

Remaniement de la Base de données longitudinales sur les immigrants (BDIM)

Rose Evra¹

Résumé

La Base de données longitudinales sur les immigrants (BDIM) combine des données du Fichier des immigrants reçus (FIR) et des fichiers annuels de l'impôt. Ce couplage d'enregistrements est effectué au moyen d'une base de données de déclarants fiscaux. Le FIR comprend tous les immigrants qui ont été admis au Canada depuis 1980. En vue de remanier la BDIM, on a étudié la possibilité d'ajouter les résidents temporaires (RT) et les immigrants qui ont été admis entre 1952 et 1979 (PRE80). L'ajout de ces renseignements donnerait un aperçu plus complet de la population immigrante vivant au Canada. Afin d'intégrer les fichiers des RT et des PRE80 dans la BDIM, on a procédé à un couplage d'enregistrements entre ces deux fichiers et la base de données des déclarants fiscaux. Cet exercice a posé un défi, en partie en raison de la présence d'enregistrements en double dans les fichiers et de liens conflictuels entre les différents couplages d'enregistrements.

Mots-clés : couplage d'enregistrements, immigration, BDIM, fichier des immigrants reçus, résident temporaire, résident permanent

1. Introduction

1.1 Description du projet

La Base de données longitudinales sur les immigrants (BDIM) combine des données du Fichier des immigrants reçus (FIR) d'Immigration, Réfugiés et Citoyenneté Canada (IRCC) et des fichiers annuels de l'impôt, les fichiers sur la famille T1 (FFT1). Ce couplage d'enregistrements est effectué au moyen du Fichier de contrôle des couplages (FCC), une base de données qui couvre tous les déclarants fiscaux au Canada depuis 1981. Le FIR comprend tous les immigrants admis au Canada depuis 1980.

En vue de remanier la BDIM, on a étudié la possibilité d'ajouter les résidents temporaires (RT) et les immigrants admis entre 1952 et 1979 (PRE80). L'extension de la BDIM pour inclure les données sur les immigrants qui ont été admis avant 1980 donnerait un meilleur aperçu de la population immigrante vivant au Canada. En outre, en incluant les données des RT, on disposerait des antécédents de migration des RT qui deviennent des résidents permanents du Canada. On pourrait ainsi comparer l'intégration socioéconomique des immigrants ayant une expérience canadienne précédant l'admission et ceux n'ayant pas une telle expérience.

Pour intégrer les fichiers RT et PRE80 dans la BDIM, on a procédé à des couplages d'enregistrements entre ces deux fichiers et le FCC. En outre, un couplage administratif entre le fichier des résidents temporaires (FRT) et le FIR a été rendu possible grâce à un numéro d'identification de personne fourni par IRCC. Cet exercice a posé un défi, en partie en raison de la présence d'enregistrements en double dans les fichiers et de liens conflictuels entre les différents couplages d'enregistrements. Cet article examinera le processus et les défis que pose le remaniement de la BDIM grâce à ces deux fichiers.

1.2 Description de la Base de données longitudinales sur les immigrants (BDIM)²

La BDIM est une source exhaustive de données sur le comportement économique de la population immigrante de déclarants fiscaux au Canada. Un immigrant désigne une personne qui est ou a déjà été un résident permanent (RP).

¹Rose Evra, Statistique Canada, 100, promenade Tunney's Pasture, Canada, K1A 0T6 (rose.evra@canada.ca).

²Pour plus de détails sur la BDIM, voir la page Web :

Cette personne a obtenu des services d'immigration le droit de vivre en permanence au Canada. La BDIM permet l'analyse des résultats économiques de différentes catégories d'immigrants au fil du temps, afin d'évaluer les répercussions des caractéristiques des immigrants au moment de l'admission, comme le niveau de scolarité et la connaissance du français ou de l'anglais, sur leur établissement.

La BDIM combine un fichier administratif sur les immigrants reçus au Fichier sur la famille T1 (T1FF), au moyen de techniques d'appariement exact pour le couplage d'enregistrements. Le couplage est effectué à partir de variables comme le nom, la date de naissance, le sexe et les dates d'admission. Le taux de couplage global du FIR au FCC est d'environ 87 % (Brennan, 2014).

Chaque année, la BDIM est mise à jour grâce à l'ajout d'une nouvelle cohorte d'immigrants reçus et de données fiscales correspondant à toutes les cohortes visées par la BDIM. Pour chaque nouvelle année d'imposition, on ajoute également de nouveaux entrants provenant de toutes les cohortes d'immigrants reçus – et non pas seulement de la nouvelle cohorte - qui ont produit une déclaration de revenus (ou sont appariés) une première fois.

