



Statistics
Canada

Statistique
Canada

Access to Statistics Canada's Business Microdata

www.statcan.gc.ca



CANADA 150

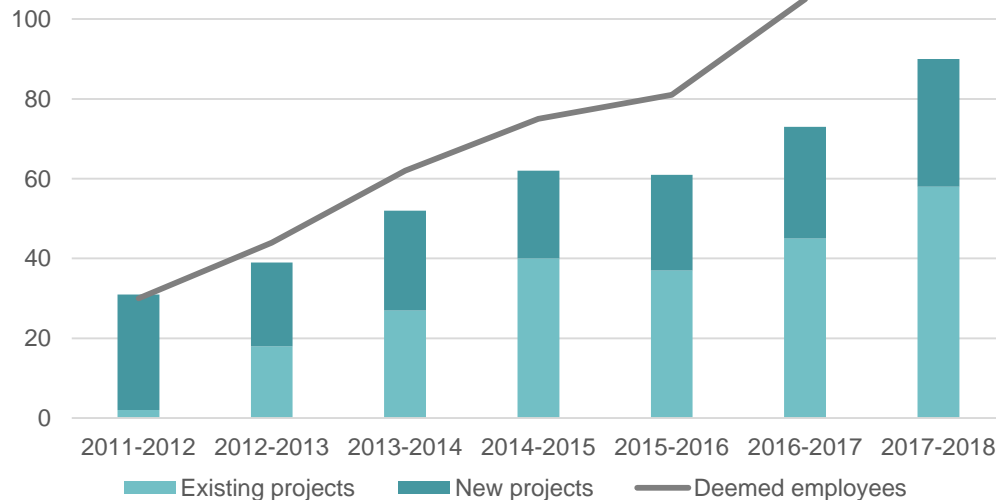
Telling Canada's
story in numbers

Lydia Couture
Program Manager

February 26, 2018

CDER overview

- Established in 2011 at Statistics Canada’s headquarters to provide secure access to business microdata for approved analytical research projects
- Institutions served:
 - About 30 Canadian and foreign academic institutions
 - 16 federal departments and agencies, provincial government and non-profit institutions





Improving access to business microdata

1) Pilot project

2) Synthetic data

- Cost of coming to Ottawa has been identified as one of the most significant barrier to access
- 2016 CDER workshop: intro to sophisticated synthetic data and methodology - Lars Vilhuber

3) Collaboration with Productivity Partnership

4) Improving application process

Pilot project

- Statistics Canada will announce a call for proposal to use synthetic business micro data in the Research Data Centres (RDCs)
- Approved research projects that focus on multivariate analysis will be carried out in the RDCs in 2018-2019
 - Researchers will develop models using the synthetic/treated data at the RDCs
 - Results from the synthetic and treated data will not be released
 - Programs will be sent to CDER by RDC analysts to be run on the actual data
 - Final results are released from CDER by e-mail



Synthetic Data Projects

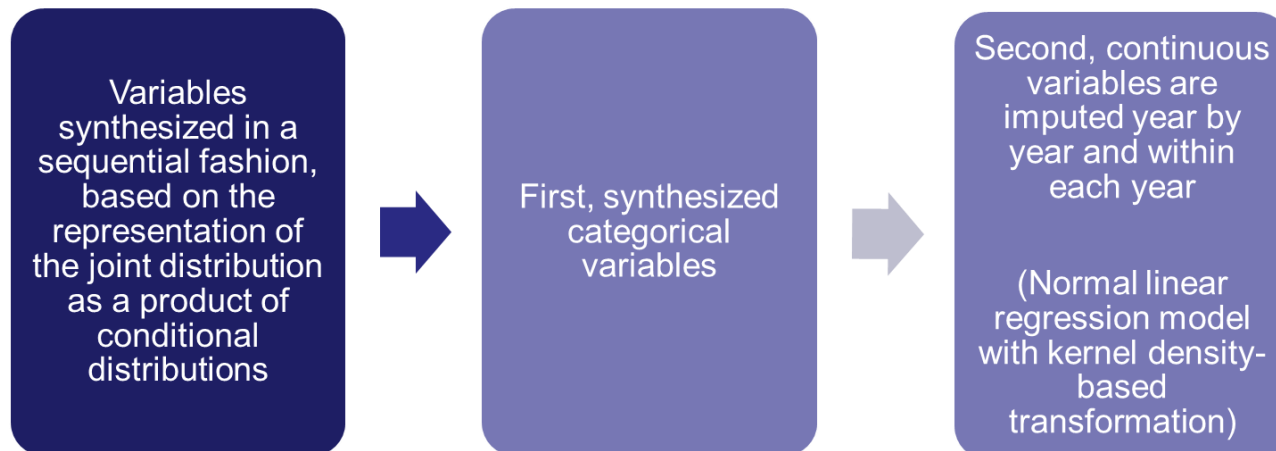
- More sophisticated synthetic data developed to facilitate access outside of headquarters
 - Synthetic Longitudinal Business Database
 - Disclosure-limited Survey of Financing and Growth of Small and Medium Sized Enterprises (SFGSME) linked to CRA tax data

