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

Redesign of the longitudinal immigration database (IMDB)

Rose Evra[1]

Abstract

The Longitudinal Immigration Database (IMDB) combines the Immigrant Landing File (ILF) with annual tax files. This record linkage is performed using a tax filer database. The ILF includes all immigrants who have landed in Canada since 1980. In looking to enhance the IMDB, the possibility of adding temporary residents (TR) and immigrants who landed between 1952 and 1979 (PRE80) was studied. Adding this information would give a more complete picture of the immigrant population living in Canada. To integrate the TR and PRE80 files into the IMDB, record linkages between these two files and the tax filer database, were performed. This exercise was challenging in part due to the presence of duplicates in the files and conflicting links between the different record linkages.

Key Words: Record linkage, Immigration, IMDB, Landing file, Temporary Resident, Permanent Resident

1. Introduction

1.1 Description of the project

The Longitudinal Immigration Database (IMDB) combines Immigration, Refugees and Citizenship Canada's (IRCC) Immigrant Landing File (ILF) with annual tax files, the T1 Family Files (T1FF). This record linkage is performed using the Linkage Control File (LCF), a tax filer database that covers all tax filers in Canada since 1981. The ILF includes all immigrants who have landed in Canada since 1980.

In looking to enhance the IMDB, the possibility of adding temporary residents (TR) and immigrants who landed inclusively between 1952 and 1979 (PRE80) was studied. Extending the IMDB to include data on immigrants who landed before 1980 would give a broader picture of the immigrant population living in Canada. In addition by including TR data, the migration history of TRs who become permanent residents would be available; it would be possible to compare the socioeconomic integration of immigrants with pre-landing experience against those without any pre-landing experience.

To integrate the TR and PRE80 files into the IMDB, record linkages between these two files and the LCF were performed. In addition, an administrative link between the temporary resident file (TRF) and ILF was possible using a person identification number provided by IRCC. This exercise was challenging in part due to the presence of duplicates in the files and conflicts between the different record linkages. This article will go over the process and challenges faced to enhance the IMDB with these two files.

1.2 Description of the Longitudinal Immigration Database (IMDB)[2]

The IMDB is a comprehensive source of data on the economic behaviour of the tax-filing immigrant population in Canada. An immigrant refers to a person who is or has ever been a permanent resident (PR). This person has been

[1] Rose Evra, Statistics Canada, 100 Tunney's Pasture Driveway, Canada, K1A 0T6 (rose.evra@canada.ca).
[2] For more details on the IMDB see web page:
http://www23.statcan.gc.ca:81/imdb/p2SV.pl?Function=getSurvey&SDDS=5057&lang=fr&db=imdb&adm=8&dis=2&fpv=30000

granted the right to live in Canada permanently by immigration authorities. The IMDB allows the analysis of economic outcome of different categories of immigrants over time to assess the impact of immigrant characteristics at landing, such as education and knowledge of French or English, on their settlement outcomes.

The IMDB combines an administrative Landing File with annual T1 Family Files (T1FF) through exact matching record linkage techniques. The linkage is done using variables such as names, date of birth, gender and landing dates. The overall linkage rate of the ILF to the LCF is approximately 87% (Brennan, 2014).

Each year the IMDB is updated with a new cohort of landings and the taxation data for all landing cohorts covered by the IMDB. In each new tax year there are new entrants from all landing cohorts, not just the newly added cohort, who have filed (or are matched) for the first time.

The Immigrant Landing File (ILF) is a census of all immigrants who landed since 1980, it contains information on their characteristics when they became permanent residents (PR). As of 2013, the ILF has over 6.8 million records. Each year the size of the ILF increases by the number of new permanent residents (approximatively 250,000). The ILF has variables such as immigrant admission category, landing date and birth country. Each immigrant is assigned a person identification number, which remains unchanged over time, and a permit number. This file contains information on the PRs at time of landing, but pre-landing Canadian experience information (such as previously obtaining work or study permits as TR) is not collected. Deaths and emigrations are not identified, since no follow-up is made after landing. Consequently, a percentage of the records on the ILF are no longer part of the population living in Canada.