Le Fichier des immigrants reçus (FIR) représente un recensement de tous les immigrants admis depuis 1980. Il comprend des renseignements sur leurs caractéristiques lorsqu'ils sont devenus des résidents permanents (RP). En date de 2013, le FIR comptait plus de 6,8 millions d'enregistrements. Chaque année, la taille du FIR augmente en fonction du nombre de nouveaux résidents permanents (environ 250 000). Le FIR comporte des variables comme la catégorie d'admission des immigrants, la date d'admission et le pays de naissance. Chaque immigrant se voit attribuer un numéro d'identification de personne, qui demeure inchangé au fil du temps, et un numéro de permis. Ce fichier comprend des renseignements sur les RP au moment de l'admission, mais on ne recueille pas de données sur leur expérience au Canada avant l'admission (p. ex., le fait d'avoir obtenu au préalable un permis de travail ou d'études comme RT). Les décès et les émigrations ne sont pas identifiés, étant donné qu'aucun suivi n'est assuré après l'admission. Par conséquent, un pourcentage des enregistrements du FIR ne font plus partie de la population vivant au Canada.

2. Amélioration de la BDIM

2.1 Avantages de l'amélioration de la BDIM

Selon l'Enquête nationale auprès des ménages (ENM) de 2011, on estime qu'en 2011 près de 70 % de la population immigrante du Canada avait été admise après 1980, ce qui fait que la BDIM ne couvre pas l'ensemble de la population immigrante du Canada. Afin d'obtenir une meilleure couverture de la population immigrante du Canada, des données sur les immigrants antérieures à 1980 sont requises. Selon l'ENM, en 2011, 98 % des immigrants avaient été admis au Canada à partir de 1952.

Grâce aux renseignements concernant les RT, il serait possible d'établir des liens entre le FRP et le FRT, ce qui devrait élargir la valeur analytique de la BDIM. Il serait possible, par exemple, de comparer l'intégration socioéconomique des immigrants ayant une expérience précédant l'établissement au Canada par rapport à ceux n'ayant pas une telle expérience.

2.2 Résidents temporaires (RT)

Plus de 7,5 millions de personnes ont obtenu un permis temporaire pour demeurer légalement au Canada, entre 1980 et 2013. Une base de données couvrant la population des RT et comprenant des renseignements, comme la date d'arrivée, des détails sur les permis et un numéro d'identification de personne (NIP), a été utilisée pour ce projet. Ce NIP devrait être le même que celui présent dans le FIR pour les RT qui ont obtenu le statut d'immigrant reçu. Dans certains cas, le NIP diffère entre les deux fichiers parce qu'il n'est pas collecté ou qu'ils ont été perdu.

Après un couplage d'enregistrements, on a estimé que le FRT incluait environ 400 000 doubles, en raison principalement de problèmes de qualité des données à la fin des années 1980 (Evra, 2015). Pour atténuer cette perte d'information, des numéros d'identification de personne fictifs facilement identifiables ont été créés. Afin de d'identifier les doubles et de les supprimer, un couplage d'enregistrements du fichier des RT avec lui-même a été effectué.

Pour évaluer les avantages de l'ajout des données sur les RT dans la BDIM, le fichier a été couplé au fichier de contrôle des couplages (FCC), qui représente une base de données sur les déclarants fiscaux postérieurs à 1981. On estime que 39 % des enregistrements du fichier ont été couplés. Le faible taux de couplage est principalement attribuable au grand nombre de RT qui ne produisent pas des déclarations de revenus (Brennan, 2014). Certains d'entre eux sont au pays pour une courte période seulement. Il a été déterminé que 19,5 % des RT de ce fichier sont devenus des RP entre 1980 et 2012 (Barayandema, 2015). Il serait possible d'ajouter des données antérieures à l'admission pour 1,7 million d'enregistrements du FIR, qui représentent une personne sur quatre du FIR, comme le démontrera la section 3.3.

2.3 Résidents permanents avant 1980 (PRE80)

Un fichier de données administratives couvrant 4,1 millions de RP qui se sont établis au Canada au cours de la période de 1952 à 1979 a été utilisé pour ce projet. Certaines variables pertinentes, comme la date d'admission et le pays de naissance, sont présentes. Toutefois, la qualité de l'information disponible pour ces immigrants est plus faible que dans le FIR et varie selon l'année d'admission. Par exemple, entre 1962 et 1972, seule l'année de naissance était disponible, ce qui a fait diminuer le taux de couplage pour les personnes admises au cours de cette période (Brennan, 2014). Certaines variables, comme la catégorie d'admission des immigrants, ne sont pas incluses. Le fichier comporte aussi certaines limites similaires au FIR, par exemple, l'absence de données postérieures à l'admission concernant les décès et l'émigration. La proportion de décès devrait être plus élevée dans ce fichier que dans le FIR, en raison de la population plus âgée.

