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

Trajectoires des étudiants et résultats des diplômés[1]

Aimé Ntwari[2]

Résumé

Les fichiers comprenant des données couplées du Système d'information sur les étudiants postsecondaires (SIEP) de Statistique Canada et des données fiscales peuvent servir à examiner les trajectoires des étudiants qui poursuivent des études postsecondaires (EPS) et leurs résultats sur le marché du travail par la suite. D'une part, les données administratives sur les étudiants couplées de façon longitudinale peuvent fournir des renseignements agrégés sur les trajectoires des étudiants pendant leurs études postsecondaires, comme les taux de persévérance, les taux de diplomation, la mobilité, etc. D'autre part, les données fiscales peuvent compléter le SIEP et fournir des renseignements sur les résultats au chapitre de l'emploi, comme la rémunération moyenne et médiane ou la progression de la rémunération selon le secteur d'emploi (industrie), le domaine d'études, le niveau de scolarité et/ou d'autres données démographiques, année après année suivant l'obtention du diplôme. Deux études longitudinales pilotes ont été menées au moyen de données administratives sur les étudiants postsecondaires d'établissements des Maritimes, qui ont été couplées de façon longitudinale et avec le fichier de données fiscales de Statistique Canada (le fichier sur la famille T1) pour les années pertinentes. Cet article met d'abord l'accent sur la qualité des renseignements compris dans les données administratives et sur la méthode utilisée pour mener ces études longitudinales et calculer des indicateurs. En deuxième lieu, elle portera sur certaines limites liées à l'utilisation de données administratives, plutôt que de données d'enquête, pour définir certains concepts.

Mots-clés : études postsecondaires, couplage d'enregistrements, données fiscales.

1. Introduction

Le Système d'information sur les étudiants postsecondaires (SIEP) fournit des données annuelles détaillées sur les inscriptions et les diplômes des établissements postsecondaires au Canada (universités et collèges), selon le niveau de scolarité, le domaine d'études et certaines variables démographiques. Le fichier sur la famille T1 (FFT1) fournit des données annuelles de l'impôt sur la rémunération d'emploi et d'autres renseignements, comme le lieu de résidence et le secteur d'emploi. La Plateforme de couplage longitudinale en éducation (PCLE) permet la combinaison d'enregistrements uniques entre les fichiers annuels et entre les sources de données. Des études longitudinales reposant sur des données obtenues par suite du couplage des données du SIEP-FFT1 peuvent aider à combler les lacunes dans les données des indicateurs de l'enseignement postsecondaire, comme les taux de persévérance et les taux de diplomation longitudinaux, ainsi que les résultats des diplômés sur le marché du travail, pour différents sous-groupes d'intérêt.

1.1 Motivation des études longitudinales reposant sur des données sur les études postsecondaires

Statistique Canada publie des chiffres annuels sur les inscriptions et les diplômes postsecondaires, ainsi qu'un certain nombre d'indicateurs pancanadiens de l'éducation. En outre, il soumet des données pour la publication annuelle

[1] Cet article n'aurait pas été possible n'eût été la précieuse contribution des personnes suivantes : Eric Fecteau, Christine Hinchley, Sylvie Gauthier Rubab Arim et Louise Marmen.
[2] Aimé Ntwari, Statistique Canada, Ottawa, Canada, K1A 0T6 (Aime.Ntwari@Canada.ca)

d'indicateurs liés à l'éducation et au marché du travail par les organismes internationaux. Ces indicateurs comportent une lacune de longue date, à savoir la production permanente d'un taux de diplomation postsecondaire permettant de suivre les personnes au fil du temps, afin de déterminer les programmes terminés. Parmi les autres domaines d'intérêt figure une meilleure compréhension des trajectoires des étudiants dans les études postsecondaires et de leur interaction avec le marché du travail après l'obtention du diplôme. Des études passées ont utilisé des données annuelles ou des enquêtes longitudinales pour obtenir ce type de renseignements. Le couplage longitudinal des données administratives annuelles existantes et l'inclusion de variables fiscales permet l'élaboration de nouveaux indicateurs de l'éducation liés à ces domaines.