2. Enhancing the IMDB

2.1 Benefits of enhancing the IMDB

According to the 2011 National Household Survey (NHS), in 2011, it was estimated that nearly 70% of Canada’s immigrant population landed after 1980, therefore the IMDB does not cover the complete immigrant population of Canada. In order to have better coverage of Canada’s immigrant population, information on pre-1980 immigrants would be required. The NHS estimates that, in 2011, 98% of immigrants landed in Canada from 1952 onwards.

With information on TRs, it would be possible to make links between the PRF and TRF, which would broaden the analytical value of the IMDB. It would be possible, for example, to compare the socioeconomic integration of immigrants with pre-landing experience in Canada against those without any pre-landing experience.

2.2 Temporary Residents (TR)

Over 7.5 million people obtained a temporary permit to legally stay in Canada between 1980 and 2013. A database covering the TR population containing information such as entry date, permit information and a person identification number (PIN) was used for this project. This PIN should be the same present on the ILF for TRs who obtained landed immigrant status. In some cases, this PIN differ between the two files because the information was not collected or it was lost.

After undertaking a record linkage, it was estimated that the TRF included approximately 400,000 duplicates, mainly due to data quality issues in the late 1980s (Evra, 2015). To mitigate this loss of information, easily identifiable dummy person identification numbers were created. In order to identify the duplicates and remove them, a record linkage of the TR file with itself was done.

To evaluate the benefit of adding TR data to the IMDB, the file was linked to the linkage control file (LCF) which is a post-1981 tax filer database. An estimated 39% of the records on the file were linked. The low linkage rate is mainly because many TRs do not file taxes (Brennan, 2014). Some of them are in the country only for a short period of time. It was determined that 19.5% of the TRs on this file became PRs between 1980 and 2012 (Barayandema, 2015). It would be possible to add pre-landing information for 1.7 million records on the ILF, which represent one out of four people on the ILF as will be described in section 3.3.

2.3 Pre-1980 Permanent Residents (PRE80)

An administrative data file covering 4.1 million PRs who landed in Canada during the period of 1952 to 1979 was used for this project. Some relevant variables, such as the landing date and country of birth, is present. However, the quality of the information available for these immigrants is lower than for the ILF and varies depending on the landing year. For example, only the year of birth was available between 1962 and 1972, which lowered the linkage rate for people who landed during that period (Brennan, 2014). Some variables, such as the immigrant admission category, are not included. The file also shares some limitations with the ILF, such as a lack of post landing information regarding deaths and emigration. The proportion of deaths is expected to be higher in this file than the ILF due to the older population.

To evaluate the benefit of adding pre-1980 immigrants' data to the IMDB, the file was linked to the linkage control file. A proportion of 43% of the records on the file were linked to the LCF (Evra, 2015). Also, for tax year 2012, approximately 1 million pre-1980 immigrants were still filling taxes, and in the early 1980s the pre-1980 immigrants accounted for the vast majority of immigrant tax filers. The percent representation of pre 1980 landers goes from about 85 % in 1982 and decreases to about 19% in 2012 (Brennan, 2014). Adding this data to the IMDB would greatly improve the coverage of the immigrant population, especially in the 1980s.

3. Processing

3.1 Purpose

As a result of the feasibility studies, it was decided to enhance the IMDB by adding the temporary residents and pre-1980 immigrants' data to the database. This will allow the creation of a more comprehensive immigration database with more detailed immigration variables. For example, the original IMDB tax files contains the number of immigrants in a family, but these only counted post-1980 immigrants; this number will now include temporary residents and pre-1980 immigrants. It will also be possible to identify the spouse of an immigrant tax filer as a temporary resident or as a person who has obtained landed immigrant status prior to 1980.

The purpose of this project was to create a person file (figure 3.1-2) with all the connections between becoming a permanent resident pre- or post-1980 and having been a temporary resident (TR), to produce tax files from 1982 to 2013 for these immigrants and TRs, and to create a permit file for the TRs The permit file would include temporary resident information at the permit level (e.g.: effective date, type of permit) while the person file would give details like the number of temporary permits (e.g.: 2 work permits and 1 student permit). The difficulties related to this process are described in the following section (figure 3.1-1).

