extensive annual update of the file in July. A monthly file of month *m* contains updates performed between the 15th day of month *m*-1 to the 15th day of month *m*. The CRA sends the file to Statistics Canada quickly so that it is ready for use at the start of month *m*+1.

3.2 Address from Ident files

Since 2013, a new source of tax data, also from the CRA, has been available to StatCan on a quarterly basis. This source is the Ident files. These files contain all T1 tax filers who have filed at least one tax return since 1983. The file provides the five most recent addresses and dates of address changes for each tax filer. We will denote quarters from January to March by *Q*1, April to June by *Q*2, and so on. The Ident *Q*1 file for year *Y* is actually made available to StatCan at the beginning of April of year *Y*+1, and every other quarterly file is available three months later. These files are distinguished by the addition of new tax filers for year *Y* and address updates following movements that may have occurred up to the date of the file's creation. For example, the Ident *Q*1 file for year *Y* reflects all address changes collected by the CRA up to April 1 *Y*+1. Each Ident file sent to StatCan is available for use the month after receipt, after undergoing a few verification and processing steps.

[2] Preliminary monthly and quarterly estimates are produced only for interprovincial migration.

[3] A CMA is an area consisting of one or more adjacent municipalities situated around a major urban core. A CMA must have a total population of at least 100,000, of which 50,000 or more must live in the core.

[4] A CD is a group of neighbouring municipalities joined together for the purposes of regional planning and managing common services (such as police or ambulance services). These groups are established under laws in effect in certain provinces of Canada.

3.3 Concordance of addresses from Ident files and CCTB files

The arrival of quarterly Ident files at StatCan makes it possible to partially evaluate the assumption that CCTB files offer the data source with the fastest update of addresses during a year. In the past, only annual sources of tax data were available, which meant this assumption could only be evaluated annually. Since Ident and CCTB files both come from the CRA, the address concept is the same, namely, the mailing address, and their updating is theoretically centralized. The CRA becomes aware of an address change mainly from the filer who informs it directly of a move or when the CRA receives a new tax return with a different address than the most recent one on record. Thus, the change of address dates in the Ident files do not correspond exactly to the actual date of the move but more to the date on which the CRA learned of the change of address. More than 40% of address updates noted in 2014 in the 2013 Ident files apparently occurred in March, April and May, which does not necessarily correspond to the actual time when the moves took place but rather to the time of year when filers must complete their tax returns.

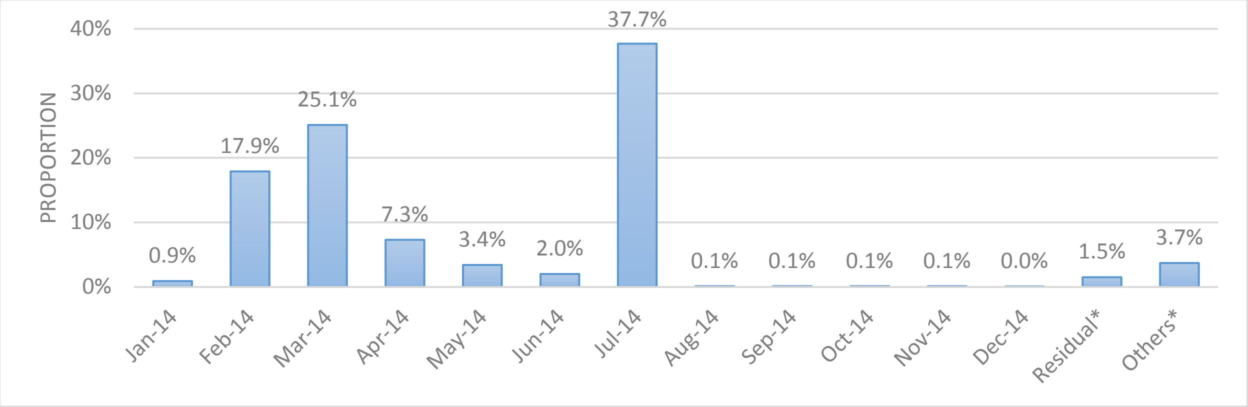
That being said, if each quarterly Ident file is compared to the monthly CCTB file covering the end of the last month of the Ident file update, the expectation is that there would be excellent agreement between the addresses of the population common to both sources. The four quarterly 2013 Ident files were tested by comparing the postal codes of filers. Based on the updating processes of the two sources described in sections 3.1 and 3.2, optimum concordance of the addresses in Ident $Q1$ 2013 should be observed with the CCTB file for April 2014. A similar assumption can be made regarding the subsequent $Q2$, $Q3$ and $Q4$ quarters being comparable respectively to the CCTB files for July 2014, October 2014 and January 2015. The postal codes of common filers from the 2013 Ident files were compared to the monthly CCTB files covering a few months before and after the deduced month for the CCTB in order to verify the optimum period of address comparability. The assumption of optimum concordance between postal codes was verified and did occur in the expected months, with an average agreement of 97.5% and reaching 99.9% for the Ident $Q2$ 2013 file and CCTB file of July 2014.

