Statistics Canada Statistique Canada

Business microdata for economic research

www.statcan.gc.ca

CANADA 150

Telling Canada's
story in numbers

Danny Leung
Economic Analysis Division

March 24, 2017

Canada

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
  - Cross-sectional survey in 2000, 2001, 2004, 2007, 2011, and 2014
  - 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)

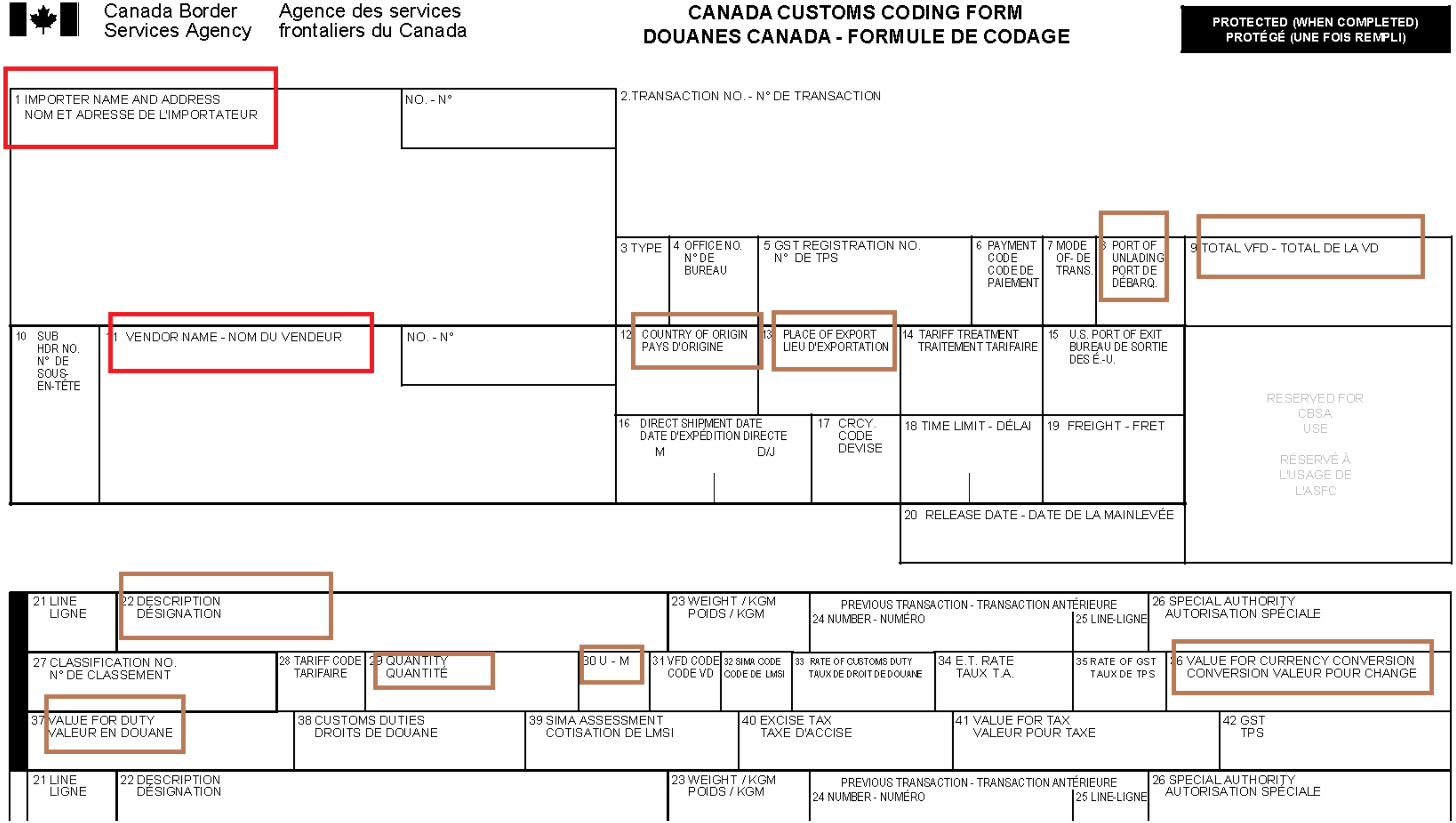
Stand-alone databases (2)

- 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:
    - 1961-1990 (1970 SIC; 2-digit level); 1970-1990 (1970 SIC; 4-digit level)
    - 1961-1999 (1980 SIC; 2-digit level); 1973-1999 (1970 SIC; 4-digiti level)
    - 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 Border Services Agency
Agence des services frontaliers du Canada

