Business microdata for economic research

www.statcan.gc.ca

Danny Leung
Economic Analysis Division

March 24, 2017
Outline

• Accessing business microdata for research purposes at the Canadian Centre for Data Development and Economic Research (CDER) at Statistics Canada
  • CDER basics
  • Data sets available for access at CDER
  • Application process
  • Future directions
  • Other information
CDER background

- CDER was created to allow Statistics Canada to make better use of its business data holdings without compromising security
- Set up at Statistics Canada HQ in Ottawa and launched in June 2011 for federal government researchers
- Access to CDER extended to academic and non-federal government researchers in October 2012
- Currently, there are roughly 80 projects in progress
CDER activities

- Provides analysts with secure access to business micro data for *research-oriented projects* that serve the mandate of Statistics Canada
- Serves as a repository for business microdata
- Leads the development of new business micro data
Key information

- CDER approval process is similar to that of the RDCs
- Access is provided at Statistics Canada’s headquarters in Ottawa
- Researchers must cover the full cost of their project
  - Project costs payable to Statistics Canada start at $7,200
- In the past, partners that have covered some of the researcher costs
  - Global Affairs Canada
  - Innovation, Science and Economic Development Canada
  - Environment and Climate Change Canada – Economic and Environmental Policy Research Network
  - SSHRC partnership grant – Firms, Productivity and Incomes
  - Economic Analysis Division – Research Affiliate Program
Data available at CDER

1) Stand-alone, research-ready data already in use
   • Examples: Survey of Innovation and Business Strategies; T2 Corporate Income Tax; T2-Longitudinal Employment Analysis Program; Annual Survey of Manufactures; Survey of Financing and Growth of SMEs linked to tax data; Customs database

2) Linkable File Environment (LFE)
   • Specific variables from a set of files where linkages have been done, but files are so large that extractions are made upon request

3) Developmental datasets and other linkage environments
   • Analytical databases containing derived variables for specific analyses (e.g., National Accounts Longitudinal Micro data File); additions to LFE; other linkage environments (e.g., Canadian Employer-Employee Database); new stand-alone data
Stand-alone databases

- Survey of Financing and Growth of SMEs
  - Linked to administrative data on firm performance before and after survey years, 2000 to 2015
  - Use of financing during start up
  - Requests for financing (term loans, mortgages, lines of credit, credit cards, government loans, equity) and outcomes (approved/rejected, collateral, term, interest rate, amounts requested/received)
  - Business information (exports, R&D, innovation, IP use, plans for growth, public procurement participation)
  - Owner information (age, education, experience, country of birth, language of primary decision maker; % female, % aboriginal, % visible minority)
Stand-alone databases (2)

- Annual Survey of Manufactures (1961-2012) – series of longitudinal datasets, that have been used for research on trade, innovation, productivity
  - Cross-sectional, establishment level survey of manufacturing
  - Principal industrial statistics (revenue, employment, payroll, cost of materials, cost of energy, water usage, inventories, exports, etc) and commodity file…100s of variables in the latest database
  - Essentially a census up to 2012, where administrative records have been used for small units; and cutoffs by province, industry and revenue size has changed over time
  - Post-2012, ASM data are survey data with a take all and take some portion
  - Although micro data exist back to 1961, various longitudinal data bases have been constructed with the help of industry and identifier concordances:
    - 1990-2010 (NAICS, standard ASM variables)
    - 2000-2012 (NAICS, UES ASML variables)
Stand-alone databases (3)

- ASM has been linked to other sources
  - National Pollutant Release Inventory (NPRI) and Greenhouse Gas Reporting Protocol (GHGRP), plant level, 2000 to 2012
    - NPRI: Canada’s legislated, publicly accessible inventory of pollutant releases (air, water and land), disposals and transfers for recycling, 300+ pollutants (criteria air contaminants, heavy metals and toxins)
    - GHGRP: carbon dioxide, methane, nitrous oxide, sulphur hexafluoride, nitrogen trifluoride, various hydrofluorocarbons, and various perfluorocarbons
  - General index of financial information (GIFI), ASM-enterprise level, 2000 to 2012
    - tangible capital stock by ASM-enterprise
  - Research and Development in Canadian Industries, ASM-enterprise level, 2000 to 2009
    - Intramural and extramural R&D expenditures at enterprise level
    - Other variables may be available
### B3 Form

