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

Using Administrative Data to Study Education in Canada

Martin Pantel[1]

Abstract

The Educational Master File (EMF) system was built to allow the analysis of educational programs in Canada. At the core of the system are administrative files that record all of the registrations to post-secondary and apprenticeship programs in Canada. New administrative files become available on an annual basis. Once a new file becomes available, a first round of processing is performed, which includes linkage to other administrative records. This linkage yields information that can improve the quality of the file, it allows further linkages to other data describing labour market outcomes, and it's the first step in adding the file to the EMF. Once part of the EMF, information from the file can be included in cross-sectional and longitudinal projects, to study academic pathways and labour market outcomes after graduation. The EMF currently consists of data from 2005 to 2013, but it evolves as new data become available. This paper gives an overview of the mechanisms used to build the EMF, with focus on the structure of the final system and some of its analytical potential.

Key Words: Education; Administrative data; Linked files.

1. Introduction

The Educational Master File (EMF) was created at Statistics Canada to increase the analytical possibilities of data on educational programs in Canada. It does this by creating a student-level identifier that is consistent across all educational datasets, and by facilitating linkages to other administrative files. The system involves many data-linkage activities, but these are not covered in detail here: instead, the focus of this article is to describe the structure of the EMF, and the processes involved when adding a new file into the system or when extracting an analytical file.

1.1 Motivation For Creating the EMF

Administrative data has become increasingly available and easy to use. When a new research question surfaces, an analyst must consider the possibility that the data already exist in some form. Sample surveys certainly still have their place in scientific research: they are more flexible, and can be designed for a specific set of research questions. However, if the data already exist in an adequate form, then its use for research is less burdensome, and typically much more timely and cost-effective than carrying out a sample survey.

Studies of post-secondary education have been carried out at Statistics Canada in the past, but most were of a cross-sectional nature. One recent source of longitudinal educational data came from the National Graduates Survey / Follow-up Survey of Graduates (NGS / FOG). The most recent data from this project targeted the graduating class of 2000: a sample of graduates were selected for the NGS in 2002, and respondents were interviewed a second time in 2005 for the FOG. Since then, two more editions of the NGS were carried out (in 2007 and 2013), but no follow-up surveys were administered for these collection years. As a result, a research need was identified: obtain data on recent graduates that allows us to study their academic careers and their life after graduation.

1.2 Data Sources

Two data sources form the primary components of the EMF: the Postsecondary Student Information System (PSIS) and the Registered Apprenticeship Information System (RAIS). Both are annual registries that track all of the enrolments in Canadian public postsecondary institutions or apprenticeship programs.

[1]Martin Pantel, Statistics Canada (Ottawa, Canada, K1A 0T6). Martin.Pantel@Canada.ca

PSIS is a census of enrolments in all Canadian public postsecondary institutions. It collects information pertaining to the programs, for example their classification and duration. Each enrolment on the file also includes data on the student, including some demographic information (name, birthdate, current address and phone number). A complete PSIS file typically includes approximately 3,000,000 records. The data are provided either by the institutions themselves, or in some cases by the provincial ministries of education. As such, there are varying levels of completeness and quality that must be taken into consideration. There are some gaps in the EMF, where a particular province might be completely missing for one year of PSIS, or have diminished quality. Sustained efforts are ongoing to affirm and improve the quality of the information gathered.

RAIS consists of all registered apprentices taking in-class or on-the-job training in trades where apprenticeship training is at least voluntary (if not compulsory). As with PSIS, each enrolment on the file includes data on the program as well as on the student. A complete RAIS file typically includes approximately 450,000 records.

In order to evaluate the relationships between education and the labour force, the PSIS and RAIS files can be linked to the T1 Family File, or T1FF. This is a database developed and managed at Statistics Canada, derived from income tax declarations (T1) and other administrative files. For a given fiscal year, the T1FF provides information on income from various sources, field of employment, tuition and education deductions, government transfers, as well as some demographic and geographic data. The linkage to the T1FF is done by using the Social Insurance Number (SIN).

At the time of this writing, the EMF consisted of PSIS files from 2008 to 2013 and RAIS files from 2008 to 2013, along with Tax data from 2006 to 2013. It also includes PSIS data for 2005-2007 for Canada's four Atlantic provinces. Any combination of these files can be used in the extraction of an analytical file for research purposes.