Afin d'évaluer les avantages de l'ajout des données sur les immigrants antérieures à 1980 dans la BDIM, le fichier a été couplé au fichier de contrôle des couplages. Une proportion de 43 % des enregistrements du fichier a été couplée au FCC (Evra, 2015). En outre, pour l'année d'imposition 2012, environ 1 million d'immigrants arrivés avant 1980 continuaient de produire des déclarations de revenus et, au début des années 1980, les immigrants arrivés avant 1980 représentaient la grande majorité des déclarants fiscaux immigrants. La représentation en pourcentage des personnes ayant obtenu le droit d'admission avant 1980 va d'environ 85 % en 1982 à environ 19 % en 2012 (Brennan, 2014). L'ajout de ces données à la BDIM améliorerait considérablement la couverture de la population immigrante, particulièrement dans les années 1980.

3. Traitement

3.1 Objectif

Suite aux études de faisabilité, il a été décidé de remanier la BDIM en ajoutant les résidents temporaires et les immigrants établis avant 1980 à la base de données. Cela permettra la création d'une base de données sur l'immigration plus exhaustive comportant plus de variables détaillées sur l'immigration. Par exemple, les fichiers fiscaux originaux de la BDIM comprennent le nombre d'immigrants dans une famille, mais uniquement ceux admis après 1980; ce chiffre comprendra maintenant les résidents temporaires et les immigrants établis avant 1980. Il sera aussi possible d'identifier le conjoint d'un déclarant fiscal immigrant comme résident temporaire ou comme personne qui a obtenu le statut d'immigrant reçu avant 1980.

L'objectif de ce projet était de créer un fichier de personnes (figure 3.1-2) comprenant tous les liens entre le fait de devenir un résident permanent avant ou après 1980 et le fait d'avoir été un résident temporaire (RT), de produire des dossiers fiscaux de 1982 à 2013 pour ces immigrants et RT, et de créer un fichier de permis pour les RT. Celui-ci comprendrait des données sur les résidents temporaires au niveau du permis (p. ex., date d'entrée en vigueur, type de permis), tandis que le fichier de personnes comprendrait des détails, comme le nombre de permis temporaires (p. ex., 2 permis de travail et 1 permis d'étudiant). Les difficultés liées à ce processus sont décrites dans la section suivante (figure 3.1-1).

Figure 3.1-1
Illustration du traitement

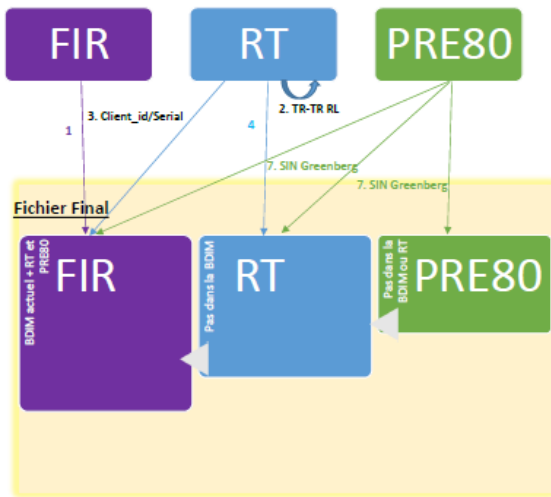
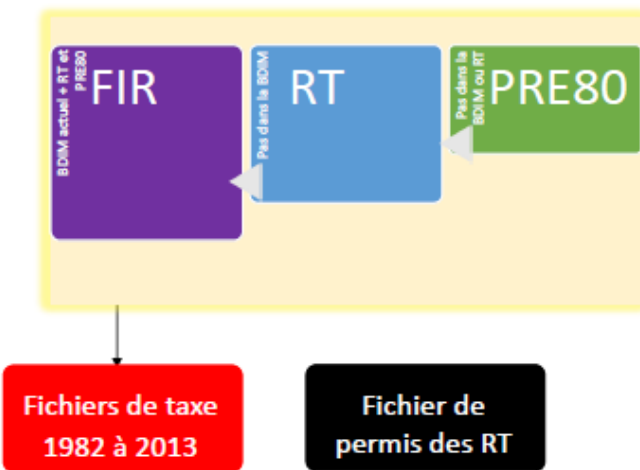


