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

Student Pathways and Graduate Outcomes[1]

Aimé Ntwari[2]

Abstract

Files with linked data from the Statistics Canada, Postsecondary Student Information System (PSIS) and tax data can be used to examine the trajectories of students who pursue postsecondary education (PSE) programs and their post-schooling labour market outcomes. On one hand, administrative data on students linked longitudinally can provide aggregate information on student pathways during postsecondary studies such as persistence rates, graduation rates, mobility, etc. On the other hand, the tax data could supplement the PSIS data to provide information on employment outcomes such as average and median earnings or earnings progress by employment sector (industry), field of study, education level and/or other demographic information, year over year after graduation. Two longitudinal pilot studies have been done using administrative data on postsecondary students of Maritimes institutions which have been longitudinally linked and linked to Statistics Canada Ttx data (the T1 Family File) for relevant years. This article first focuses on the quality of information in the administrative data and the methodology used to conduct these longitudinal studies and derive indicators. Second, it will focus on some limitations when using administrative data, rather than a survey, to define some concepts.

Key Words: Postsecondary education, record linkage, tax.

1. Introduction

The Postsecondary Student Information System (PSIS) data provides detailed annual information on enrolments and graduations from Canadian postsecondary institutions (universities and colleges) by education level, field of study and certain demographic variables. The T1 Family File (T1FF) provides detailed annual taxation data on employment earnings and other information such as place of residence and industry of employment. The recently developed Education Longitudinal Linkage Platform (ELLP) allows data for unique records between the annual files and between the data sources to be combined. Longitudinal studies using information obtained through PSIS-T1FF data linkages can help to fill data gaps in postsecondary education indicators such as persistence and longitudinal graduation rates, as well as graduates' labour market outcomes, for different subgroups of interest.

1.1 Motivation for longitudinal studies using postsecondary education data

Statistics Canada publishes annual counts of postsecondary enrolments and graduates and a number of pan-Canadian education indicators. In addition, Statistics Canada submits data for the annual publication of education- and labour market-related indicators by international organizations. A long-acknowledged gap in these indicators is the ongoing production of a postsecondary graduation rate that traces individuals over time to identify program completion. Other areas of interest include a better understanding of students' pathways into and through postsecondary schooling and their interaction with the labour market after graduation. Past studies have used annual data or used longitudinal surveys to obtain this type of information. The longitudinal linkage of existing annual administrative data and the inclusion of tax variables allows the development of new education indicators related to these areas.

[1] This article would not have been possible without the tremendous contribution of Eric Fecteau, Christine Hinchley, Sylvie Gauthier Rubab Arim and Louise Marmen.
[2] Aimé Ntwari, Statistics Canada, Ottawa, Canada, K1A 0T6 (Aime.Ntwari@Canada.ca)

1.2 Data sources

There are two data sources form the primary components of this project: the Postsecondary Student Information System and the T1 Family File. PSIS is an annual registry that tracks all enrolments and graduations in Canadian public colleges, cégeps and universities and the T1FF includes annual tax information for all tax filers.

1.2.1 PSIS

The Postsecondary Student Information System is a national survey that enables Statistics Canada to provide detailed information on enrolments and graduates of Canadian public postsecondary institutions in order to meet policy and planning needs in the field of postsecondary education. PSIS collects information pertaining to the programs offered at an institution, as well as information regarding the students themselves and the program(s) in which they were registered, or from which they have graduated. PSIS is also designed to collect continuing education data. This is a mandatory survey. The data are provided either by the institutions themselves, or in some cases by the provincial ministries of education or another centralized organization. Results presented in this article concern only data from Maritime Postsecondary Information System for the 6 year database (reporting years 2005/2006 to 2011/2012) provided by the Maritimes Provinces High Education Commission (MPHEC), containing only the provinces of New Brunswick, Nova Scotia and Prince Edward Island.

1.2.2 T1FF

In order to evaluate earnings outcomes and geographic mobility following graduation, the PSIS files can be linked to the T1 Family File. The T1FF 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 tuition and education deductions, government transfers, as well as some demographic and geographic data.

