Synthetic Data for Canadian Longitudinal Business Data

M. Jahangir Alam, Benoit Dostie, Lars Vilhuber

November 25, 2019

Abstract

======= Data on businesses collected by statistical agencies are challenging to protect. Many businesses have unique characteristics, distributions of employment, sales, and profits are highly skewed, and most disclosure avoidance mechanisms fail to strike an acceptable balance between usefulness and confidentiality protection. Often, only very few aggregate statistics are released, and access to confidential microdata can be burdensome.

This paper documents the creation of a synthetic data version of Statistics Canada's Longitudinal Employment Analysis Program (LEAP). Since the LEAP has a structure similar to the U.S. Longitudinal Business Database (LBD), this allows us to adapt the procedures that were used by add citation to create the synthetic LBD in a Canadian context. We show the synthetic LEAP is analytically valid for a wide range of commonly used statistical analyses, while maintaining respondents' confidentiality.

Contents

1 Introduction

While Canada maintains a network of research data centers similar to the one in the U.S., there is virtually no firm-level data sets amongst the data holdings of the Canadian Research Data Center Networks (CRDCN). Researchers who need access to confidential micro-level business data have to go to the Canadian Center for Data Development and Economic Research (CDER) located in the headquarters of Statistics Canada. One reason commonly heard for this limited access has to do with the fact that Canada is a small country with a highly skewed distribution of firms, thus compounding problems linked to confidentiality protection.

Confidentiality protection with CDER is done through a variety of means, including remote execution and research monitoring. But more importantly, researchers accessing CDER data holdings must become *deemed employees* of Statistics Canada and sign an asset-freeze agreement to ensure they do not profit financially from their research.

Since 2018, Statistics Canada has undertaken a so-called Modernization initiative. One of the objective of that Initiative is to improve access to researchers to confidential firm-level micro data. As part of this initiative, we have been asked by CDER to explore how the creation of synthetic data could help achieve part of that objective. Synthetic data are created by replacing sensitive value from the original data with repeated draws from a model fit to the original data (Little, 1993; Rubin, 1993). This approach is closely related to multiple imputations.

We use this approach on the Canadian Longitudinal Employment Analysis Program (LEAP) 2015 that contains retrospective data on firms from 1991 to 2014. Our implementation adapts the approach used to create synthetic data version of the Longitudinal Business Database (LBD) from the United States since both data sets have a similar structure and are used for similar research questions. We assess the performance of the newly created synthetic along two dimensions: analytical validity and confidentiality protection.

We verify the analytical validity of synthetic data set so created along a variety of measures. First, we show that more average firm characteristics (gross employment, total payroll) in the synthetic data closely match those from the original data. Second, we also find the synthetic data close replicates various measures of firm dynamics (entry and exit rates) and job flows (gross and net job creation rate) from the original data. Finally, we assess whether measures of economic growth vary between both data sets using a dynamic panel data models and find that both data sets yield similar predictions.

To provide evidence on the confidentiality properties this newly created synthetic database, we estimate the probability that the synthetic first year equals the true first year given the synthetic fist year and find that those probabilities are quite low except for the first year of LEAP database. This is because of censoring and lack of previous information.

The rest of the paper is organized as follows. Section 2 provides a detailed description of the LEAP database and its uses. Section 3 defines what we mean by synthetic data and describe how we created the synthetic LEAP. In section 4, the analytical validity of synthetic database is assessed. Section 5 concludes.

BD: I have no idea what this means, please explain better.

2 Data Description

2.1 LEAP database

The LEAP contains information on annual employment for each employer in Canada. It covers incorporated and unincorporated businesses that issue at least one annual statements of remuneration paid (T4) in any given calendar year, but excludes self-employed individuals or partnerships with non-salaried participants. One advantage of the LEAP is that it covers all sectors of the Canadian economy.