**Canada Customs Coding Form**

**Douanes Canada - Formule de Codage**

1. **Importer Name and Address**
   - **Nom et adresse de l'importateur**
   - **No. - N°**

2. **Transaction No. - N° de transaction**
   - **No. - N°**

3. **Type**
   - **Type**
   - **Office No. N° de bureau**

4. **Gst Registration No. - N° de TPS**
   - **N° de TPS**

5. **Payment Code - Code de paiement**
   - **Code de paiement**

6. **Mode of Transport - Mode de transport**
   - **Mode de transport**

7. **Port of Unloading - Port de débarquement**
   - **Port de débarquement**

8. **Total Weight - Total de la vo**
   - **Total de la vo**

9. **Vendor Name - Nom du Vendeur**
   - **Nom du Vendeur**
   - **No. - N°**

10. **Country of Origin - Pays d'origine**
    - **Pays d'origine**

11. **Place of Export - Lieu d'exportation**
    - **Lieu d'exportation**

12. **Tariff Treatment - Traitement tarifaire**
    - **Traitement tarifaire**

13. **Cycy Code - Code de devise**
    - **Code de devise**

14. **Time Limit - Délai**
    - **Délai**

15. **Freight - Fret**
    - **Fret**

16. **Direct shipment date - Date d'expédition directe**
    - **Date d'expédition directe**

17. **Release Date - Date de la mainlevée**
    - **Date de la mainlevée**

18. **Weight (kgm) - Poids (kgm)**
    - **Poids (kgm)**

19. **Previous Transaction - Transaction antérieure**
    - **Transaction antérieure**

20. **Special Authority - Autorisation spéciale**
    - **Autorisation spéciale**

21. **Line - Ligne**
    - **Ligne**

22. **Classification No. - N° de classement**
    - **N° de classement**

23. **Tariff Code - Code tarifaire**
    - **Code tarifaire**

24. **Quantity - Quantité**
    - **Quantité**

25. **Value for Duty - Valeur en douane**
    - **Valeur en douane**

26. **Value for Currency Conversion - Valeur pour change**
    - **Valeur pour change**

27. **Value for Tax - Valeur pour tax**
    - **Valeur pour tax**

28. **Gst - Tps**
    - **Tps**

---

**Canada Border Services Agency**

**Agence des services frontaliers du Canada**

**STATISTICS CANADA • STATISTIQUE CANADA**

**PROTECTED (WHEN COMPLETED) PROTEGE (UNE FOIS REMPLI)**

**RESERVED FOR CESA USE**

**RESERVE À L'USAGE DE L’ASFC**
Stand-alone databases (4)

- ASM linked to import data, ASM enterprise-level, 2002-2011
  - Import data includes: import value by HS-10 commodity classification and country of origin
  - Linkage up to 2012 possible
- Canadian Border Service Agency Customs Database, Business Number level, July 2002 to June 2008
  - Similar source to import data, B3 customs form
  - Transactions-level file – value, country of export, country/state of origin, HS-10 commodity classification, currency of transaction, unit values, limited information on identity of exporter
  - Can be linked to other Canada Revenue Agency data at the BN level, e.g., GIFI or T2 corporate income data
Stand-alone databases (5)

• Trade by Exporter Characteristics (TEC)
  • Enterprise-level, 2010 to 2015
  • Industry, province, CMA, employment of enterprise
  • Value of shipments by HS-8 commodity codes and country of destination