2. Structure of the EMF

Conceptually, the EMF can be described as consisting of three elements:

1) A student-level identifier (called MASTERID) that can be tied to every PSIS or RAIS file in the system, allowing all files to be linked. There is also a record of every SIN that is associated with each MASTERID, to allow linkages to Tax data. The MASTERIDs and SINs are considered highly sensitive data: access to them is restricted to the few people that are responsible for the maintenance of the EMF and the extraction of analytical files. The system meets all of the requirements set forth by Statistics Canada's strict policies and guidelines regarding record linkage and confidentiality.

2) A mechanism to add a new file to the system that, at the same time, allows corrections and maintenance to the existing system.

3) A mechanism to extract anonymous analytical files. These can be cross-sectional or longitudinal, and can include multiple years of PSIS, RAIS and T1FF data.

Computationally, the EMF can be seen as three collections of files, each stored separately:

1) The Contact files include all of the contact information, such as names, addresses and phone numbers. When adding a new file to the EMF, the set of existing Contact files are used to identify records that already exist in the system. If new contact information becomes available (someone gets a new phone number, for example), the old information is retained in the system. A potential future use of the set of Contact files could be to build a frame, or to obtain contact / tracing information.

2) The Program files include all of the information on the educational program, the institution, etc. – basically all of the educational data that will be part of the extracted file delivered to the analyst.

3) The Keys files include all of the linkage keys. For a given year, the file will include a RecordID variable (essentially the line number) that corresponds directly to the same variable on the Contact and Program files.

Associated with each record on the Keys file is a MASTERID, so the same student can be identified elsewhere on the same file and on any other file in the EMF. The Keys file also include any SINs associated with that MASTERID, and for each SIN, a score that reflects the quality of the SIN by indicating how it was obtained and the strength of the linkage.

3. Processing a New File

A good portion of the processing is due to the fact that not every record on the file is for a different student. Some students are enrolled in multiple programs in the same year, and in other cases, a record is duplicated on the file due to an administrative error.

Upon receiving a new PSIS or RAIS file, the first step is to assign a RecordID to each line on the file. Then, the file is linked to administrative records to obtain as many SINs as possible for each RecordID. The information on the file, including the newly obtained SINs, is then used to identify cases where two or more records are for the same person.

Once the records referring to the same person within the new file have been identified, the file is ready to be added to the EMF. Using all available information, each person on the new file is sought out on all of the files that already exist in the system. A student-level identifier called the MASTERID is assigned, that is consistent among all files that form the EMF. Finally, all of the information is stored in a secure and confidential manner, ready to be used in the extraction of analytical files. The following sections describe each step of this process in more detail.

3.1 Identifying Common Students on the New File

After having been assigned the RecordID, each record is linked to an administrative database which, for our purposes, can be described as a repository of SINs for all Canadians (one which also includes names and other contact information). This linkage is accomplished using an in-house hierarchical deterministic linkage program called MixMatch. The software allows the user to compare the names, birthdates, phone numbers and postal codes from the two files, and quantify the similarity between the records. MixMatch uses a set of matching tools that focus particularly on the comparison of character strings, like names and birthdates.

After this linkage, a typical RAIS file will have SINs for more than 98% of its records. For PSIS, the rates vary widely: if the PSIS file is complete (very little or no item non-response), then rates are around 91%. However, some files have very little information; without a name, for example, it becomes difficult to link records with any kind of confidence.

With the SINs now on the file, multiple records belonging to the same student can be sought out. This is accomplished using three main tools. As a unique identifier of individual Canadians, the SIN is an excellent tool for identifying pairs of records that refer to the same person. Within an institution, the StudentID is also a unique student-level identifier. The third tool is to use MixMatch once again, but this time it's used to do an "internal linkage": instead of seeking out common records between two separate files, this time the linkage is between the file and an exact copy of itself. The result is the list of RecordIDs within the file that describe the same person.

There are two other important reasons for obtaining as many SINs as possible. The value of an extracted analytical file can be improved by linking it with data describing labour force activities. The SIN allows the educational files to be linked to data from the T1FF. The other important reason for having the SIN is that it allows the development and refinement of MixMatch linkages strategies: when developing an algorithm that compares names and birthdates, the available SINs can be used as the baseline to calculate false positive and false negative rates.