1.2 Sources des données

Il existe deux sources de données pour les composantes principales de ce projet : le Système d'information sur les étudiants postsecondaires et le fichier sur la famille T1. Le SIEP est un registre annuel qui suit l'ensemble des inscriptions et des diplômes dans les collèges publics, les cégeps et les universités au Canada, et le FFT1 comprend des données fiscales annuelles de tous les déclarants fiscaux.

1.2.1 SIEP

Le Système d'information sur les étudiants postsecondaires est une enquête nationale qui permet à Statistique Canada de publier de l'information détaillée sur les effectifs et les diplômés des établissements postsecondaires publics canadiens, afin de répondre aux besoins en matière d'élaboration de politiques et de planification dans le domaine de l'éducation postsecondaire. Le SIEP recueille des renseignements relatifs aux programmes offerts dans un établissement, ainsi que des renseignements concernant les étudiants proprement dits et les programmes dans lesquels ils sont inscrits, ou dont ils sont diplômés. Le SIEP est aussi conçu pour recueillir des données sur l'éducation permanente. Il s'agit d'une enquête obligatoire. Les données sont fournies par les établissements proprement dits ou, dans certains cas, par les ministères provinciaux de l'Éducation ou un autre organisme centralisé. Les résultats compris dans la présente communication concernent uniquement les données du système d'information postsecondaire des Maritimes pour la base de données de six ans (années de déclaration de 2005-2006 à 2011-2012) fournies par la Commission de l'enseignement supérieur des provinces Maritimes (CESPM), qui comprend uniquement le Nouveau-Brunswick, la Nouvelle-Écosse et l'Île-du-Prince-Édouard.

1.2.2 FFT1

Afin d'évaluer les résultats au chapitre de la rémunération et la mobilité géographique suivant l'obtention du diplôme, les fichiers du SIEP peuvent être couplés au fichier sur la famille T1. Le FFT1 est une base de données élaborée et gérée à Statistique Canada, qui est établie à partir des déclarations de revenu T1 et d'autres fichiers administratifs. Pour un exercice donné, le FFT1 fournit des renseignements sur les déductions au titre des frais de scolarité et des études, les transferts gouvernementaux, ainsi que certaines données démographiques et géographiques.

2. Aperçu de la méthode de couplage

La Plateforme de couplage longitudinale en éducation (PCLE) a été élaborée pour conserver un registre des clés qui permettent le couplage des données du SIEP de façon longitudinale, de même qu'avec le FFT1 et le Système d'information sur les apprentis inscrits (SIAI). À l'avenir, d'autres projets de couplage seront aussi possibles, par exemple le couplage des données de l'Enquête nationale auprès des diplômés (END), du Recensement et de l'Enquête nationale auprès des ménages (ENM), etc. avec celles de la PCLE.

Le couplage du SIEP pour les universités des Maritimes et du FFT1 nécessaire pour créer la PCLE se fait initialement au moyen du Fichier de contrôle des couplages (FCC) de Statistique Canada, qui représente une combinaison de nombreuses années d'identificateurs de dossiers fiscaux. Le principal objectif de ce couplage est d'obtenir une clé unique pour identifier les personnes dans les ensembles de données. Si un numéro d'assurance sociale (NAS) existe, le NAS du FCC est apparié aux enregistrements du SIEP. Dans le cas des enregistrements sans NAS, une autre clé est élaborée à partir d'identificateurs comme la concaténation des noms, de la date de naissance, du code postal, etc. Le couplage est effectué par itération de plusieurs méthodes probabilistes et directes.