• Longitudinal Employment Analysis Program
  • Firm entry and exit, job creation and job destruction, and payroll
  • Labour tracking
  • 2001 to 2014 vintages covering 1983 to 2014
  • T2-LEAP – LEAP linked to core administrative data (including capital investment program) from the corporate tax system covering 1983 to 2014; 1997, 2004, 2007, 2008 to 2014 vintages
Stand-alone databases (6)

- Survey of Innovation and Business Strategies
  - Cross-sectional survey in 2009 and 2012
  - Can be linked to administrative data through the Linkable File Environment
  - Research projects linking the common respondents in the two surveys have been approved

Content
- Strategic decisions – locations, outsourcing, global value chain participation, lost-cost/product differentiation,
- Innovation activities – advanced technology use, product/process/marketing/organizational innovation, obstacles
- Operational tactics – production and human resource management, business practices, relationship with suppliers
Linkable File Environment

- The LFE is an environment that contains datasets from administrative and surveyed sources that are linkable (the links have been done, proven and documented), but because of the size of the databases involved are not stored as one database.
- Statistics Canada’s Business Register is the “central source” of the LFE environment.
- Depending on the research or analytic project, records with the required variables are extracted from the required databases and a “custom research dataset” is produced.
- The cost of an extraction is approximately $4000.
Administrative Datasets in the LFE

- Business Register (BR)
- Longitudinal Employment Analysis Program (LEAP)
- General Index of Financial Information (GIFI)
- T1 business and T4
- PD7 (Payroll Deduction Accounts)
- Research and Development in Canadian Industry (RDCI)
- Value of Foreign Direct Investment
- Canadian Direct Investment Abroad
- Trade in Commercial Services
- Trade in Exporter Characteristics
Survey Datasets in the LFE

- Survey of Electronic Commerce and Technology
- Survey of Innovation
- Survey of Innovation and Business Strategy
- Survey of Advanced Technology
- Survey of Commercialization of Innovation
- Survey of Intellectual Property Management
- Survey of Financing and Growth of SME’s
- Survey of Digital Technology and Internet Use
Developmental datasets and other linkage environments

- Includes: new linkages, creation of micro data from survey, micro data bases in progress with derived analytical variables, complex extractions
- In certain cases, limited and incomplete documentation available
- Additional costs may be applied
- Examples of data bases in progress:
  - National Accounts Longitudinal Micro data File
  - Canadian Employer-Employee Dynamics Database
  - Surface Transportation File
  - Patents database linked to administrative data
  - Longitudinal Census of Agriculture
National Accounts Longitudinal Microdata File (NALMF)

- Longitudinal database of Canadian enterprises covering the 2000-2014 period
  - Successor to T2-LEAP file
  - Tracks a richer set of firm characteristics over time, such as employment, payroll, revenue, profit, assets, tangible capital stock, R&D capital stock, investment, value-added and productivity
  - Updated longitudinal structure; reconcile micro data with concepts and aggregates used and produced in the Macroeconomic Accounts

- Main data sources:
  - Statistics Canada’s Business Register
  - Corporation Income Tax: T2
  - Employment: Payroll Account Deductions (PD7) and Statements of Remuneration Paid (T4 slip)
  - Goods and Services Tax: GST
Canadian Employee-Employer Dynamics Database (CEEDD) - Overview

• A set of linkable files to provide matched data between employees and employers in the Canadian labour market.

• Analysis can be done with the CEEDD data at
  1. Cross-sectional basis: at a given point in time based on covariates drawn from the same year across different component files; or
  2. Longitudinal basis: tracking firms and employees over time across different component files

• All CEEDD files contain 100% of the respective population from the administrative sources.
CEEDD - Component files