Table 3.1-1 is a simplified example of what an incoming file may resemble, and how the additional information is linked. The first step is to add a RecordID (line number). Then, the SINs are added to the file; some of these may already be on the file, but most will have been obtained by linking to the SIN repository. Finally, the SIN, the StudentID and the contact information (names, birthdates, etc.) can be used to identify multiple records belonging to the same person. A UniqueID (or UID) is derived as a person-level identifier.

In this example, if one were to look strictly at the incoming data (the first four columns), it may not be evident that the first two records refer to the same person. However, the linkage to tax data may have yielded the same SIN for both of these records, thereby establishing the link. The UID was assigned to reveal that records L001 and L002 refer to the same person, U001.

Table 3.1-1
Example of a new PSIS file, with the SIN and MASTERID

Institution	StudentID	Name	Program	←	RecordID	←	SIN	←	UID
UofM	(missing)	Robert	Economics		L001		112233		U001
UofM	12345	Bob	Statistics		L002		112233		U001
W.Coll.	GD89	Marie	Engineering		L003		445588		U002

The SIN and the UniqueID are considered highly sensitive and are guarded very closely. In order to respect the strict regulations governing record linkage and confidentiality policies, identifying data and linkage keys must be stored separately. Hence, once the file shown in the above table is complete, it gets broken into three pieces, which will be stored in separate locations. The RecordID is the common thread that would allow the original file to be recreated. For brevity, only one variable is shown in the Program and Contact files in Table 3.1-2, but these files would include many other variables of a similar nature.

Table 3.1-2
Breakdown of new PSIS file for storage

Program File	
RecordID	**Program**
L001	Economics
L002	Statistics
L003	Engineering

Contact File	
RecordID	**Name**
L001	Robert
L002	Bob
L003	Marie

Keys File		
RecordID	**SIN**	**UID**
L001	112233	U001
L002	112233	U001
L003	445588	U002

3.2. Adding the File to the EMF

The three files depicted in Table 3.1-2 (Program, Contact and Keys) all refer to the final storage state of one file (PSIS or RAIS for a given year). In fact, the EMF system is the collection of all such files: all Program files, for all yearly PSIS and RAIS files, are stored together in one location. All Contact files and Keys files are similarly grouped together, and all three sets are stored in separate locations.

When the first round of processing for a new file is done (RecordID, SINs, and UniqueID have been added), then the file is ready to be added to the EMF. The process can be difficult and time-consuming to execute, but it is simple to describe: the UniqueID is updated into a MASTERID that is consistent across all files. To do this, each new record is compared to all records that already exist on the EMF. If a match is found, then the existing MASTERID is assigned to the new record(s), overriding its (their) UniqueID. Once all linked records have been assigned pre-existing MASTERIDs, then the remaining UniqueIDs can be replaced by new MASTERIDs.

The linkages are done in the same series of steps that were described in section 3.1. The SINs of the new records are compared to the SINs belonging to all of those that currently exist on the EMF. The StudentIDs are similarly compared, under the condition that the institutions are also the same. Finally, all of the existing Contact files are combined, and the names / birthdates / phone numbers / postal codes are compared with the same information coming from the new file.

When different strategies lead to conflicting linkages, only the “strongest” match is retained. The measure of strength depends on the source of the linkage. If the linkage was obtained by MixMatch, then the score reflects the similarity of the records as defined by the MixMatch rules. If the linkage was obtained by matching SINs, then the strength measure indicates whether the SIN was on the original file, or if it was itself obtained by one (or more) of the linkage processes.

3.3 Updates and Corrections to the EMF

Adding a new file to the EMF brings new information into the system, which can bring to light some errors within the new file, or within the existing version of the EMF. There are two types of errors:

1) False matches: Two different people linked due to similar names, or due to the same SIN being inadvertently assigned. For example, an error could be due to twins (same birthdate and last name) with similar given names, who set off to college and decide to live together (same phone number and postal code).

2) Missed matches: Two records that refer to the same person might be missed, if the data on one or both of the records is missing, erroneous, or has changed. For example, some institutions did not provide students' names, making it very difficult to find them elsewhere in the system.

Corrections can be brought to the EMF (or to the incoming file) when such errors are identified. Correcting the EMF consists of updating the various Keys files, by changing or updating MASTERIDs and by deleting SINs when necessary. The majority of the updates are due to missed matches, and these can be resolved automatically by setting the MASTERIDs to the same value. False matches are a much smaller group, but they can still represent thousands of cases – and since their resolution requires manual verification and correction, the task can be quite time-consuming.