CANADA CUSTOMS CODING FORM
DOUANES CANADA - FORMULE DE CODAGE

PROTECTED (WHEN COMPLETED)
PROTÉGÉ (UNE FOIS REMPLI)

1 IMPORTER NAME AND ADDRESS
NOM ET ADRESSE DE L'IMPORTATEUR

NO. - N°

2.TRANSACTION NO. - N° DE TRANSACTION

3 TYPE

4 OFFICE NO.
N° DE BUREAU

5 GST REGISTRATION NO.
N° DE TPS

6 PAYMENT CODE
CODE DE PAIEMENT

7 MODE OF- DE TRANS.

8 PORT OF UNLADING
PORT DE DÉBARQ.

9 TOTAL VFD - TOTAL DE LA VD

10 SUB HDR NO.
N° DE SOUS-EN-TÊTE

11 VENDOR NAME - NOM DU VENDEUR

NO. - N°

12 COUNTRY OF ORIGIN
PAYS D'ORIGINE

13 PLACE OF EXPORT
LIEU D'EXPORTATION

14 TARIFF TREATMENT
TRAITEMENT TARIFAIRE

15 U.S. PORT OF EXIT
BUREAU DE SORTIE DES É.-U.

RESERVED FOR CBSA USE

RÉSERVÉ À L'USAGE DE L'ASFC

16 DIRECT SHIPMENT DATE
DATE D'EXPÉDITION DIRECTE
M D/J

17 CRCY. CODE
DEVISE

18 TIME LIMIT - DÉLAI

19 FREIGHT - FRET

20 RELEASE DATE - DATE DE LA MAINLEVÉE

21 LINE
LIGNE

22 DESCRIPTION
DÉSIGNATION

23 WEIGHT / KGM
POIDS / KGM

PREVIOUS TRANSACTION - TRANSACTION ANTÉRIEURE
24 NUMBER - NUMÉRO
25 LINE-LIGNE

26 SPECIAL AUTHORITY
AUTORISATION SPÉCIALE

27 CLASSIFICATION NO.
N° DE CLASSEMENT

28 TARIFF CODE
TARIFAIRE

29 QUANTITY
QUANTITÉ

30 U - M

31 VFD CODE
CODE VD

32 SIMA CODE
CODE DE LMSI

33 RATE OF CUSTOMS DUTY
TAUX DE DROIT DE DOUANE

34 E.T. RATE
TAUX T.A.

35 RATE OF GST
TAUX DE TPS

36 VALUE FOR CURRENCY CONVERSION
CONVERSION VALEUR POUR CHANGE

37 VALUE FOR DUTY
VALEUR EN DOUANE

38 CUSTOMS DUTIES
DROITS DE DOUANE

39 SIMA ASSESSMENT
COTISATION DE LMSI

40 EXCISE TAX
TAXE D'ACCISE

41 VALUE FOR TAX
VALEUR POUR TAXE

42 GST
TPS

21 LINE
LIGNE

22 DESCRIPTION
DÉSIGNATION

23 WEIGHT / KGM
POIDS / KGM

PREVIOUS TRANSACTION - TRANSACTION ANTÉRIEURE
24 NUMBER - NUMÉRO
25 LINE-LIGNE

26 SPECIAL AUTHORITY
AUTORISATION SPÉCIALE

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.