- The CEEDD linkage environment contains information from the following component files from 2001 onwards:
  1. T1 Personal Master File (T1PMF)
  2. T1 Historical Personal Master File (T1H)
  3. T1 Family File (T1FF)
  4. T1 Financial Declaration File (T1FD)
  5. T1 Business Declaration File (T1BD)
  6. T2 Schedules (T2)
  7. T2 Schedule 50
  8. T4 Statement of Remuneration Paid Files (T4)
  9. Record of Employment (ROE)
  10. Raw Import Data for Research Purpose (Import data)
  11. Trade by Exporter Characteristics (TEC)
  12. National Accounts Longitudinal Microdata File (NALMF)
  13. Longitudinal Immigration Database (IMDB)
  14. Temporary Residents File (TR)
CEEDD – Linkage environment

CEEDD linkage environment

- **IMDB** (landing file)
- **TIFF** (T1 Family File)
- **T1 PMF or T1 HPMF**
- **TR** (temporary residents)

**Incorporated business owners**
- Schedule 50 (private corporations)

**Unincorporated business owners**
- NALMF
  - Business Register
  - T2
  - T4
  - PD7
  - GST

- **T4**
- **T1 BD and FD** (unincorporated businesses)

**Exports and imports**

Legend:
- **Yellow**: Individual and family identifiers
- **Blue**: Business identifiers
- **Green**: Individual and business identifiers
CEEDD – Linkage environment

• Individual-level data:
  • *From T1 files*: Demographic information and reported earnings of individual tax filers. Multiple SIN holders are processed in the files so that information from different SINs of the same individual can be linked over time.
  • *From IMDB files*: Immigration-related information for foreign-born individuals who became landed immigrants in Canada.

• Family-level data:
  • *From T1FF files*: Individual tax filers can be linked to their spouse and children at the census family level
CEEDD – Linkage environment

• Job-level and Firm-level data:
  • For employees: From T4 and ROE files, information related to payroll and job separation is available.
  • For incorporated business owners: Information from T2 Schedule 50, T1, and T4 are linked to identify information related to the business owners and their businesses including employment, revenue, profit, and industry code.
  • For unincorporated business owners: Information from T1 self-employment income report is used to identify information related to the unincorporated business owners and information from T1BD can be used to extract information for the unincorporated businesses (2005 onwards).
  • For firms: Information from the NALMF and Import and Export files. The NALMF is a comprehensive longitudinal database of Canadian enterprises that links annual employment and administrative data from T4, PD7, T2, T2 Schedule 50, GST, and Import & Export files.
CEEDD – Linkage environment

- Geography data:
  - Province variables are available from the T1 files for individual tax filers.
  - Province of employment is available from the T4 files.
  - Province of business for unincorporated firms is available from T1BD files.
  - Province of operation for incorporated firms is available from Business Register through the NALMF.
  - Sub-provincial variables based on standard geographical classification from Census are also available at individual level.
<table>
<thead>
<tr>
<th>Output Analytical Files</th>
<th>Source Files</th>
<th>2015 vintage</th>
<th>2017 vintage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Individual-level data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1 Personal Master Files</td>
<td>T1 PMF</td>
<td>2001 to 2013</td>
<td>2001 to 2015</td>
</tr>
<tr>
<td>T1 Historical Files</td>
<td>T1 H</td>
<td>2001 to 2011</td>
<td>2001 to 2013</td>
</tr>
<tr>
<td>IMDB Files</td>
<td>Landing Files &amp; Temporary Residents Files</td>
<td>2001 to 2013</td>
<td>2004 to 2015</td>
</tr>
<tr>
<td><strong>Family-level data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1 Family Files</td>
<td>T1 PMF, T4, Canada Child Tax Benefit (CCTB) Files</td>
<td>2001 to 2013</td>
<td>2001 to 2015</td>
</tr>
<tr>
<td><strong>Job-level data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edited T4 Files</td>
<td>T4</td>
<td>2001 to 2011</td>
<td>2001 to 2013</td>
</tr>
<tr>
<td>Business owners' module</td>
<td>T1 H, T1FD, T1BD, T2 Schedule 50, T2 Corporate Income Tax, T4, IMDB</td>
<td>2001 to 2011</td>
<td>2001 to 2013</td>
</tr>
<tr>
<td>Raw T4 - ROE - LEAP</td>
<td>T4, ROE, LEAP</td>
<td>2001 to 2013</td>
<td>2001 to 2015</td>
</tr>
<tr>
<td>Edited T4 - ROE - NALMF</td>
<td>Edited T4, ROE, NALMF</td>
<td>2001 to 2015</td>
<td></td>
</tr>
<tr>
<td><strong>Firm-level data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NALMF</td>
<td>BR, T2, T4, PD7, GST, Import &amp; Export Files</td>
<td>2001 to 2013</td>
<td>2001 to 2015</td>
</tr>
<tr>
<td>Import Files</td>
<td>Raw import data for research purpose</td>
<td>2002 to 2012</td>
<td>2002 to 2012</td>
</tr>
<tr>
<td>Export Files</td>
<td>Trade by Exporter Characteristics</td>
<td>2010 to 2014</td>
<td>2010 to 2015</td>
</tr>
<tr>
<td><strong>Geography data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sub-provincial indicators</td>
<td>Postal code information from the T1 PMF</td>
<td>2001 to 2013</td>
<td>2001 to 2015</td>
</tr>
</tbody>
</table>
Surface Transportation File