To construct the LEAP Statistics Canada draws from three distinct sources: (1) T4 statements from Canada Revenue Agency, (2) Statistics Canada's Business Register, and (3) Statistics Canada's Survey of Employment, Payrolls and

Hours (SEPH). In Canada, employing businesses are required to register with Canada Revenue Agency using their Business Number, and issue to each of their employees a T4 statement summarizing earnings received in the current year. This process creates a link between the employee and the business through the Business Number that is the backbone of LEAP. Reported payrolls from SEPH allows estimates of annual employment to be added to the data set. The payroll is converted to employment (called ALUs or Average Labour Units, defined later) using conversion factors derived from the SEPH.

The LEAP essentially contains four variables (1) A Longitudinal Business Register Identifier (LBRID), (2) Industry, (3) Employment and (4) Payroll. This still allows research on multiple themes, like employment growth, industry turnover, firm survival, job creation and job destruction, etc. We discuss each of those variables in turns.

LBRID: This is the unique identifier assigned to each enterprise. The LBRID tracks the enterprise across all years in which it has employees, for the period covered by the LEAP vintage. It is derived from the Business Register enterprise identifier (BRID).

For various administrative reasons, an enterprise's identifier in the Business Register may sometimes change from year to year. This would lead to the appearance of false deaths and births in the LEAP file. To avoid this, a system of Labour Tracking is used to track the movements of workers between firms. This is used to detect false births and deaths and link firms by a common LBRID. Labour tracking can lead to many different types of linkages between firms.

The simplest would be a one-to-one linkage between a death and a birth record. For example, if a business changes from incorporated to limited business, the Business Register may remove the original business from the register and create a new one. In this case, the only action necessary is to assign a common LBRID to the two businesses over time.

A more complex case would be a merger between two firms, where most employees from the previous two firms are at a new firm. Here, all three entities

are given the same LBRID, and the past records of the two merged firms would be combined into a single record. The employment of the two firms is added together, and the current NAICS code for the new firm is assigned to the combined, synthetic past record. In other words, it would be as if the newly merged firm already existed in the past. Similarly, acquisitions and spin offs lead to the combination of firms and the creation of synthetic records.

Industry: The 4-digit North American Industrial Classification System (NAICS) code that is assigned to a firm nationally. This is the dominant NAICS code for firms that have activity in multiple industries. One of the characteristics of the LEAP is that the industry code for the most recent year that the firm is in operation is pushed back in time, so that an enterprise has the same industry code each year within the same vintage.

Employment: Employment of each firm is measured by its average labour units (ALUs). ALUs are the average employment an enterprise would have if it paid its workers the average annual earnings (AAE) of a typical worker in the enterprise's particular industry, province and enterprise size class. AAE are derived using information from the SEPH.

Payroll: Sum of payroll from all T4 slips issued by the enterprise. We next turn to the methods we used to create a synthetic version of the LEAP.

3 Methodology

3.1 Overview

There is growing demand for firm-level data allowing detailed studies of firm dynamics. Recent examples include Bartelsman, Haltiwanger, and Scarpetta [3] who use cross-country firm-level data to study average post-entry behavior of young firms. Sedláček and Sterk [11] use the BDS to show the role of firm size in firm dynamics.

BD: define BDS

However, such studies are made difficult due to the limited or restricted access to firm-level data.To provide better access to establishment data in the

United States, Kinney et al. [10] describe an approach to create and release synthetic data for Longitudinal Business Database (LBD), which was created in the early 2000s (see Jarmin and Miranda [7] for details). The variables currently available in the LBD are industry, annual payroll, employment, geography, birth year, death year, and firm structure.

BD: can we be more precise than geography and firm structure?

We can currently distinguish between two methods to create synthetic data. The general approach to data synthesis is to generate a joint posterior predictive distribution of $Y|X$ where Y are variables to be synthesized and X are unsynthesized variables. In the Phase 1 version of the method, variables are synthesized in a sequential fashion, with categorical variables being generally processed first using a variant of Dirichlet-Multinomial. Continuous variables are then synthesized using a normal linear regression model with kennel density-based transformation (Woodcock and Benedetto [13]).