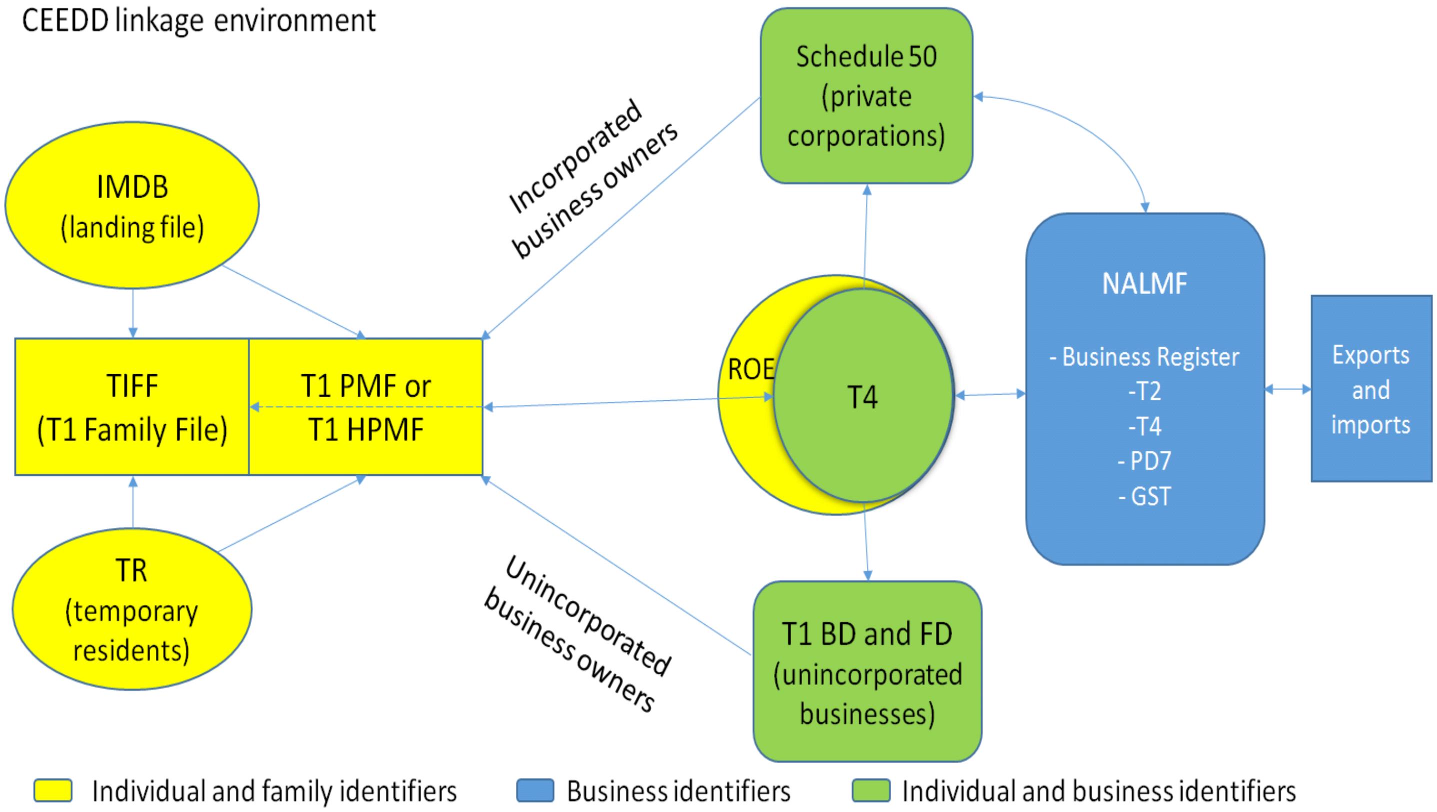
- Analysis can be done with the CEEDD data at
  1. Cross-sectional basis: at a given point in time based on covariates drawn from the same year across different component files; or
  2. Longitudinal basis: tracking firms and employees over time across different component files

- 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
Unincorporated business owners
Schedule 50 (private corporations)
ROE
T4
T1 BD and FD (unincorporated businesses)
NALMF
- Business Register
-T2
-T4
- PD7
- GST
Exports and imports
🟨 Individual and family identifiers
🟦 Business identifiers
🟩 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.