Figure 3.1-1
Processing Illustration

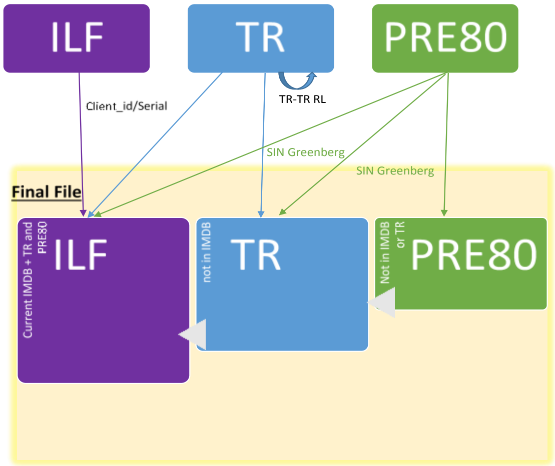
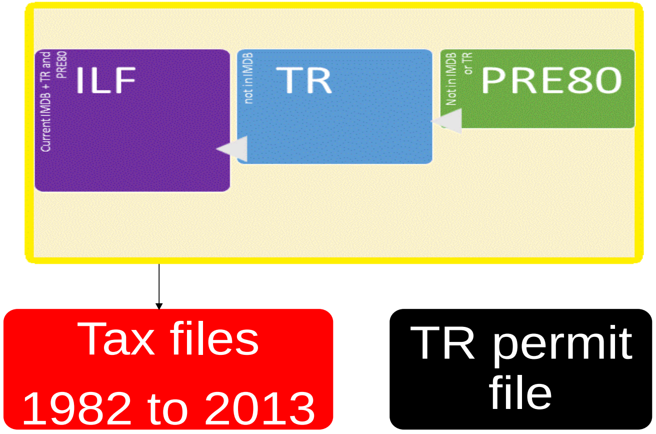


Figure 3.1-2
Final Product



3.2 Challenges

As mentioned earlier in this paper, the initial record linkage consisted of linking the ILF to the LCF (tax filer database) to identify the immigrant tax filers. Similarly, the TRF and PRE80F were linked to the LCF separately after the initial record linkage. The results of these record linkages were files containing linkage keys that connected the immigration files (ILF, TRF and PRE80F) to the tax files, with the Tax Identification Number (TIN) and the person identification number acting as the keys. The TRF was later linked to itself to identify duplicates. Additionally, the TRF was linked to the ILF using the person identification number (PIN) and landing permit number, which are common to the two files.

The main challenge was the need to combine the results of four record linkages to produce the person file. These record linkages were performed separately at different times, resulting in conflicting results. For example, in figure 3.2-1, the first person had two different PINs'. It was only possible to link the records using the TIN. The second person was only linked using the person identification number and the record linkage of the IFL to the LCF did not provide a link while the linkage of the TRF to the LCF did.

Figure 3.2-1
Fictive Examples of Record Linkage Results

Person	PIN ILF	PIN TRF	TIN ILF	TIN TR
1	9012345	9012346	12345678	12345678
2	4567890	4567890	Not linked	23456789
3	5678901	5678901	34567890	Not linked
4	6789012	6789012	78901234	89012345

These differences are the results of many factors. Linkage error explains part of the discrepancies. The quality of the linkage fields differed from a file to another. Also, the tools and record linkage methods evolve with time. Measurement error arose from data capture error and respondent error. For example, the date of birth might differ between the TRF and ILF, causing a link to different tax filers for a given person. Note that only a small proportion of the records resulted in conflicting results, each of the inconsistencies being applicable to less than 1% of the records included in the enhanced IMDB (Evra, 2015).

Some of the data variation is explained by the elapsed time between the collections of information included in different files. For example, tax identification number (TIN) can change over time. Many immigrants begin with a temporary TIN and later are assigned a permanent TIN. About 20% of permanent residents, 44% of temporary residents and 5% of pre-1980 permanent residents had multiple TINs assigned to them (Brennan, 2014). In some instances, immigrants change person identification number and cannot be found without record linkage using other variables (Figure 3.2-1: Person 1).

After the temporary resident record linkage to itself was done it was discovered that a few TR duplicates were linked to different tax filers (less than 1000 records; Evra 2015) according to the results of the other record linkages (TRF-LCF and ILF-LCF record linkages). The results from the ILF-LCF record linkage were considered to be of better quality and any conflicts that would result in a change of these results were disregarded.