Figure 3.1-2
Produit final



3.2 Défis

Comme il est mentionné précédemment dans le présent document, le couplage d'enregistrements initial comprenait le couplage du FIR au FCC (base de données des déclarants fiscaux), afin d'identifier les déclarants fiscaux immigrants. De même, le FRT et le PRE80F ont été couplés au FCC séparément, après le couplage d'enregistrements initial. Ces couplages d'enregistrements ont permis la production de fichiers comprenant des clés de couplage permettant de joindre les fichiers de l'immigration (FIR, FRT et PRE80F) aux dossiers fiscaux, avec le numéro d'identification fiscal (NIF) et le numéro d'identification de personne comme clés. Le FRT a par la suite été couplé avec lui-même pour identifier les doubles. Par ailleurs, le FRT a été couplé au FIR au moyen du numéro d'identification de personne (NIP) et du numéro de permis d'établissement, qui sont communs aux deux fichiers.

Le principal défi était la nécessité de combiner les résultats des quatre couplages d'enregistrements, en vue de produire le fichier de personnes. Ces couplages d'enregistrements ont été exécutés séparément à différents moments, ce qui a donné lieu à des résultats contradictoires. Par exemple, dans la figure 3.2-1, la première personne avait deux NIP différents. On ne pouvait que coupler les enregistrements au moyen du NIF. La deuxième personne a été couplée uniquement au moyen du numéro d'identification de personne, et le couplage d'enregistrements du FIR au FCC n'a pas produit de lien, contrairement au couplage du FRT et du FCC.

Tableau 3.2-1
Exemples fictifs de résultats du couplage d'enregistrements

Personne	NIP FIR	NIP FRT	NIF FIR	NIF FRT
1	9012345	9012346	12345678	12345678
2	4567890	4567890	Non couplé	23456789
3	5678901	5678901	34567890	Non couplé
4	6789012	6789012	78901234	89012345

Ces différences sont le résultat de nombreux facteurs. L'erreur de couplage explique une partie des disparités. La qualité des champs de couplage diffèrait d'un fichier à l'autre. Par ailleurs, les outils et les méthodes de couplage d'enregistrements évoluent avec le temps. Les erreurs de mesure peuvent découler d'erreurs de saisie des données ou d'erreurs des répondants. Par exemple, la date de naissance peut différer entre le FRT et le FIR, entraînant un couplage avec des déclarants fiscaux différents pour une personne donnée. À noter que seulement une petite proportion des enregistrements a donné lieu à des résultats contradictoires, chacune des incohérences s'appliquant à moins de 1 % des enregistrements inclus dans la BDIM remaniée (Evra, 2015).

Une partie de la variation dans les données s'explique par le temps écoulé entre la collecte de l'information comprise dans les différents fichiers. Par exemple, le numéro d'identification fiscal (NIF) peut changer au fil du temps. De nombreux immigrants obtiennent un NIF temporaire puis, par la suite, un NIF permanent. Environ 20 % des résidents permanents, 44 % des résidents temporaires et 5 % des résidents permanents avant 1980 avaient des NIF multiples (Brennan, 2014). Dans certains cas, les immigrants changent de numéro d'identification de personne et ne peuvent être trouvés sans un couplage d'enregistrements au moyen d'autres variables (figure 3.2-1 : personne 1).

Une fois que le couplage des enregistrements du fichier des résidents temporaires avec lui-même a été complété, on a découvert que quelques RT en double étaient couplés à des déclarants fiscaux différents (moins de 1 000 enregistrements; Evra, 2015), selon les résultats des autres couplages d'enregistrements (FRT-FCC et FIR-FCC). Les résultats du couplage des enregistrements FIR-FCC ont été considérés comme étant de meilleure qualité, et toutes les différences donnant lieu à un changement dans ces résultats ont été ignorées.

Lorsqu'on a couplé le FIR au FRT au moyen des numéros d'identification de personne, certains cas ont été couplés à un numéro d'identification fiscal (NIF) différent, comme pour la personne 4 dans le tableau 3.2-1. Cela pourrait venir d'une erreur de couplage ou du fait que la personne avait plusieurs NIF. Lorsque cette situation s'est produite, les résultats du couplage FIR-FCC ont été conservés. Pour résoudre le problème de NIF multiples, la base de données des NIF a servi à identifier tous les NIF liés à une personne donnée.

Certains enregistrements comportant le même NIF avaient des numéros d'identification de personne différents. Cela pourrait être attribuable à une erreur de couplage ou à un changement valide dans le numéro d'identification de personne.

Certaines des variables sociodémographiques différaient entre le FIR, le FRT et le PRE80F. Par exemple, le pays de naissance pouvait différer. Les données du FIR ont été conservées lorsqu'elles étaient disponibles, suivies par celles du FRT et du PRE80. Cette décision a été prise en fonction de la qualité des données comprises dans les différents fichiers.