The Phase 2 version of the method has shifted to a Classification and Regression Trees (CART) model with Bayesian bootstrap. For the United States, the phase 2 version is currently in its final stages of implementation (Kinney, Reiter, and Miranda [9]).

To evaluate whether synthetic data algorithms developed in the U.S. can be adapted to generate similar synthetic data for other countries, Drechsler and Vilhuber [6] implement the Phase 1 version of the method to the German Longitudinal Business Database (GLBD).

3.2 Implementation in the Canadian context

To create a Canadian synthetic database, we use the 2015 LEAP vintage. As for the U.S. synthetic database for LBD, we synthesize categorical variables first, followed by continuous variables, controlling for the firm ID and industry classification at 4-digit NAICS (see Table 1.)

BD: Is the name SynLBD official, and should we use it?

Table 1: CanSynLEAP variable descriptions

Name	Type	Description	Notation	Action
synid	Identifier	Unique random number for enterprise		Created
NAICS	Categorical	4 digit industry code	x_1	Unmodified
Firstyear	Categorical	First year enterprise is observed	y_1	Synthesized
Lastyear	Categorical	Last year enterprise is observed	y_2	Synthesized
Year	Categorical	Year dating of annual variables		Created
ALU	Continuous	Average Labor Unit (annual)	y_3	Synthesized
Payroll	Continuous	Payroll (annual)	x_4	Synthesized

Note: Variables denoted with y_i are synthesized and variables denoted with x_i are not synthesized.

After implementing the U.S. synthetic LBD code, we follow four steps to create a Canadian synthetic database.

1. We exclude the public sector (NAICS 61, 62, and 91) because Statistics Canada does not produce any statistics for those sectors.

2. We exclude industries for which the algorithm is not converging.These industries are NAICS 4481, 4482, 4483, 4511, 4513, 4841, 4842, 5241, and 5242. These industries represent approximately 7 percent of the total number of observations and are labeled as "not synthesized" in Table 2.

3. We drop some industries, from the synthesized industries, which have sensitive information.

4. We drop observation for the last year each firm was observed since the SynLBD code does not properly approximate the last year of the data.

. After the implementation of these steps, we have around 22 million observations in the CanSynLBD database during the period of 1991 - 2014.

Which one? What is meant by sensitive?

Need to mention suppressed industries. Note presence of 0000 industries. Discuss.

Table 2: Synthesized observations

Category	# of Observations (millions)	Percentage
Synthesized	22.01	93.35
Not synthesized	1.57	6.65
Total	23.58	100.00

Note: Not synthesized industries are NAICS 4481, 4482, 4483, 4511, 4513, 4841, 4842, 5241, and 5242. These industries are not converging for each time of implementation We drop industries, from the synthesized industries, which have sensitiveness information.We do not synthesize the public sector (NAICS 61, 62, and 91).

4 Analytical validity

4.1 Firm Characteristics

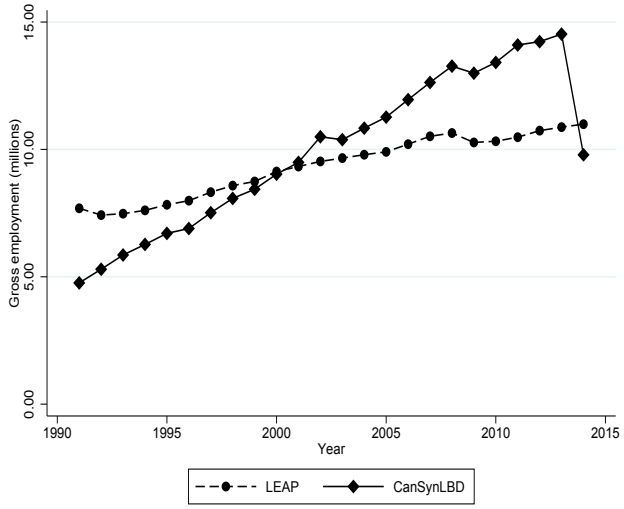
The CanSynLBD and LEAP generally provide comparable inferences on aggregate means and correlations. For example, Figures 1 and 2 show that gross employment levels for each year in the CanSynLBD are very close to those in the LEAP. However, the manufacturing sector shows closer patterns than private sector.We find similar results for total payroll (Figures 3 and 4) .