À partir des clés uniques de la PCLE pour chaque étudiant, des fichiers analytiques peuvent alors être élaborés et comprennent des variables des dossiers fiscaux et du SIEP aux fins de la recherche. Toutefois, ces fichiers ne comprennent pas d'identificateurs personnels.

3. Définitions de cohortes

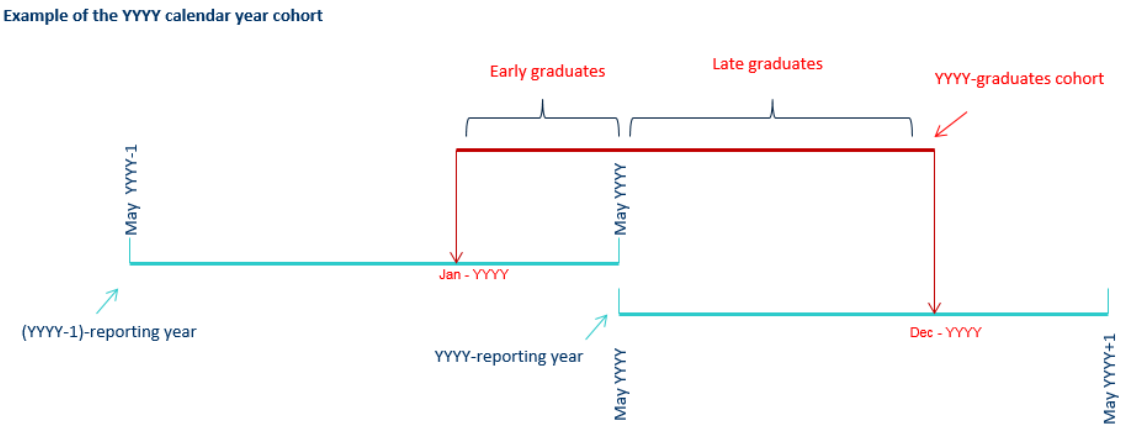
Deux études pilotes sont entreprises en vue de calculer et d'élaborer certains indicateurs de l'éducation et du marché du travail, au moyen des couplages longitudinaux de la PCLE. Dans ces projets, on a recours à deux types différents de cohortes. La première est fondée sur les personnes nouvellement inscrites dans un programme au cours d'une année de déclaration donnée du SIEP (d'avril à mai de l'année suivante environ), et la deuxième, sur les étudiants qui obtiennent leur diplôme au cours d'une année civile donnée. En outre, le projet est axé sur les étudiants individuels au fil du temps, plutôt que sur les enregistrements de programmes d'études dénombrés dans la diffusion annuelle de données du SIEP. La présente section fait aussi état de la façon dont les deux types de cohortes ont été définis et calculés.

3.1 Cohortes de trajectoires d'étudiants

Chaque cohorte est établie en couplant chaque nouvel enregistrement de programme d'études du fichier annuel du SIEP à toutes les années de déclaration subséquentes disponibles, afin d'identifier tous les enregistrements de programme se rapportant à des personnes en particulier. À cette étape, un étudiant appartenant à une cohorte de première année donnée (p. ex. 2005-2006) peut apparaître une fois, deux fois ou plusieurs fois dans le fichier couplé. Il peut apparaître pour des années de déclaration distinctes, mais peut aussi avoir plus d'un enregistrement de programme une année donnée. Par exemple, les « inscriptions multiples » constatées à l'automne dans les données des universités des Maritimes correspondent principalement à des étudiants inscrits simultanément dans deux ou trois établissements. Plusieurs spécifications sont requises pour suivre un enregistrement par étudiant par année et obtenir des données uniformes au niveau longitudinal. Les étudiants nouvellement inscrits sont définis comme des cohortes et sont suivis dans les fichiers subséquents du SIEP. Sept cohortes d'étudiants nouvellement inscrits (années scolaires 2005-2006 jusqu'à la fin de 2011-2012) ont été conservées à partir des données des années de déclaration de 2005-2006 à 2012-2013 du SIEP dans les universités des Maritimes.