2. Overview of the linkage methodology

The Statistics Canada Education Longitudinal Linkage Platform (ELLP) was developed to contain a register of keys that links PSIS data longitudinally and to the T1FF and the Registered Apprenticeship Information System (RAIS). In the future, other linkage projects will also be possible; e.g., by linking data from the National Graduates Survey (NGS), the Census and the National Household Survey (NHS), etc. to the ELLP.

The linkage between PSIS for universities in the Maritimes and the T1FF needed to create the ELLP is initially performed using the Statistics Canada Linkage Control File (LCF), which is a combination of many years of tax file identifiers. The main objective of this linkage is to obtain a unique key to identify individuals across data sets. If a Social Insurance Number (SIN) exists, the SIN from the LCF is matched to the records in PSIS. For records without a SIN, another key is built using identifiers such as a concatenation of names, date of birth, postal code, etc. The linkage is done using iterations of several probabilistic and direct methods.

Using the ELLP unique keys for each student, analytical files can then be built that include variables from the tax and PSIS files for research purposes. These files do not, however, contain personal identifiers.

3. Cohorts definitions

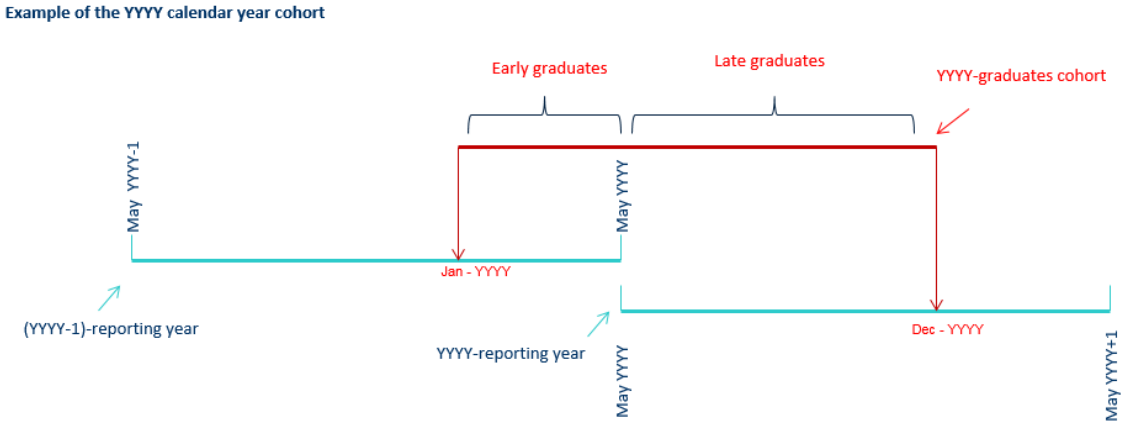
Two pilot studies are being undertaken with the goal to derive and develop some education and labour market indicators using the ELLP longitudinal linkages. In these projects, two different types of cohorts are used. The first is based on new entrants to a program during a given PSIS reporting year (approximately from April to May of the next year) and the second is based on students graduating in a given calendar year. Additionally, the project focuses on individual students over time rather than the student program records counted in the annual PSIS data release. This section further outlines how the two types of cohorts were defined and derived.

3.1 Student pathways cohorts

Every cohort is built by linking each new student program record in the annual PSIS file to all subsequent reporting years available in order to identify all program records belonging to unique individuals. At this step, a student of a given starting year cohort (e.g. 2005/2006) may appear in the linked file once, twice or more times. They may appear in separate reporting years, but also may have more than one program record in a given year. For example, 'simultaneous enrollees' in the fall in the Maritime university data are primarily comprised of students enrolled in two or three institutions at the same time. Several specifications are required in order to follow one record per student per year and have longitudinally consistent information. Newly enrolled student are defined as cohorts and followed in subsequent PSIS files. Seven cohorts of new entrants (the academic years 2005/2006 through 2011/2012) were retained from the Maritime universities PSIS 2005/2006 to 2012/2013 reporting year data.

3.2 Graduates outcome cohorts

Graduate cohorts were defined based on date of graduation before they were linked to all subsequent tax files. Since the tax information in the T1FF is provided from January to December, for comparison it was decided to define the graduate cohort based on graduation within a calendar year. Each graduate cohort was built from two consecutive PSIS reporting years. This was done by selecting the records corresponding to students who officially received a qualification and completed their program in that calendar year as shown in the following chart.