BD: adjust footnote using previous answers to questions.

BD: Is what is included in the Private sector defined somewhere? Is the Manufacturing sector included in the Private sector? Please provide details

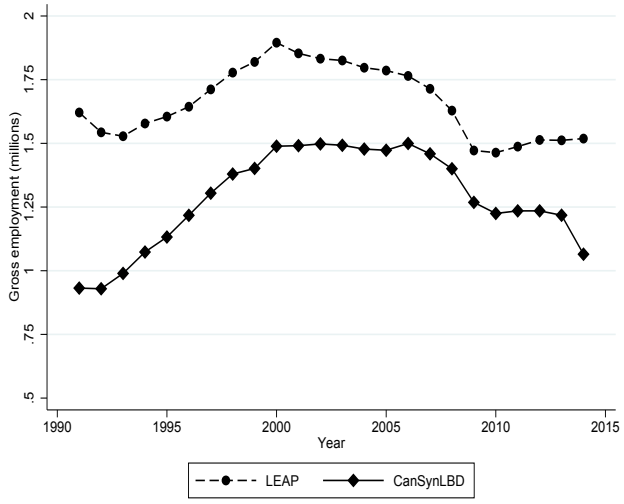
Redo Graph 1 omitting the last year

Figure 1: Gross employment level by year (private)



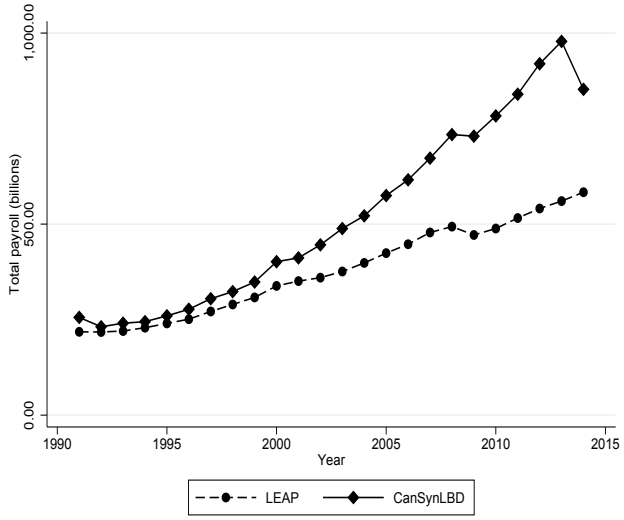
Note: *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. In this graph, we use the 2015 vintage of LEAP for private sector and drop the last year of observation for each firm.

Figure 2: Gross employment level by year (manufacturing)



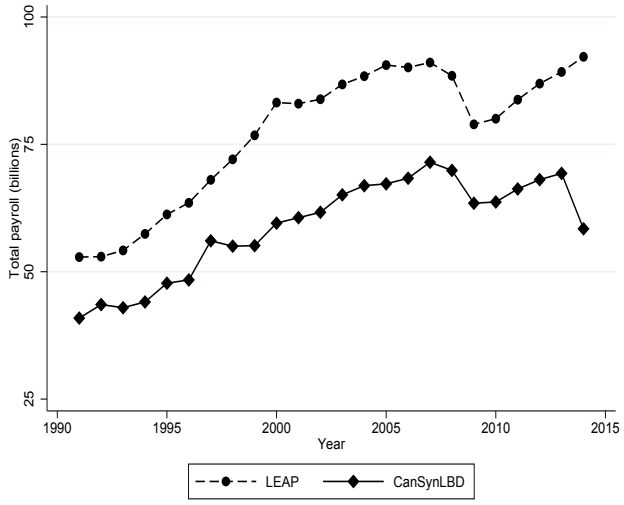
Note: *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. In this graph, we use the 2015 vintage of LEAP for the private sector and drop the last year of observation for each firm.