3.2 Cohortes de résultats de diplômés

Les cohortes de diplômés ont été définies sur la base de la date d'obtention du diplôme, avant d'être couplées à tous les dossiers fiscaux subséquents. Comme les données fiscales du FFT1 sont fournies de janvier à décembre, à des fins de comparaison, il a été décidé de définir la cohorte de diplômés sur la base de l'obtention du diplôme à l'intérieur d'une année civile. Chaque cohorte de diplômés a été constituée à partir de deux années de déclaration consécutives du SIEP. À cette fin, on a sélectionné les enregistrements correspondants aux étudiants qui ont officiellement obtenu un titre de compétences et terminé leur programme au cours de cette année civile, comme le montre le graphique suivant.

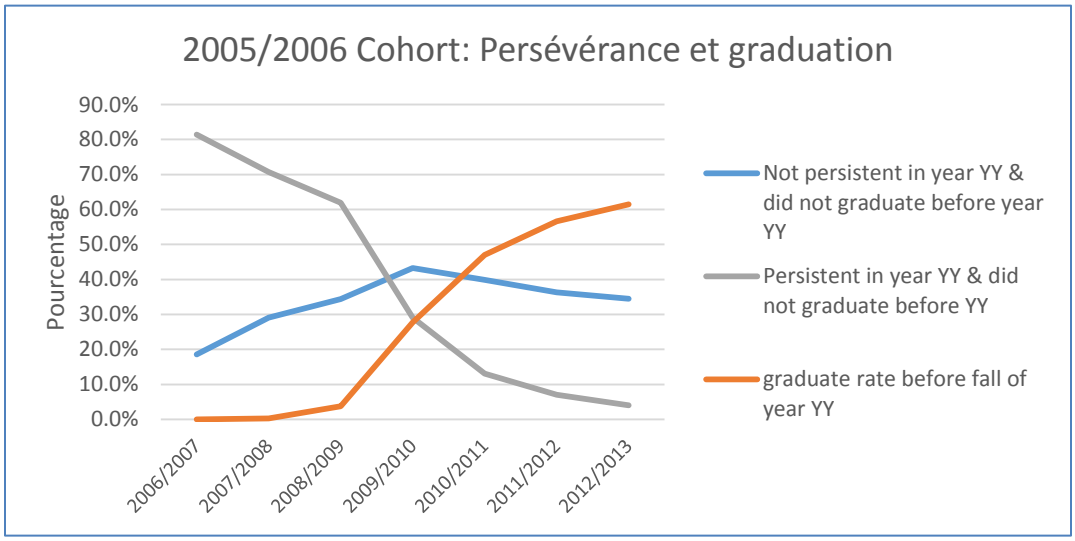


Six cohortes de diplômés (années civiles 2006 jusqu'à la fin de 2011) ont été conservées à partir des données des années de déclaration 2005-2006 à 2012-2013 du SIEP pour les universités des Maritimes.

4. Indicateurs possibles des trajectoires d'étudiants

Les données longitudinales sur les étudiants postsecondaires permettent de suivre les différentes trajectoires des étudiants et d'élaborer des indicateurs de l'éducation concernant ces trajectoires. Les trajectoires des étudiants affichent des tendances diverses; certains étudiants poursuivent le même programme jusqu'à l'obtention du diplôme; d'autres passent à un programme différent dans le même établissement; certains changent d'établissement, soit dans le même programme ou dans un programme différent; et d'autres abandonnent leurs études au cours de leur première année d'inscription ou pendant les années suivantes et ne reviennent pas dans une université des Maritimes; certains abandonnent et reviennent à une date ultérieure; etc. Ces concepts seront résumés comme des indicateurs des différents types de taux de persévérance, de mobilité et de diplomation. Dans la présente étude, tous les types de programmes d'études font l'objet d'un suivi, grâce aux années subséquentes disponibles, et différents indicateurs sont calculés.