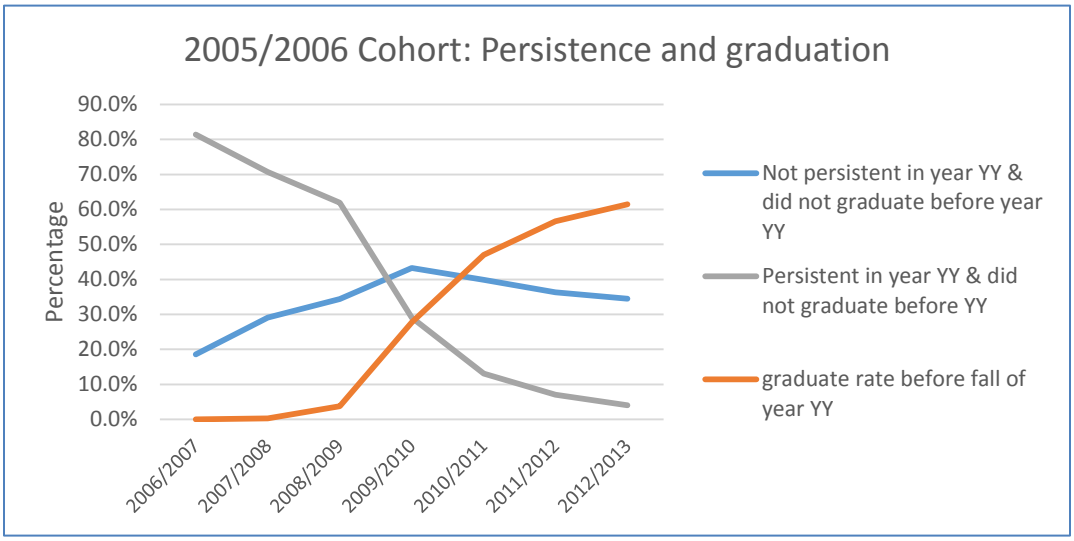
Six cohorts of graduates (the calendar years 2006 through 2011) were retained from the Maritime universities PSIS 2005/2006 to 2012/2013 reporting year data.

4. Possible student pathways indicators

Longitudinal postsecondary student data allow the tracking of different pathways of students and the development of education indicators on these pathways. Student pathways show various patterns; some students continue in the same program until graduation; others switch to a different program at the same institution; some change institutions, either to the same program or to a different program; still other students abandon their studies in the first year of registration or in following years and do not return to a Maritime university; some abandon and return at a later date; etc. These concepts will be summarized as indicators of different types of persistence, mobility and graduation rates. In this study, all types of student programs are traced through available subsequent years and different indicators are derived.

Chart **4.1**

Persistence and graduation rates for the 2005/2006 bachelor's degree new entrants cohort



A breakdown of these student pathways indicators has been done in order to identify different socio-demographic factors associated to them. The table below shows, for example results of logistic regression of persistence indicator on gender variable.

Table 4.1

Results of logistic regression model for persistence - factors associated with first year persistence (* sign at p)

Cohort year	Effect	O.R.	Lower CI	Upper CI
2005/2006	Female **vs** Male	1.152*	1.037	1.280
2006/2007	Female **vs** Male	1.188*	1.067	1.323
2007/2008	Female **vs** Male	1.172*	1.062	1.293
2008/2009	Female **vs** Male	1.097	0.99	1.22
2009/2010	Female **vs** Male	1.217*	1.093	1.356
2010/2011	Female **vs** Male	1.175*	1.054	1.308
2011/2012	Female **vs** Male	1.054	0.946	1.174