When the ILF was linked to the TRF using the person identification numbers, some cases were linked to different tax identification number (TIN) as showcased by Person 4 in figure 3.2-1. This could be the result of linkage error or the person having multiple TINs. When this situation occurred, the results from the ILF-LCF record linkage were kept. To resolve the multiple TIN issue, the TIN database was used to identify all TINs related to a given person.

Some records with the same TIN had different person identification numbers. This could be due to record linkage error or a valid change in the person identification number.

Some of the sociodemographic variables differed between the ILF, TRF and PRE80F. For example, the birth country might differ. The information from the ILF was kept when available, followed by TRF and the PRE80 data. This decision was made based on the quality of the data contained in the different files.

3.3 Results

The final result of this project is a person file containing over 16.5 million records and tax files from 1982 to 2013 including data on people who became permanent residents (from 1952 to 2013) and temporary residents (from 1980 to 2013); see table 3.3-1. More than 8.9 million records were identified as belonging to people who filed taxes at least once from 1982 to 2013. According to the results of this project 1.7 million immigrants were temporary residents prior to becoming permanent residents; of these 94% were linked to the LCF.

For records that appear in the ILF, TRF, and PRE80F the linkage rate is 100% because the tax identification number was the only variable available to connect all three. The PRE80-only category has a linkage rate of 43%. Using the results from a record linkage linking the LCF to the Canadian Mortality Database (CMDB), it was established that over 25% of the linked pre-1980 immigrants had died by 2013. The proportion of deceased is potentially higher for the non-linked, since the data goes back to 1952. Someone who landed in Canada in 1952 at the age of 25 would have been 86 years old in 2013.

Records that appear only on the TRF have a linkage rate of 22%. This linkage rate can be explained by the fact that 2.2 million of the TRs only had visitor permits (other TRs could also have stayed in the country for a short period of time). Many, TRs do not file T1 forms (Brennan, 2014).

Table 3.3-1
Distribution of Records after Processing

Record Type	Records	Linked to tax	%Linked
ILF-PRE80-TR	4500	4500	100%
ILF-TR	1707300	1599800	94%
ILF only	5175900	4385300	85%
ILF-PRE80	12200	12200	100%
TR-PRE80	6400	6400	100%
TR only	5601700	1225600	22%
PRE80 only	4091700	1745800	43%
Total	**16599700**	**8979500**	**54%**

4. Conclusion

The longitudinal immigration database could be enhanced by including temporary residents and pre-1980 immigrants' data. This would result in new analytical capabilities using the IMDB and a broader coverage of Canada's immigrant population.

After resolving multiple issues related to combining the results of multiple record linkages to produce a file, a strategy to combine the content of the three datasets was developed. In the future, the quality of the dataset will have to be validated, and reference material will need to be developed. The quality assessment should include evaluation of inconsistencies such as those who only have visitor permits filing taxes, pre-filers who are not included on the TR file. In addition, the date of death will be added to the IMDB using the results from another record linkage. For subsequent years, a single record linkage combining TR and PR data should be performed. The content of the final file will need to be confirmed.

Acknowledgement

I would like to acknowledge the contribution of the methodologists who did the record linkages – James Brennan, Alexandr Diaz-Papkovich, Martin Lessard and Émilie Mayer – and my colleagues from the Social and Aboriginal Statistics Division (SASD) Scott McLeish, Athanase Barayandema and Ian Marrs.

References

Barayandema, A. (2015) “Income and mobility of immigrants, 2013”, Statistics Canada

Barayandema, A. (2015) “Feasibility study: Adding immigrants landed between 1952 and 1979 to the IMDB”, unpublished document, Statistics Canada

Brennan, J. (2014) “Landed Immigrants 1952 to 1979 linkage to Tax”, unpublished document, Statistics Canada

Brennan, J. (2014) “Temporary Residents Linkage to T1 tax”, unpublished document, Statistics Canada

Evra, R. (2015) “Feasibility Study: Redesign of the IMDB”, unpublished document, Statistics Canada

Evra, R. (2015) “Summary of processing steps”, unpublished document, Statistics Canada

Statistics Canada (2015) “Longitudinal Immigration Database (IMDB)”. Web site: http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5057