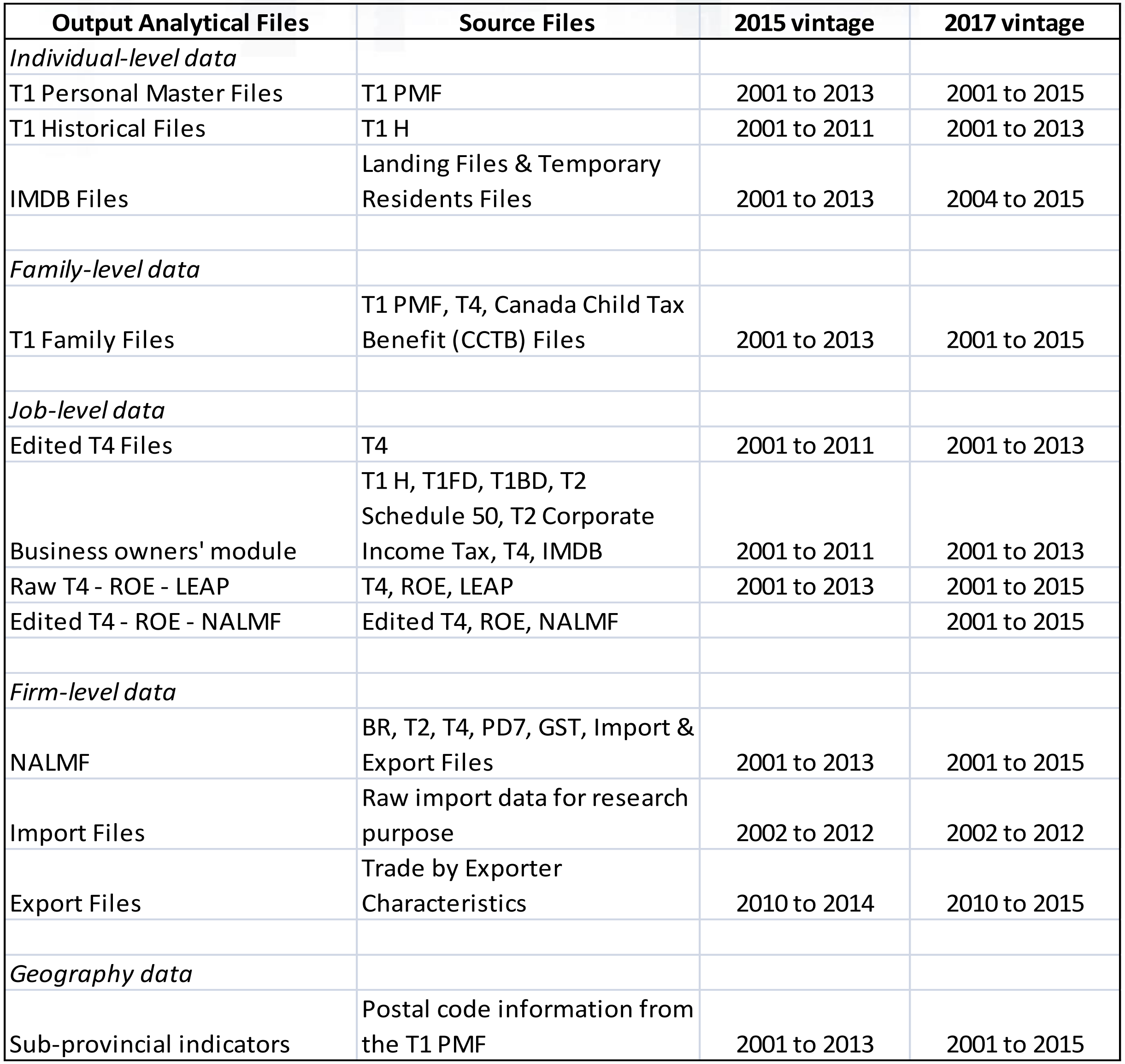
- Family-level data:
  - *From T1FF files*: Individual tax filers can be linked to their spouse and children at the census family level

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.



Output Analytical Files	Source Files	2015 vintage	2017 vintage
Individual-level data			
T1 Personal Master Files	T1 PMF	2001 to 2013	2001 to 2015
T1 Historical Files	T1 H	2001 to 2011	2001 to 2013
IMDB Files	Landing Files & Temporary Residents Files	2001 to 2013	2004 to 2015
Family-level data			
T1 Family Files	T1 PMF, T4, Canada Child Tax Benefit (CCTB) Files	2001 to 2013	2001 to 2015
Job-level data			
Edited T4 Files	T4	2001 to 2011	2001 to 2013
Business owners' module	T1 H, T1FD, T1BD, T2 Schedule 50, T2 Corporate Income Tax, T4, IMDB	2001 to 2011	2001 to 2013
Raw T4 - ROE - LEAP	T4, ROE, LEAP	2001 to 2013	2001 to 2015
Edited T4 - ROE - NALMF	Edited T4, ROE, NALMF		2001 to 2015
Firm-level data			
NALMF	BR, T2, T4, PD7, GST, Import & Export Files	2001 to 2013	2001 to 2015
Import Files	Raw import data for research purpose	2002 to 2012	2002 to 2012
Export Files	Trade by Exporter Characteristics	2010 to 2014	2010 to 2015
Geography data			
Sub-provincial indicators	Postal code information from the T1 PMF	2001 to 2013	2001 to 2015

