Request for Proposals
Research Data Centre Program
Statistics Canada
Longitudinal Administrative Databank (LAD)
Fall 2015

A joint working group from Statistics Canada and the Canadian Revenue Agency was created in 2011 to address the issue of the availability of CRA tax filer information in Statistics Canada's Research Data Centres. The benefits of increased access to and analysis of LAD included:

- Support for evidence based policy making, through the provision of data access
- Increased use of administrative data sources for analysis and research thereby reducing the cost of data collection through subsequent surveys
- Provision of a common research database will reduce the need for parallel database construction exercises, and
- Making full use of information provided by Canadians will reduce the burden on respondents for surveys

At that time, researchers external to Statistics Canada had limited access to LAD through client requested cost-recoverable custom tabulations or as a deemed employee at Statistics Canada's head office in Ottawa. It was the decision of the joint working group that a two year LAD pilot project be conducted at the Federal Research Data Center (FRDC) to address the feasibility of supporting the use of the LAD in all Research Data Centres.

In the fall of 2012, a group of federal government researchers was invited to submit research proposals and four were accepted. After the successful completion of the pilot in December 2014, Statistics Canada is now conducting a gradual roll-out of the LAD to the RDCs. This will allow further testing of the confidentiality vetting rules and testing of the IT resources necessary to transfer and analyze these large data files.

The Data

Longitudinal Administrative Databank (LAD)

The LAD is a sample of individual taxfilers with a longitudinal design. Currently data are available from 1982-2012 and 2013 will be available early in the roll-out. The frame is constructed from the annual T1 Family File (Annual Estimates for Census Families and Individuals (T1 Family File)) which makes use of information from administrative files. Only individual records that have social insurance numbers can be selected for the LAD and these are sampled at a 20% rate. Also included in the LAD are a set of immigration variables, drawn from the Longitudinal Immigration Data Base (IMDB), relating to information collected at landing, as well as a set of variables describing Tax Free Saving Account usage.

The LAD survey units are individuals but limited information about the characteristics of their family during the reference year is also kept (e.g. spouse/parent, family, and children). No stratification is performed consequently the sampling weight is equal across all units. The sampling is done once on

each record in such a way that if someone is selected in a particular reference year, they will be selected in any other later (or earlier) years in which they are present in the T1 Family File.

Researchers unfamiliar with administrative tax data are cautioned that not all LAD data are internally or externally coherent, in part, because tax data are not subject to the same edit and imputation procedures as survey data. Consequently, many researchers have found it takes some time to become familiar with the LAD and to be able to operationalize it in their research.

Submissions

While a limited number of proposals for both cross sectional and longitudinal analyses will be considered, research that includes the following types of analyses or results is of particular interest:

- Proposals making use of the longitudinal aspects of the databank
- Proposals making use of an individual taxfiler as the unit of analysis are preferable to family based analyses. However, individuals can be characterized by certain characteristics of their family or spouses
- A mix of SAS users and STATA users.

Researchers are invited to submit proposals for consideration by **October 5th, 2015**. The submitted research proposals will be assessed based on the research areas of interest as well as the proposed types of analysis, including the viability of the proposed research. The proposals that are better aligned with the above described research areas of interest, types of analysis and results will receive more favorable consideration.

All researchers will be notified by **November 6th, 2015**. The researchers whose proposals have been accepted *should* be able to access the data by late November.

We will be testing the robustness of the vetting rules, over the next year, with these proposals. Therefore, vetting could be delayed. Researchers should bear this in mind when considering the appropriateness of applying for access to LAD, at this time, because of the potential impact on the timely completion of their research.

Proposals should be submitted by October 5th, 2015 to:
Donna Dosman, Chief
Research Data Centre Program
Donna.Dosman@statcan.gc.ca

Demande de propositions
Programme des centres de données de recherche
Statistique Canada
Banque de données administratives longitudinales
Automne 2015

Un groupe de travail conjoint de Statistique Canada et de l'Agence du revenu du Canada a été formé en 2011 pour se pencher sur la question de la disponibilité des renseignements sur les déclarants de l'Agence du revenu du Canada dans les centres de données de recherche de Statistique Canada. Voici quelques avantages qui ressortaient d'un accès accru à la Banque de données administratives longitudinales et de l'analyse de celle-ci :

- appuyer l'élaboration de politiques fondée sur des données probantes, en fournissant un accès aux données;
- accroître l'utilisation des sources de données administratives aux fins d'analyse et de recherche, et ainsi réduire le coût de la collecte de données par des enquêtes subséquentes;
- fournir une base de données de recherche commune pour réduire la nécessité de procéder à des exercices de construction de bases de données parallèles;
- utiliser pleinement les renseignements fournis par la population canadienne pour réduire le fardeau des répondants lié aux enquêtes.