Synthetic longitudinal business database

- A wholly-synthetic version of the Longitudinal Employment Analysis Program (LEAP) database (2001 to 2015)
- LEAP is a database of employer firms
 - T4 Statement of Remuneration Paid
 - Business Register
 - Survey of Employment Payroll and Hours (SEPH)
 - Variables: entry, exit, payroll, average labour unit (payroll divided average annual earnings from SEPH), and industry
 - Used to study business and employment dynamics (e.g., firm entry and exit, job creation and destruction, firm survival and growth, high-growth firms, firm transitions)

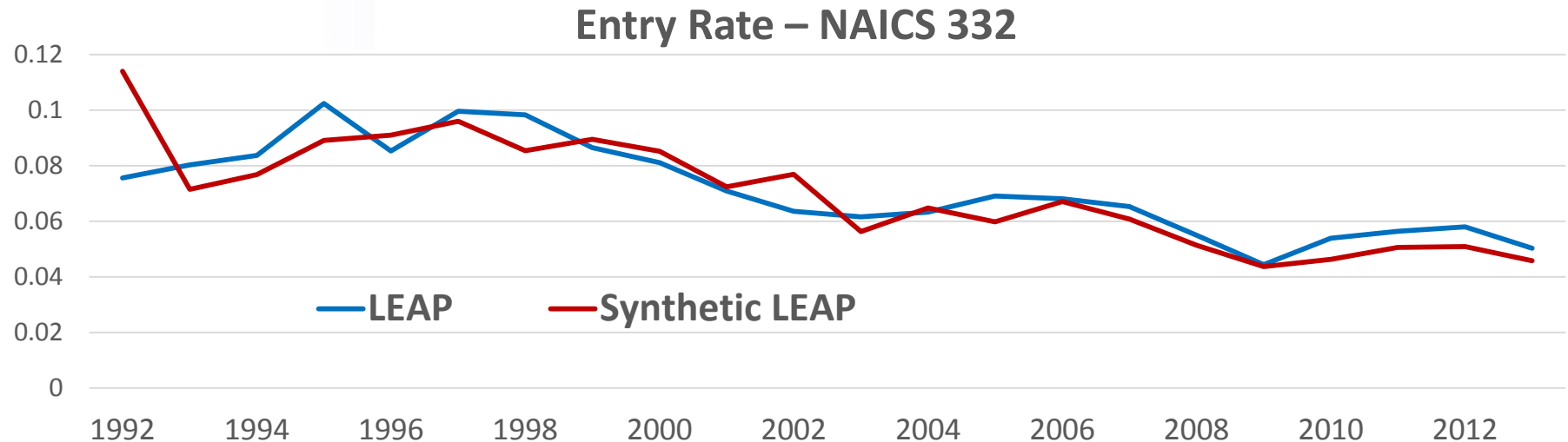
Methodology

- Follows approach developed at the U.S. Census Bureau
 - Worked with Methodology Branch, Cornell University, Productivity Partnership
- All variables are modelled and then simulated by industry
 - There is no risk of direct re-identification as each observation is simulated
 - The synthetic data will be able to be used to produce detailed aggregates that resemble the ones in the actual data, including ones that are considered confidential
- Synthetic data are created by replacing actual values with draws from a model fit to the original data.
- Approach closely related to multiple imputation





Analytic Validity of the Synthetic Data



- Similar magnitude and trend for entry, exit, employment, and payroll over time

Analytic Validity of the Synthetic Data

- Test of the validity using regression analysis

$$emp_i = \alpha + \beta emp_{i,t-1} + \delta pay_i + \vartheta age_i + \theta Naics_i + \epsilon_i$$

Variable	LEAP	Standard Error	Synthetic LEAP	Standard Error
Intercept	-9.08	0.05	-6.83	0.06
Emp_t-1	0.12	0.00	0.30	0.01
pay	0.86	0.00	0.65	0.01
Age 4-7	-0.04	0.01	-0.05	0.01
Age 8 -10	-0.03	0.01	-0.04	0.01
Age 10-22	-0.02	0.01	-0.05	0.01
Age 23 and more	-0.02	0.01	-0.04	0.01



Confidentiality

- No risk of direct re-identification
- Synthetic file will have limited industry detail
- Final runs on master file will allow the use of more detailed industry and geography, and other variables from the broader T2-LEAP database



Productivity Partnership

- Team of experts from the academic, private, and public sectors
- Lowers barriers to access business data for academics and students by providing funding for productivity research projects
- Statistics Canada liaison researcher: Beryl Li

<https://productivitypartnership.ca>



Productivity Partnership (2)

- Important partner in the development of synthetic datasets
 - Post-doc: M. Jahangir Alam
- Collaboration and support to clarify the CDER application process steps

Metrics

Average length – CDER Application process

- Federal government projects: 3 months
 - 2 months when projects that required record linkage approval are excluded
- Academic projects: 6 months
- Difference: mainly due to security clearance & peer review



Website – upcoming revision in 2018

- Clear application process steps
- Revised proposal requirements
- Security screening guidance
- Conflict of interest section

Next steps

- Develop additional synthetic data:
 - Business surveys: SFGSME, etc.
 - Tax data
- Expand the use of synthetic data and access to business micro data outside of CDER and StatCan HQ in the short term
- Upcoming Webinar-Seminar on Pilot



Thank you! Merci !

CDER e-mail:

statcan.cder-cdre.statcan@canada.ca