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

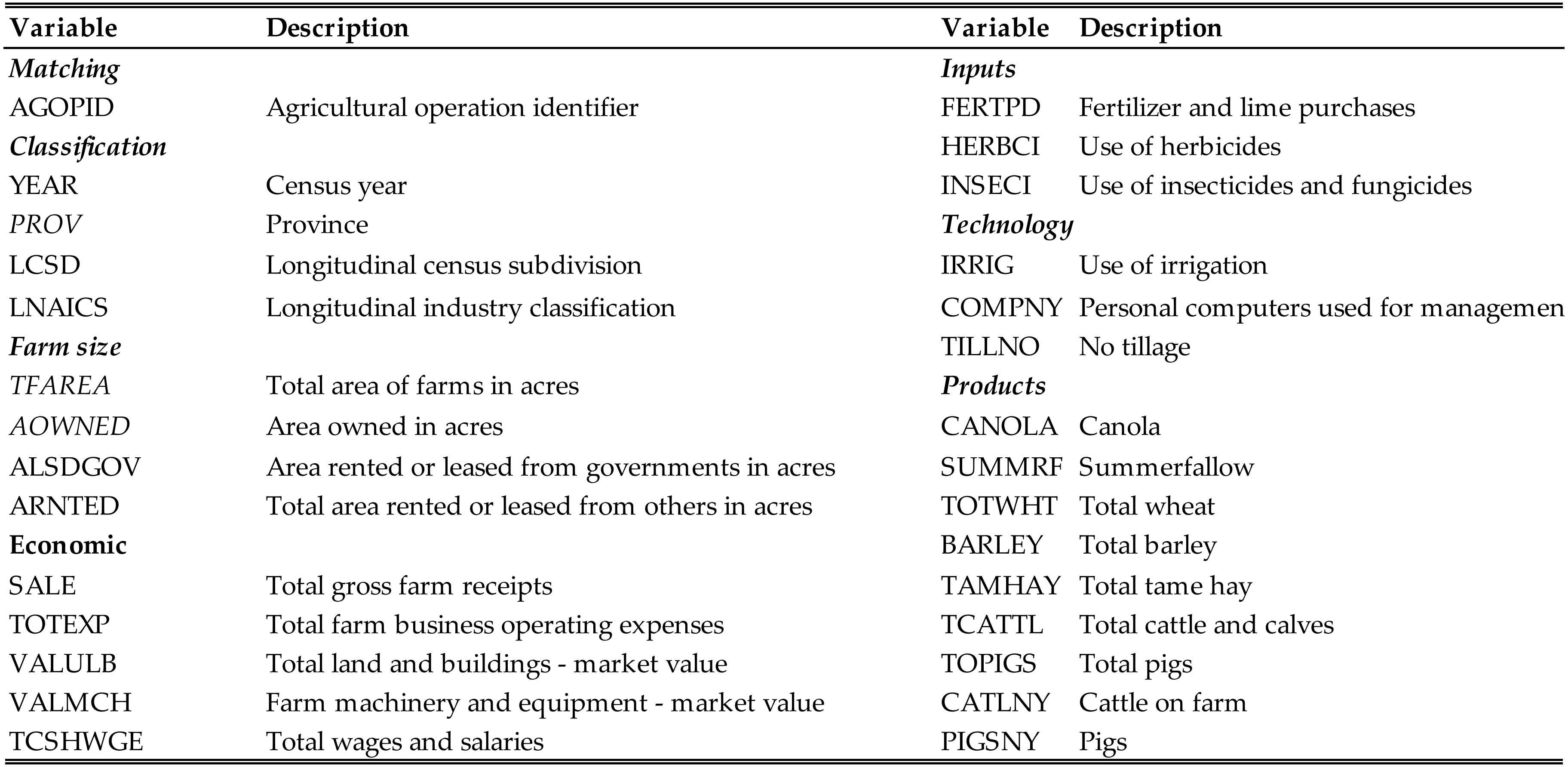
Workplace Employee Survey

- 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.
- Connects multiple censuses: 1986, 1991, 1996, 2001, 2006, and 2011.
- 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



Variable	Description	Variable	Description
Matching		***Inputs***	
AGOPID	Agricultural operation identifier	FERTPD	Fertilizer and lime purchases
Classification		HERBCI	Use of herbicides
YEAR	Census year	INSECI	Use of insecticides and fungicides
PROV	Province	***Technology***	
LCSD	Longitudinal census subdivision	IRRIG	Use of irrigation
LNAICS	Longitudinal industry classification	COMPNY	Personal computers used for managemen
Farm size		TILLNO	No tillage
TFAREA	Total area of farms in acres	***Products***	
AOWNED	Area owned in acres	CANOLA	Canola
ALSDGOV	Area rented or leased from governments in acres	SUMMRF	Summerfallow
ARNTED	Total area rented or leased from others in acres	TOTWHT	Total wheat
Economic		BARLEY	Total barley
SALE	Total gross farm receipts	TAMHAY	Total tame hay
TOTEXP	Total farm business operating expenses	TCATTL	Total cattle and calves
VALULB	Total land and buildings - market value	TOPIGS	Total pigs
VALMCH	Farm machinery and equipment - market value	CATLNY	Cattle on farm
TCSHWGE	Total wages and salaries	PIGSNY	Pigs

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.

Application process (3)

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

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