À l'époque, les chercheurs de l'extérieur de Statistique Canada avaient un accès limité à la Banque de données administratives longitudinales par l'intermédiaire de tableaux personnalisés à recouvrement des coûts, produits à la demande du client, ou à titre de personne réputée être employée au bureau central de Statistique Canada à Ottawa. C'est le groupe de travail conjoint qui a décidé qu'un projet pilote de deux ans concernant la Banque de données administratives longitudinales serait mené au Centre fédéral de données de recherche pour se pencher sur la faisabilité d'appuyer l'utilisation de la Banque de données administratives longitudinales dans l'ensemble des centres de données de recherche.

À l'automne 2012, un groupe de chercheurs du gouvernement fédéral ont été invités à soumettre des propositions de recherche, et quatre d'entre elles ont été acceptées. Par suite de l'achèvement réussi du projet pilote en décembre 2014, Statistique Canada procède actuellement au déploiement graduel de la Banque de données administratives longitudinales dans les centres de données de recherche. Il sera alors possible de faire d'autres essais quant aux règles de contrôle de la confidentialité et de mettre à l'essai les ressources de technologie de l'information nécessaires au transfert et à l'analyse de ces volumineux fichiers de données.

Les données

Banque de données administratives longitudinales

La Banque de données administratives longitudinales présente un échantillon des déclarants selon un plan longitudinal. Actuellement, les données sont disponibles pour la période de 1982 à 2012, et les

données de 2013 seront disponibles au début du déploiement. La base de sondage provient du Fichier des familles T1 annuel (Estimations annuelles du revenu des familles de recensement et des particuliers (Fichier des familles T1)), qui utilise les renseignements des fichiers administratifs. Seuls les enregistrements individuels comportant un numéro d'assurance sociale peuvent être sélectionnés aux fins d'intégration dans la Banque de données administratives longitudinales, et l'échantillonnage est fait à un taux de 20 %. La Banque de données administratives longitudinales comprend aussi un ensemble de variables liées à l'immigration, tirées de la Base de données longitudinales sur les immigrants, qui se rapportent aux renseignements recueillis lors de l'établissement, ainsi qu'un ensemble de variables décrivant l'utilisation du compte d'épargne libre d'impôt.

Les unités d'enquête de la Banque de données administratives longitudinales sont les particuliers, mais des renseignements limités sur les caractéristiques de leur famille pendant l'année de référence sont aussi conservés (p. ex. conjoint/conjointe ou parent, famille et enfants). Aucune stratification n'est effectuée, donc le poids d'échantillonnage est égal parmi toutes les unités. L'échantillonnage est fait une fois dans chaque enregistrement de façon que si une personne est sélectionnée dans une année de référence donnée, elle sera sélectionnée toutes les années suivantes (ou précédentes) où elle figure dans le Fichier des familles T1.

Les chercheurs qui connaissent peu les données fiscales administratives sont mis en garde : les données de la Banque de données administratives longitudinales ne présentent pas toutes une cohérence interne et externe, en partie parce que les données fiscales ne sont pas assujetties aux mêmes procédures de contrôle et d'imputation que les données d'enquête. Par conséquent, de nombreux chercheurs ont constaté qu'il faut un certain temps avant de se familiariser avec la Banque de données administratives longitudinales et d'être en mesure de l'intégrer de façon opérationnelle à leurs recherches.

Soumissions

Même si un nombre limité de propositions pour des analyses transversales et longitudinales seront prises en compte, les recherches comportant les types d'analyses ou de résultats suivants revêtent un intérêt particulier :

- propositions utilisant les aspects longitudinaux de la banque de données;
- propositions utilisant un déclarant individuel comme unité d'analyse, qui sont préférables aux analyses fondées sur la famille. Toutefois, les particuliers peuvent être caractérisés par certaines caractéristiques de leur famille ou de leur conjoint/conjointe;
- combinaison d'utilisateurs de SAS et d'utilisateurs de STATA.

Les chercheurs sont priés de soumettre leurs propositions aux fins d'examen au plus tard le **5 octobre 2015**. Les propositions de recherche soumises seront évaluées en fonction des domaines de recherche d'intérêt ainsi que des types d'analyses proposés, y compris la viabilité de la recherche proposée. Les propositions qui cadrent le mieux avec les domaines de recherche d'intérêt et les types d'analyses et de présentation de résultats décrits ci-dessus seront pris en compte plus favorablement.

Tous les chercheurs seront avisés au plus tard le **6 novembre 2015**. Les chercheurs dont les propositions auront été acceptées *devraient* être en mesure d'accéder aux données d'ici la fin de novembre.

Nous mettrons à l'essai la robustesse des règles de contrôle de la confidentialité au cours de la prochaine année à l'aide de ces propositions. Par conséquent, il se pourrait que le contrôle soit retardé. Les chercheurs devraient garder cela à l'esprit lorsqu'ils détermineront s'il est pertinent pour eux de demander l'accès à la Banque de données administratives longitudinales en ce moment, puisque cela pourrait avoir une incidence sur le moment où ils pourront achever leurs recherches.

Les propositions doivent être soumises au plus tard le 5 octobre 2015 à :
Donna Dosman, chef
Programme des centres de données de recherche
Donna.Dosman@statcan.gc.ca