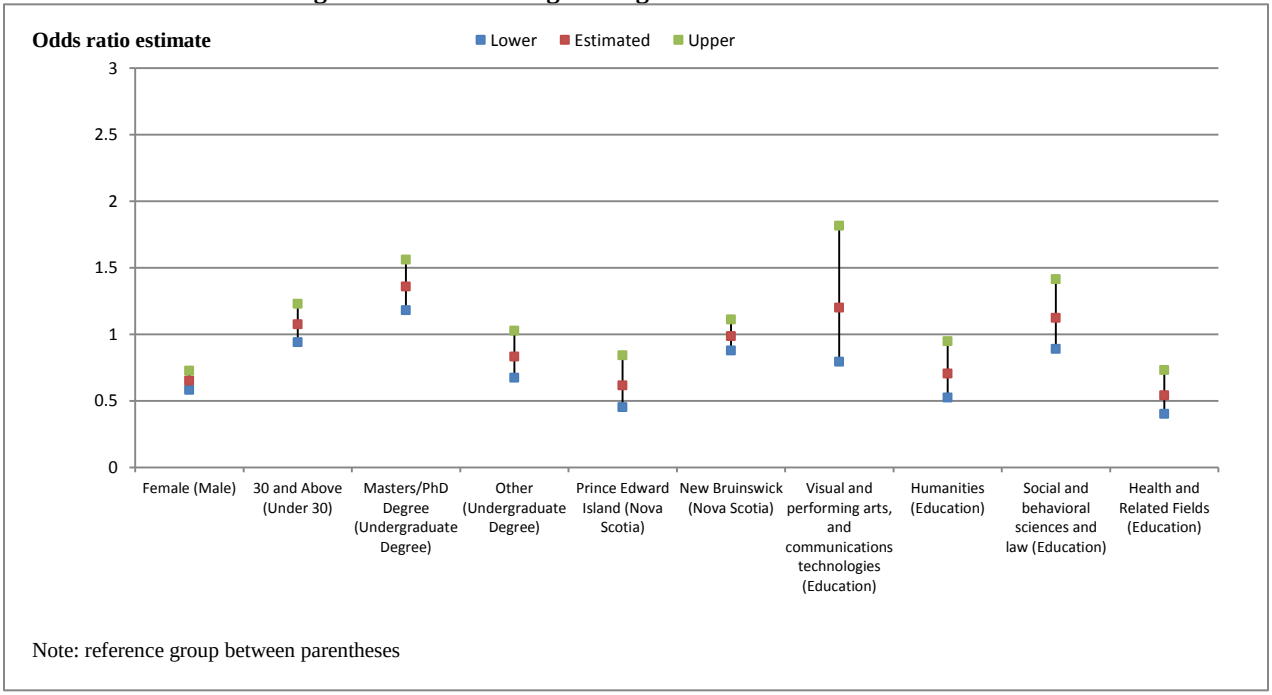
*significant at α=0.05

5. Possible graduate outcome indicators

The linkage between the PSIS files and the T1FF files allows a study of the progress of employment income over time for a cohort of graduates. The impact of an event that happened at a given time can also be measured by comparing transition patterns. The geographic mobility of graduates in the labour market can be traced using the tax information.

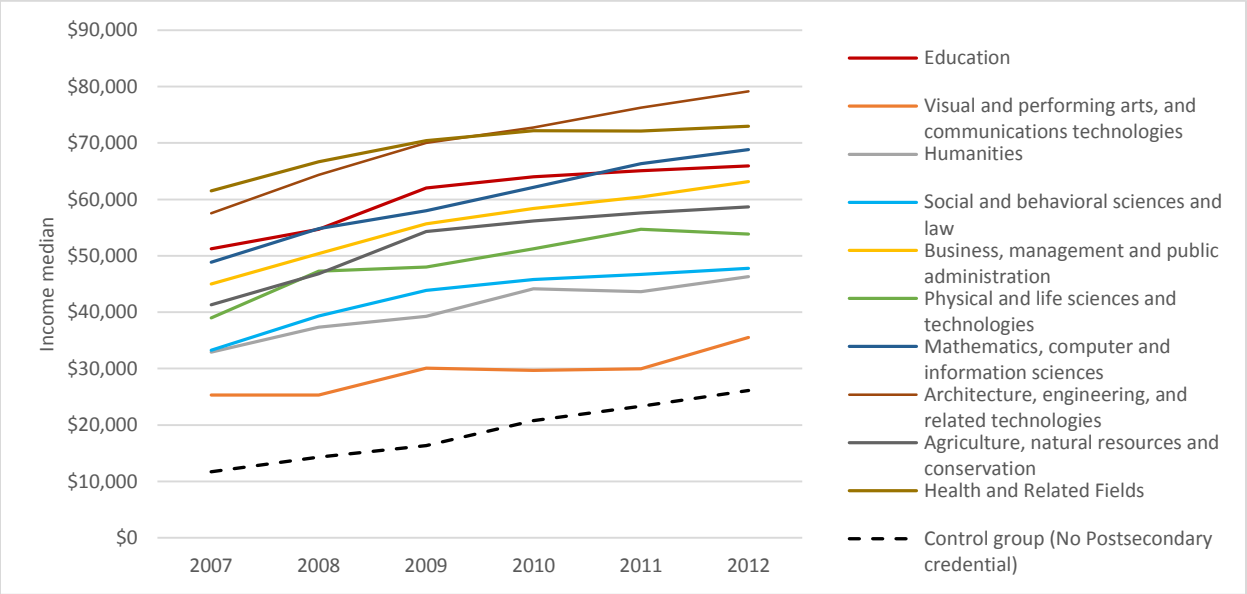
Data linkage is a well-accepted approach to find matches between two sets of records but it can lead to inexact matches (incorrect links and unlinked records). Small proportions of these errors are expected between records referring to the same individual, whether due to linking a pair of records that do not belong to the same unit (incorrect links or false positive) or to missing pair of records that belong to the same unit (unlinked records or missed pairs), however the aim is to minimize them where possible. The additional challenge is to identify among unlinked records, false negatives from true negatives. The chart below which gives logistic regression results shows that the linkage process has distorted the initial cohort and may have introduced a bias. An adjustment for false negative records may be explored.