Graphique 4.1
Taux de persévérance et de diplomation pour la cohorte des nouveaux inscrits au baccalauréat de 2005-2006



Une ventilation de ces indicateurs de trajectoires d'étudiants a été effectuée, afin de déterminer les différents facteurs sociodémographiques qui y sont associés. Le tableau ci-après montre, par exemple, les résultats de la régression logistique de l'indicateur de la persévérance sur la variable du sexe.

Tableau 4.1
Résultats du modèle de régression logistique pour la persévérance – facteurs associés à la persévérance la première année (* signe à p)

Année de la cohorte	Effet	RC	IC inférieur	IC supérieur
2005-2006	Femmes **par rapport** à hommes	1,152*	1,037	1,280
2006-2007	Femmes **par rapport** à hommes	1,188*	1,067	1,323
2007-2008	Femmes **par rapport** à hommes	1,172*	1,062	1,293
2008-2009	Femmes **par rapport** à hommes	1,097	0,99	1,22
2009-2010	Femmes **par rapport** à hommes	1,217*	1,093	1,356
2010-2011	Femmes **par rapport** à hommes	1,175*	1,054	1,308
2011-2012	Femmes **par rapport** à hommes	1,054	0,946	1,174

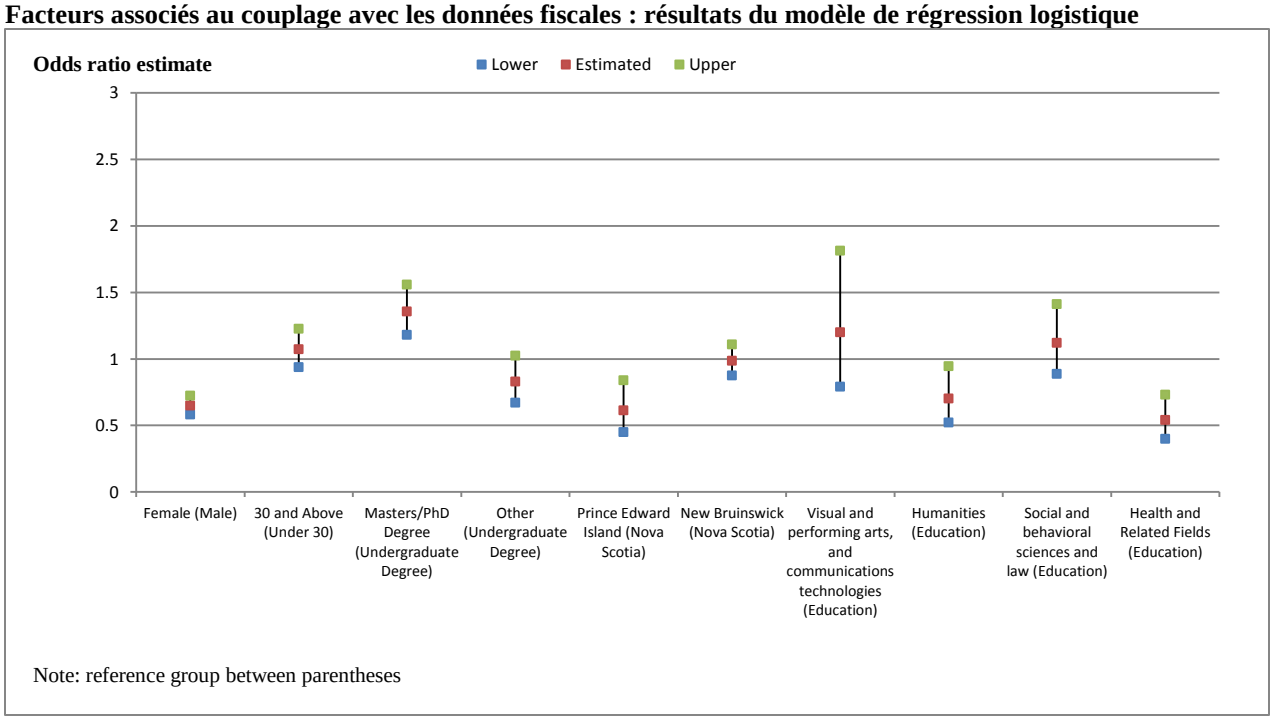
*significatif au seuil de 0.05

5. Indicateurs possibles des résultats des diplômés