3.4 Timeliness of addresses from Ident files and CCTB files

Overall, there is very good agreement of addresses by postal code, but migration is a rare event. To have a better understanding of the quality of the addresses of migrants, addresses were compared for the subpopulation of common filers with a change of address date in 2014 on the Ident 2013 file. The concordance between postal codes of the Ident files and CCTB files was not as expected for this subpopulation. A more detailed evaluation was performed comparing the CCTB file for the month of the change of address on the Ident 2013 file and subsequent months to identify when the postal code was the same in both sources. The evaluation was carried out month by month to cover address changes from January 2014 to December 2014. Figure 3.4-1 shows the distribution of the month in which the postal code became the same in the monthly CCTB files for address changes from January 2014 identified in the Ident 2013 files. The first month, January, covers address changes that may also have been observed in CCTB files before that time.

This evaluation revealed that the update of addresses in the Ident and CCTB files is not synchronized. Some address changes are observed in the CCTB files in the months following the change of address date in the Ident file but a major portion of those changes are not observed until the more extensive updates are performed on the July CCTB file. The pace of address updates in the CCTB files seems to depend on the time of the address change, despite the files being monthly. The evaluation of the full year confirms this finding. For preliminary estimates of annual migration, the impact is minor given that the July Y to July Y+1 files are compared and these files contain the expected address updates. However, for quarterly and monthly migration, the results of this evaluation indicate that the estimates are actually shifted over months and quarters.

Figure 3.4-1
Distribution by month of correspondence of CCTB postal code for address changes observed in January 2014 in the Ident 2013 file



* The “Residual” category includes changes of address that are still not found in the CCTB files at the end of the period evaluated and the “Others” category represents cases where there was inconsistency in the postal code before or after the observed change.

4. Impact of changes in identification number on final estimates

4.1 Link files for identification numbers of a given individual

The identification number commonly used in tax data files is the social insurance number (SIN) for tax filers and the dependent identification number (DIN) for dependents. An individual’s SIN may change over time. The main reason for this is that a non-resident is first assigned a temporary SIN and later receives a permanent SIN. In terms of dependents, the shift from a DIN to a SIN occurs systematically when a child reaches an age when he or she becomes fiscally active, usually between the ages of 17 and 19. Recently, the CRA has been sending StatCan two types of files: the SIN-SIN file showing the link between SINs over time, and the DIN-SIN file showing the link between DINs and SINs. StatCan assumes responsibility for performing the updates and these derived tax files are then made available for various purposes. These files are historical with updates produced twice yearly.

It should be remembered that, for the final estimate of annual migration, the population at risk of migrating is the population common to two consecutive T1FF files, which corresponds to the population matched by the SIN. To clarify, tax filers are matched and the migration of children is derived based on the filers with which they are associated. On this basis, the fact that an individual may be identified by more than one identification number could impact the final estimates in two ways. First, this situation may lead to overcoverage due to duplicates. Second, if individuals are identified by different identification numbers in the matched consecutive T1FF files, these individuals are eliminated from the population at risk of migrating, leading to undercoverage.