3.3 Résultats

Le résultat final de ce projet est un fichier de personnes comprenant plus de 16,5 millions d'enregistrements et des dossiers fiscaux de 1982 à 2013, y compris des données sur les personnes qui sont devenues des résidents permanents (de 1952 à 2013) et des résidents temporaires (de 1980 à 2013); voir le tableau 3.3-1. Plus de 8,9 millions d'enregistrements ont été identifiés comme appartenant à des personnes qui ont produit des déclarations de revenus au moins une fois de 1982 à 2013. Selon les résultats de ce projet, 1,7 million d'immigrants étaient des résidents temporaires avant de devenir des résidents permanents; parmi eux, 94 % ont été couplés au FCC.

Dans le cas des enregistrements figurant dans le FIR, le FRT et le PRE80F, le taux de couplage est de 100 % parce que le numéro d'identification fiscal est la seule variable disponible pour relier les trois. La catégorie PRE80 seulement comporte un taux de couplage de 43 %. À partir des résultats d'un deuxième couplage, entre le FCC et la Base canadienne de données sur la mortalité (BCDM), il a été établi que plus de 25 % des immigrants couplés d'avant 1980 étaient décédés en 2013. La proportion de personnes décédées est peut-être plus élevée pour les enregistrements non couplés, parce que les données remontent à 1952. Une personne qui s'est établie au Canada en 1952, à l'âge de 25 ans, aurait 86 ans en 2013.

Les enregistrements qui figurent seulement dans le FRT comportent un taux de couplage de 22 %. Ce taux de couplage peut être expliqué par le fait que 2,2 millions des RT avaient des permis de visiteur seulement (les autres RT pouvaient être demeurés au pays pendant une courte période). De nombreux RT ne produisent pas de formulaires T1 (Brennan, 2014).

Tableau 3.3-1
Répartition des enregistrements après le traitement

Type d'enregistrement	Enregistrements	Couplés aux données fiscales	% couplés
FIR-PRE80-RT	4500	4500	100 %
FIR-RT	1707300	1599800	94 %
FIR seulement	5175900	4385300	85 %
FIR-PRE80	12200	12200	100 %
RT-PRE80	6400	6400	100 %
RT seulement	5601700	1225600	22 %
PRE80 seulement	4091700	1745800	43 %
Total	16599700	8979500	54 %

4. Conclusion

La Base de données longitudinales sur les immigrants pourrait être remaniée en incluant des données sur les résidents temporaires et les immigrants d'avant 1980. Cela donnerait lieu à de nouvelles capacités d'analyse à partir de la BDIM et à une plus grande couverture de la population immigrante du Canada.

Après avoir résolu plusieurs problèmes liés à l'union des résultats de plusieurs couplages pour produire un fichier, une stratégie en vue de combiner le contenu des trois ensembles de données a été élaborée. À l'avenir, la qualité de l'ensemble de données devra être validée, et des documents de référence devront être élaborés. L'évaluation de la qualité devrait comprendre une évaluation des incohérences, par exemple, les personnes qui ont seulement des permis de visiteur et qui produisent des déclarations de revenus et les déclarants préadmission qui ne sont pas inclus dans le fichier des RT. En outre, la date du décès sera ajoutée dans la BDIM à partir des résultats d'un autre couplage d'enregistrements. Pour les années subséquentes, un couplage d'enregistrements simple, combinant les données sur les RT et le RP, devrait être effectué. Le contenu du fichier final devra être confirmé.

Remerciements

J'aimerais souligner la contribution des méthodologistes qui ont procédé aux couplages d'enregistrements – James Brennan, Alexandr Diaz-Papkovich, Martin Lessard et Émilie Mayer – ainsi que de mes collègues de la Division de la statistique sociale et autochtone (DSSEA), Scott McLeish, Athanase Barayandema et Ian Marrs.

Bibliographie

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

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

Brennan, J. (2014) “Landed Immigrants 1952 to 1979 linkage to Tax”, document non publié, Statistique Canada

Brennan, J. (2014) “Temporary Residents Linkage to T1 tax”, document non publié, Statistique Canada

Evra, R. (2015) “Feasibility Study: Redesign of the IMDB”, document non publié, Statistique Canada

Evra, R. (2015) “Summary of processing steps”, document non publié, Statistique Canada

Statistique Canada (2015) “Base de données longitudinales sur l’immigration (BDIM)”. Site internet: http://www23.statcan.gc.ca/imdb/p2SV_f.pl?Function=getSurvey&SDDS=5057