- Built in collaboration with the Environment, Energy and Transportation Statistics Division, using micro data from the Trucking Commodity Origin Destination Survey and similar data for railways.

- Definition:
  - The Surface Transportation File (STF) measures the value, tonnage and cost associated with shipments of goods by truck and rail between domestic and Canada-U.S. origins and destinations from 2004 to 2012.

- Characteristics:
  - Shipments add to known provincial and Canada-U.S. trade totals, transforming a logistics file into a trade file consistent with the provincial accounts.
  - Origins and destinations are geocoded with a latitude and longitude, allowing flows between any combination of origins and destinations to be analyzed.
  - Transportation costs can be measured on an ad valorem basis (percentage of the value of the good), a measure of costs that reflects their influence on prices.

- Uses: regional trade, size of markets, effect of infrastructure investment, firm networks-investment and interprovincial trade.
Patents database

- Canadian Intellectual Property Office (CIPO):
  - 1990 to 2012
  - Filing date, address, country, grant date, lapsed date, expired date, IPC classification
- US Patent and Trademark Office (USPTO):
  - 2000 to 2011
- Firm-level administrative data from NALMF
- Uses:
  - Patenting behavior – country of filing, joint filing, patent characteristics, firm characteristics
  - Innovation inputs and innovation outputs; innovation outputs and firm performance; IP use
Workplace Employee Survey

- Explores issues relating to establishments and their employees
  - Sheds light on:
    - relationships among competitiveness, innovation, technology use and human resource management on the employer side
    - technology use, training, job stability and earnings on the employee side
    - Questions on fraction of sales to different markets, desires to expand markets, and businesses’ perception of competition
- A file with limited industry and geographic detail is available in the RDCs; Master file is available at CDER.
- Can be linked to GIFI – capital stock, and other firm performance measures
- Can be linked to ASM
Longitudinal Agriculture Census

- Longitudinal administrative database of farms and farm operators.


- 1,531,706 records and currently has 101 variables.

- Longitudinally-consistent Agricultural Operation Identifier (AGOPID), industry- and geography-based classifying variables, and analytical variables.
Select variables in the L-CEAG