4.2 Duplicate elimination in T1FF files

The availability of link files makes it possible to clean up the T1FF files before they are matched. Individuals present more than once in each of two T1FF files to be matched are identified using the same versions of the link files. The term “duplicate” is used for simplicity even though some individuals are present more than twice. In the case of SIN to SIN duplicates, the objective is to retain only one record. The record must be chosen strategically to optimize the possibility of linking the individual between the two T1FF files afterwards. The logical choice is to retain the record with the most recently assigned SIN. Where there are DIN to SIN duplicates, the goal is to retain the record with a SIN because the migration of an individual with a DIN is not measured directly. These procedures were used to eliminate duplicates in the 2012 and 2013 T1FF files. The following table shows how many duplicates were eliminated in each of these files.

Table 4.2-1
Summary of elimination of duplicates in the 2012 and 2013 T1FF files

T1FF	Population initially eligible	SIN/SIN duplicates eliminated	DIN/SIN duplicates eliminated	More complex cases eliminated	Total eliminations
2012	33,533,211	18,449	150,106	1,694	170,249
2013	33,927,601	18,584	139,858	1,649	160,091

Most duplicates are the result of the shift from a DIN to a SIN. It should be remembered that the T1FF file is created by StatCan from the T1 file and that children are mainly added through the CCTB. Without the DIN-SIN link files that have only recently become available, it is impossible to easily identify individuals covered by both types of files under different identification numbers, which explains the presence of a larger number of duplicates for that category. The DIN-SIN link file, and the SIN-SIN link file, are therefore very useful and may make it possible to eliminate duplicates from the T1FF.

4.3 Conversion of T1FF identification numbers

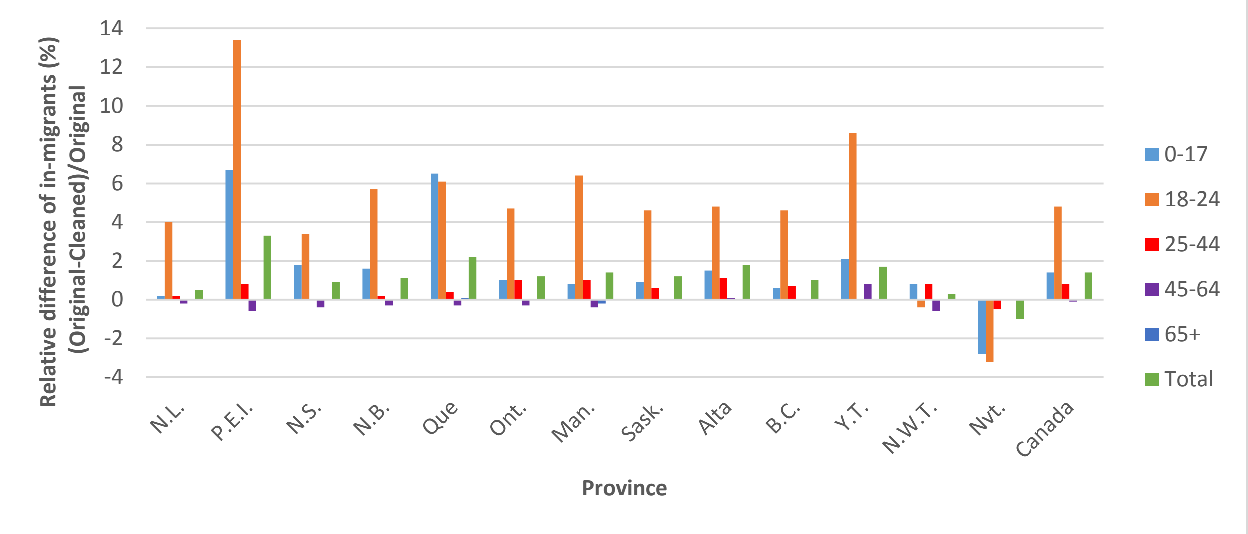
In theory, each individual appears only once in each of the T1FF files at this stage. However, some individuals may appear under different identification numbers in the T1FF files to be matched. It can be assumed that this situation will apply in particular to dependents that become newly fiscally active between the two years covered by the T1FF files. To optimize the match, the DIN-SIN and SIN-SIN link files are again used. The purpose here is to identify each individual who may appear under more than one identification number over time and ensure that, in a given T1FF, the identification number under which the individual appears is the most recent in the case of SIN-SIN links or that the individual appears under the SIN in the case of DIN-SIN links. If this is not the case, the identification number is converted. The same exercise is repeated in the T1FF of the following year using the exact same versions of the link files. A number of variable adjustments also need to be made so that an individual who may have been processed as a dependent can be processed as a tax filer when estimating migration. The conversion of the identification numbers in the 2012 and 2013 T1FF files in this manner produced 382,101 additional links, more than 86% of which were cases of individuals changing from a DIN to a SIN.