Le couplage des fichiers du SIEP et de ceux du FFT1 permet d'étudier les progrès du revenu d'emploi au fil du temps pour une cohorte de diplômés. L'incidence d'un événement survenu à un moment donné peut aussi être mesurée en comparant les profils de transition. La mobilité géographique des diplômés sur le marché du travail peut aussi être suivie au moyen des données fiscales.

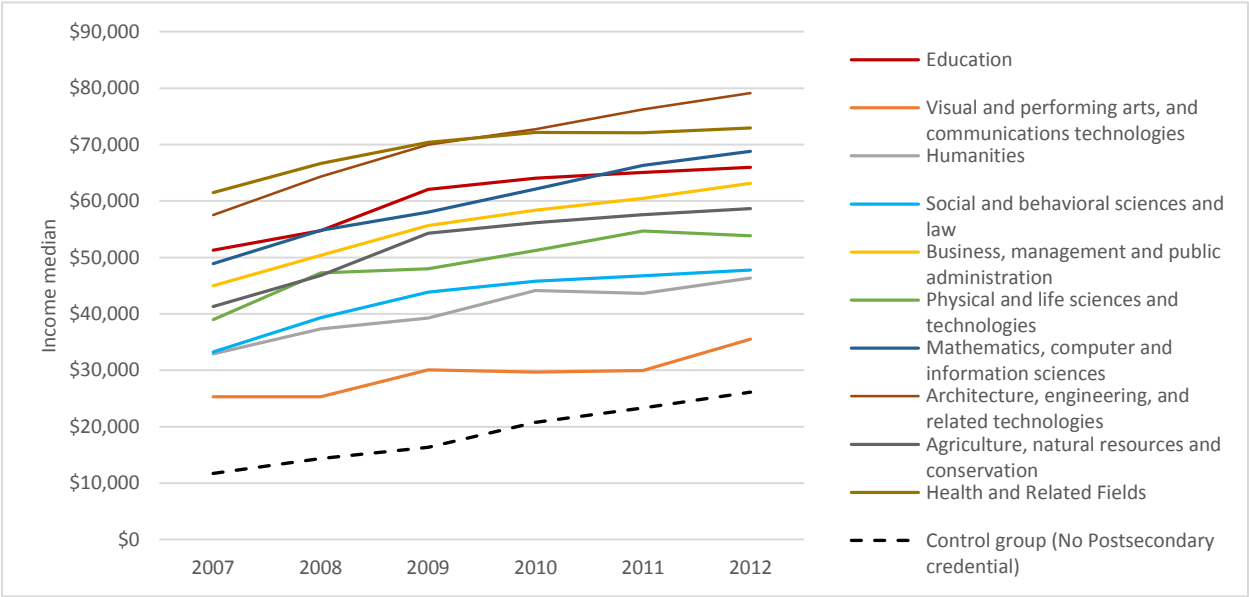
Le couplage des données représente une approche reconnue pour trouver des correspondances entre deux ensembles d'enregistrements, mais il peut donner lieu à des appariements inexacts (liens incorrects et enregistrements non couplés). De petites proportions de ces erreurs sont susceptibles de se produire entre des enregistrements se rapportant à la même personne, soit en raison de l'appariement de deux enregistrements n'appartenant pas à la même unité (couplages incorrects ou faux positifs), soit en raison d'une paire manquante d'enregistrements appartenant à la même unité (enregistrements non couplés ou paires manquées). Toutefois, l'objectif est de réduire ces cas dans la mesure du possible. Le défi additionnel consiste à déterminer, parmi les enregistrements non couplés, les faux négatifs et les vrais négatifs. Le graphique ci-après, qui comprend les résultats de la régression logistique, montre que le processus de couplage a créé une distorsion la cohorte initiale et pourrait avoir introduit un biais. On peut aussi explorer un rajustement pour tenir compte des faux négatifs.