Chart 5.1
Factors associated to linking to tax: results of logistic regression model



Note: reference group between parentheses

Chart 5.2
2006 Cohort: Median of employment income of bachelor's and master's/doctorate degree holders, by field of study (CIP primary grouping)



The chart shows a pattern change in 2009. What is the evolution of employment income for the 2006 longitudinal population between 2007 and 2009 and 2012? It is possible to measure the impact of an event occurred in 2009, possibly the economic recession, by comparing transition matrices. The table 5.1 gives a test for equality of the 2 transition matrices.

Table 5.1
Income Quartiles change between the 2007 to 2009 period and the 2009 to 2011 period for the Education field of study qualification holders

Status at time T	Time	Status at time T+2			
		Lowest	Lower-Middle	Upper-Middle	Highest
Lowest	T=2007	48.4%	37.1%	9.4%	5.0%
	T=2009	52.1%	35.2%	8.5%	4.2%
Lower-Middle	T=2007	18.8%	66.1%	13.8%	1.3%
	T=2009	23.5%	65.8%	10.3%	0.4%
Upper-Middle	T=2007	13.5%	18.0%	58.6%	9.8%
	T=2009	7.2%	13.8%	61.8%	17.1%
Highest	T=2007	1.9%	1.6%	9.3%	87.1%
	T=2009	1.7%	4.1%	31.7%	62.4%

After conducting a chi-square test of independence between Time and Post-result for each level of Pre-result and adding the respective chi-square and the respective degrees of freedom, the difference appear to be statistically significant ($p < 0.00001$).

6. Summary and Future Work

Longitudinal linked administrative data have a high potential of analysis and have the advantage to have a low cost compare to longitudinal surveys but are not themselves free of errors. Naively treating a linked file as if it were linked without errors will, in general, lead to biased estimates. An approach of weighting will be tested to address the issue of unlinked records for the employment earnings analysis.

References

Ross, Theresa (2009), "Moving Through, Moving On: Persistence in Postsecondary Education in Atlantic Canada, Evidence from the PSIS", *Statistics Canada Catalogue no. 81-595-M-No.072*, Research paper, Statistics Canada.

Ross, Dejan (2014), "Analysis of long-term outcomes for university graduates in Information and Communication Technology programs", Education Policy Research Initiative, Canada: University of Ottawa
.

Frenette, Marc and Yuri Ostrovsky. (2014), "The cumulative Earnings of Postsecondary Graduates over 20 Years: Results by Field of Study", *Statistics Canada Catalogue no.11-626-X – No. 040,* Statistics Canada.

Saidi, Ntwari. (2014), "Weighting Adjustment for False Negatives in Record Linkage", paper presented at the International Health Data Linkage Conference, Vancouver, Canada. .