4.4 Impact of cleanup of the T1FF files on the internal migration estimate

The cleanup of the 2012 and 2013 T1FF files potentially impacts the final internal migration estimate. While various adjustments are applied when estimating migration, including adjustments for undercoverage, it is preferable to have the best possible coverage. For this reason, a study of the coverage was conducted. This study looked at the population at risk of migrating resulting from the linkage of the 2012 and 2013 T1FF files at each stage of the cleanup. Coverage was measured by comparing the population at risk of migrating to the population estimates for various age groups. Overall, the global coverage rate for all age groups rose from 90.9% to 91.5%. The main impact was observed with the 0-to-17 and the 18-to-24 age groups, which represent the ages associated with the change from the DIN to the SIN. Coverage of the 18-to-24 population, which is well below that of other age groups, fell from 77% to 75% after the elimination of duplicates, and then climbed again to almost 80% after conversion of the identification numbers. While modest, this coverage gain is appreciable since the 18-to-24 population is more mobile than the population as a whole.

Finally, it is important to evaluate the direct impact of the cleanup of the T1FF files on migration estimates. To this end, the complete process for the final estimate of internal migration was applied using the final cleaned 2012 and 2013 T1FF files. The impact was measured against the original estimates obtained in production. The following figure summarizes the impact on the final interprovincial estimate by age group for interprovincial in-migrants.

Figure 4.4-2
Interprovincial in-migrant estimates using the original and cleaned 2012 and 2013 T1FF files, by age group



Once again, the impact is more pronounced for the population aged 18 to 24 years and, to a lesser degree, that aged 0 to 17 years. The differences vary from province to province and are perceivable at the national level, indicating an increase in the number of interprovincial in-migrants of almost 5% for the 18-to-24 population. This translates into an increase in in-migrants of 1.4% for the total population. This result tends to show that the elimination of more mobile subpopulations may bias migration estimates. Individuals who change from a DIN to a SIN between two consecutive years are often more mobile because this change may be related to a move for education or to entering the workforce. At present, these individuals are eliminated from the population at risk of migrating at the time when they are likely to be migrants. They will be included in future years when they appear in two consecutive T1FF files under a single identification number. However, the mobility they may have experienced while eliminated will never be identified, resulting is a certain underestimation of migration and justifies the conversion of DINs to SINs. For this reason, the cleanup steps for the T1FF files will be integrated in the production process of the final internal migration estimates at the time of the next historical revision.

5. Conclusion

The greater availability of administrative files definitely benefits many StatCan programs. In the case of the PEP, the two evaluations presented in this article demonstrate the value of analysing the new sources available to us. They raise numerous questions in order to fully understand these new sources. They also help to verify assumptions related to existing sources or may contribute to their improvement. Good communication with the organizations providing administrative files, such as the CRA, is crucial. Effective means must also be found to keep abreast of new administrative sources available to StatCan and the evaluations of those sources. Evaluations conducted by one user group may potentially benefit another group. That being said, with the arrival of new administrative sources, PEP methodology, and the migration component in particular, requires ongoing evaluation, adaptation and development.

References

Statistics Canada (2016), Population and Family Estimation Methods at Statistics Canada, Demography Division, Catalogue No 91-528-X, 94 p.