Graphique 5.1
Facteurs associés au couplage avec les données fiscales : résultats du modèle de régression logistique



Note: reference group between parentheses

Graphique 5.2
Cohorte de 2006 : Médiane du revenu d'emploi des titulaires d'un baccalauréat et d'une maîtrise/d'un doctorat, selon le domaine d'études (regroupements principaux de la CPE)



Le graphique montre une modification de la tendance en 2009. Quelle est l'évolution du revenu d'emploi de la population longitudinale de 2006 entre 2007, 2009 et 2012? Il est possible de mesurer l'incidence d'un événement qui s'est produit en 2009, par exemple, la récession économique, en comparant les matrices de transition. Le tableau 5.1 fournit un test de l'égalité des deux matrices de transition.

Tableau 5.1
Changement de quartiles de revenu entre 2007 et 2009 et entre 2009 et 2011 pour les titulaires d'un titre de compétences dans le domaine de l'éducation

Statut au temps T	Temps	Statut au temps T+2			
		Le plus faible	Faible-moyen	Moyen-élevé	Le plus élevé
Le plus faible	T=2007	48,4 %	37,1 %	9,4 %	5,0 %
	T=2009	52,1 %	35,2 %	8,5 %	4,2 %
Faible-moyen	T=2007	18,8 %	66,1 %	13,8 %	1,3 %
	T=2009	23,5 %	65,8 %	10,3 %	0,4 %
Moyen-élevé	T=2007	13,5 %	18,0 %	58,6 %	9,8 %
	T=2009	7,2 %	13,8 %	61,8 %	17,1 %
Le plus élevé	T=2007	1,9 %	1,6 %	9,3 %	87,1 %
	T=2009	1,7 %	4,1 %	31,7 %	62,4 %

Après avoir effectué un test chi-carré de l'indépendance entre le temps et le post-résultat pour chaque niveau de pré-résultat et ajouté le chi-carré et les degrés de liberté respectifs, la différence semble statistiquement significative ($p < 0{,}00001$).

6. Sommaire et travaux à venir

Les données administratives longitudinales couplées comportent un potentiel d'analyse élevé et ont pour avantage d'avoir un faible coût comparativement aux enquêtes longitudinales, mais elles ne sont pas exemptes d'erreur. Le fait de traiter naïvement un fichier couplé comme s'il ne contenait pas d'erreurs mène, en général, à des estimations biaisées. Une approche de pondération sera mise à l'essai, afin de résoudre la question des dossiers non couplés pour l'analyse du revenu d'emploi.

Bibliographie

Ross, Theresa (2009), "Transition et progression : persévérance dans les études postsecondaires dans la région de l'Atlantique, données du SIEP", *Statistique Canada Catalogue no. 81-595-M-No.072,* Document de recherche, Statistique Canada.

Ross, Dejan (2014), "Analysis of long-term outcomes for university graduates in Information and Communication Technology programs", Initiative de recherche sur les politiques de l'éducation, Canada : Université d'Ottawa

Frenette, Marc and Yuri Ostrovsky. (2014), "Aperçus économiques – Les gains cumulatifs des diplômés postsecondaires sur 20 ans: résultats selon le domaine d'études", *Statistique Canada Catalogue no.11-626-X – No. 040,* Statistique Canada.

Saidi, Ntwari. (2014), "Weighting Adjustment for False Negatives in Record Linkage", article présenté à International Health Data Linkage Conference, Vancouver, Canada.