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Matching</strong></td>
<td></td>
<td><strong>Inputs</strong></td>
<td></td>
</tr>
<tr>
<td>AGOPID</td>
<td>Agricultural operation identifier</td>
<td>FERTPD</td>
<td>Fertilizer and lime purchases</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HERBCI</td>
<td>Use of herbicides</td>
</tr>
<tr>
<td></td>
<td></td>
<td>INSECI</td>
<td>Use of insecticides and fungicides</td>
</tr>
<tr>
<td><strong>Classification</strong></td>
<td></td>
<td><strong>Technology</strong></td>
<td></td>
</tr>
<tr>
<td>YEAR</td>
<td>Census year</td>
<td>IRRIG</td>
<td>Use of irrigation</td>
</tr>
<tr>
<td>PROV</td>
<td>Province</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCSD</td>
<td>Longitudinal census subdivision</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LNAICS</td>
<td>Longitudinal industry classification</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Farm size</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TFAREA</td>
<td>Total area of farms in acres</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AOWNED</td>
<td>Area owned in acres</td>
<td>CANOLA</td>
<td>Canola</td>
</tr>
<tr>
<td>ALSDGOV</td>
<td>Area rented or leased from governments in acres</td>
<td>SUMMRMF</td>
<td>Summerfallow</td>
</tr>
<tr>
<td>ARNTEDE</td>
<td>Total area rented or leased from others in acres</td>
<td>TOTWHT</td>
<td>Total wheat</td>
</tr>
<tr>
<td><strong>Economic</strong></td>
<td></td>
<td>BARLEY</td>
<td>Total barley</td>
</tr>
<tr>
<td>SALE</td>
<td>Total gross farm receipts</td>
<td>TAMHAY</td>
<td>Total tame hay</td>
</tr>
<tr>
<td>TOTEXP</td>
<td>Total farm business operating expenses</td>
<td>TCATTTL</td>
<td>Total cattle and calves</td>
</tr>
<tr>
<td>VALULB</td>
<td>Total land and buildings - market value</td>
<td>TOPIGS</td>
<td>Total pigs</td>
</tr>
<tr>
<td>VALMCH</td>
<td>Farm machinery and equipment - market value</td>
<td>CATLNY</td>
<td>Cattle on farm</td>
</tr>
<tr>
<td>TCSHWGE</td>
<td>Total wages and salaries</td>
<td>PIGSNY</td>
<td>Pigs</td>
</tr>
</tbody>
</table>
Other databases

- Retail Trade Survey
- Wholesale Trade Survey
- Consumer Price Research Database
- Producer Price Database
- Bankruptcy Database
Application process

Step 1: Contact CDER and draft a proposal

a) The justification for the research:
   • context;
   • the research question;
   • contribution to the literature.

b) The analytical framework and the data requirements:
   • detailed data requirements;
   • proposed methodology;
   • justification for using micro data;
   • expected outputs;
   • software requirements;
   • expected length of project.
Application process (2)

Step 2: Submit project proposal

a) Application for accreditation:
   • CV that demonstrates experience and technical competence;

b) A letter from lead researcher indicating:
   • how project costs will be covered;
   • how the peer review of the project will be handled;
   • their ability to abide all the terms and conditions of becoming a deemed employee and the conditions in the research contract;
   • that they have no conflicts of interest to declare;
   • that they can commit to producing a research paper for Statistics Canada.
Step 3: Evaluation of proposal

1. Peer review
   a) **Project is being funded by SSHRC:** Statistics Canada will take this as recommendation that the project has been peer reviewed and that researchers are qualified
   b) **Require Statistics Canada to conduct a peer review:** Statistics Canada will solicit two reviews from an external panel of experts at a cost of $200 that will be paid by applicants

2. An internal Statistics Canada committee will review to ensure projects falls under the mandate of Statistics Canada
Application process (4)

Step 4: Complete the security screening process
Step 5: Sworn in as deemed employees of Statistic Canada and sign a micro data research contract
Future directions

Databases that can potentially be used in the RDC are being developed:

- In partnership with SSHRC-sponsored network to study Productivity, Firms and Incomes, a Canadian Synthetic Longitudinal Database is being developed, similar to the one in the U.S. Census Bureau

- Version of Survey of Financing and Growth of SMEs
More information

- Website
  - [http://www.statcan.gc.ca/cder](http://www.statcan.gc.ca/cder)

- E-mail
  - statcan.cder-cdre.statcan@canada.ca
  - Danny.Leung@canada.ca, Director, Economic Analysis Division
  - Lydia.Couture@canada.ca, CDER program manager