4. Example: Extracting an Analytical File from the EMF

Suppose we wish to study the people that were enrolled in postsecondary education (i.e., in PSIS) in 2010, 2011 and 2012, along with their tax data from 2012. We would begin by examining the Keys files for all three PSIS years of interest, and seeing which MASTERIDs exist on all three files. Table 4.1-1 shows that two MASTERIDs satisfy this condition. For any selected cases that have more than one SIN on file (this occurs for approximately 4% of cases), we would at this time select the most appropriate one for linkage to 2012 Tax data. Occasionally, a record with a SIN cannot be linked to tax, either because we were unable to link them (an error in the recorded SIN, for example) or the person did not file taxes for that given year.

Table 4.1-1
Example of Keys files from three versions of PSIS

Keys PSIS 2010		
Rec.ID	**MASTERID**	**SIN**
L11	M01	101
L12	**M02**	**102**
L13	**M02**	**102**
L14	M03	103
L15	**M04**	**104**
L16	M05	105

Keys PSIS 2011		
Rec.ID	**MASTERID**	**SIN**
L21	M06	201
L22	M07	202
L23	**M02**	**102**
L24	M08	203
L25	**M04**	**104**
L26	M05	105

Keys PSIS 2012		
Rec.ID	**MASTERID**	**SIN**
L31	M09	301
L32	**M02**	**102**
L33	M10	302
L34	M11	303
L35	**M04**	**104**
L36	M12	304

Each file (PSIS year) can now be handled independently: the selected RecordIDs are used to fetch the analytical data from the appropriate Program file, and the SINs are used to fetch the appropriate data from the Tax file.

If necessary, indicator variables can also be created and added to the files. Some common indicator variables are ones that describe characteristics surrounding the SIN. For example: provisional SINs are sometimes assigned to recent immigrants or victims of identity theft. These provisional SINs can be easily identified by their first digit. They can be of interest to analysts, since there may be a different (permanent) SIN on another file that would refer to the same person. Since confidentiality concerns prohibit the release of the SIN itself on the final file, the analyst may request a binary flag indicating simply whether or not the SIN was a provisional one.

The final files delivered to the analyst could resemble the set of files depicted here:

Table 4.1-2
Example of Analytical files

PSIS 2010			
RequestID	**Program data**	**Tax data 2012**	**Indicators**
R01	**Educational program, socio-demographic data, etc. 2010**	**Income, tuition, etc. 2012**	**Flags for multi-SINs, temporary SINs, etc.**
R01			
R02			

PSIS 2011			
RequestID	**Program data**	**Tax data 2012**	**Indicators**
R01	**Educational program, socio-demographic data, etc. 2011**	**Income, tuition, etc. 2012**	**Flags for multi-SINs, temporary SINs, etc.**
R02			

PSIS 2012			
RequestID	**Program data**	**Tax data 2012**	**Indicators**
R01	**Educational program, socio-demographic data, etc. 2012**	**Income, tuition, etc. 2012**	**Flags for multi-SINs, temporary SINs, etc.**
R02			

The anonymity of the files can be emphasized here again: the analytical files do not include SINs nor any personal identifiers, only analytical data and indicator variables. A RequestID is created to mask the real MASTERID: the mapping between the two can only be decrypted by the few people that have access to the EMF. Furthermore, the mapping is only valid for the one request; the same MASTERID would get a different RequestID if extracted to another analytical file.

5. Future Work

In the current system, every linked record has a score that reflects its strength, but only at the time that specific linkage was established. This score is particularly useful when bringing updates or corrections to the EMF. However, since the EMF is dynamic, it would be useful if the score could change to reflect any corrections made that relate to that record. The form of such a quality indicator is difficult to define, but it could be a useful tool for analysts.

The EMF will continue to evolve as files are added to the system; it is expected that new PSIS, RAIS and T1FF files will become available every year. It's also possible that the analytical value of the EMF could be further expanded by linking its data to those coming from other sectors, such as a Census of population, or databases describing health, justice or other sectors.

References

Caron, P., Ellison, J., and Faucher, D. (2013) "Linkage of Education Files", unpublished document, Ottawa, Canada: Statistics Canada.

Pantel, M. (2015) "Educational Master File (EMF) Methodology", unpublished document, Ottawa, Canada: Statistics Canada.