Figure 3: Total payroll by year (private)



Note: *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. In this graph, we use the 2015 vintage of LEAP for the private sector and drop the last year of observation for each firm.

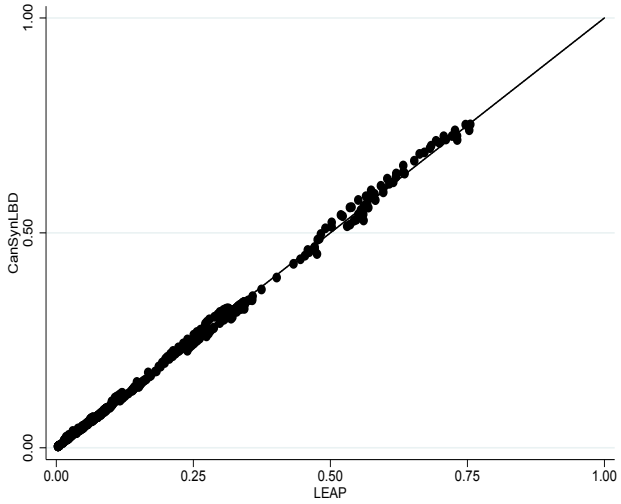
Figure 4: Total payroll by year (manufacturing)



Note: *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. In this graph, we use the 2015 vintage of LEAP for the private sector and drop the last year of observation for each firm.

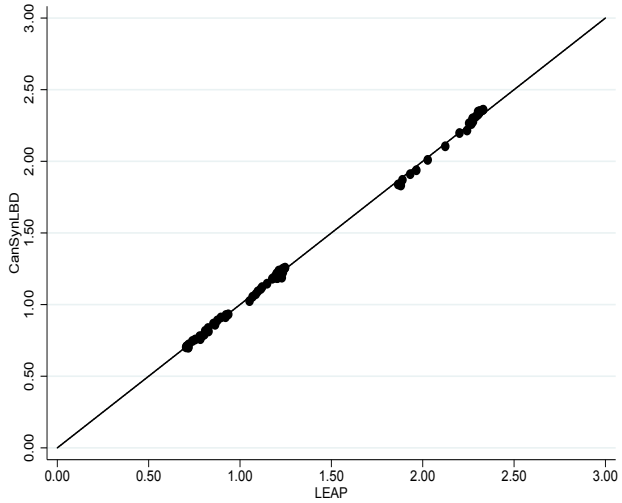
Figures 5 and 6 plot the share of firms by two-digit industry and year for both the CanSynLBD and the LEAP databaseand show that those shares clustering along the 45-degree line.

Figure 5: Share of firms by NAICS two-digit and year (private)



Note: *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. In this graph, we use the 2015 vintage of LEAP for the private sector and drop the last year of observation for each firm.

Figure 6: Share of firms by NAICS two-digit and year (manufacturing)



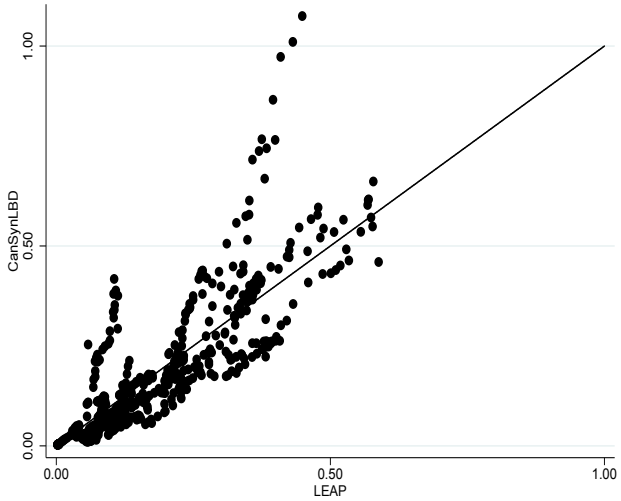
Note: *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. In this graph, we use the 2015 vintage of LEAP for the private sector and drop the last year of observation for each firm.

Figures 7 and 8 plot the share of employment by two-digit industry and year for both the CanSynLBD and the LEAP database [1] and show that those shares do not cluster along the 45-degree line. However, this hides significant differences between sectors as, for the share of employment for the manufacturing sector, we do observe clustering along the 45-degrees.

Why is manufacturing always below, but overall employment crosses? Which industries are driving that?

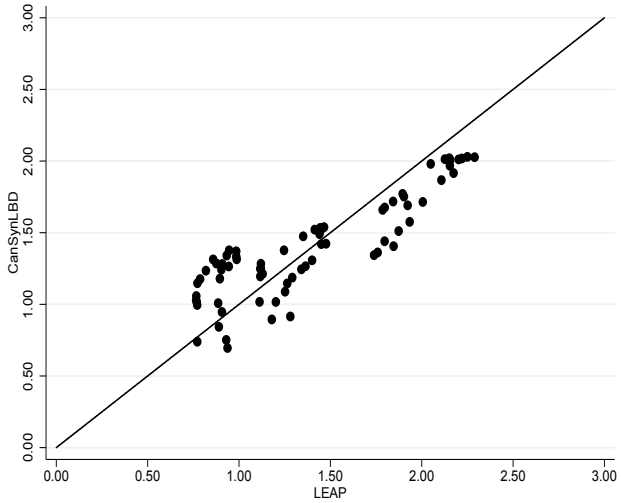
[1] $x_{its} = X_{its}/\sum_i \sum_t X_{its}$,where i are two-digit NAICS industries, t are the years in-sample, and s indicates whether it is in the synthetic or confidential data.

Figure 7: Share of employment by NAICS two-digit and year (private)



Note: *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. In this graph, we use the 2015 vintage of LEAP for the private sector and drop the last year of observation for each firm.

Figure 8: Share of employment by NAICS two-digit and year (manufacturing)

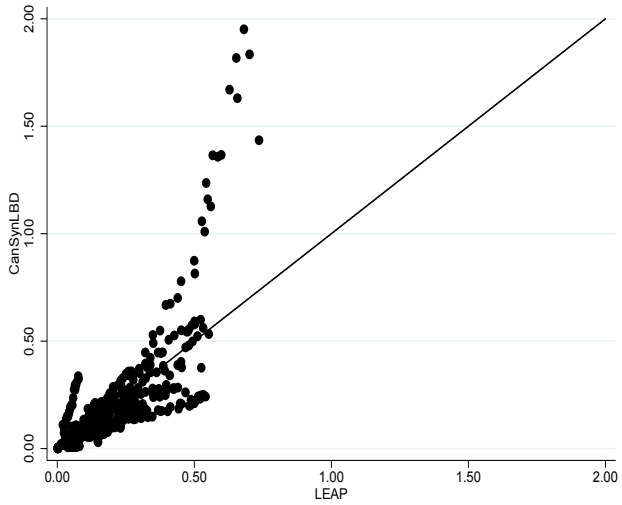


Note: *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. In this graph, we use the 2015 vintage of LEAP for the private sector and drop the last year of observation for each firm.

Figures 9 and 10 plot the share of payroll by two-digit industry and year for both CanSynLBD and LEAP database and show that those shares do not cluster along the 45-degree line. Again, we do notice that for the share of employment for the manufacturing sector, we do observe clustering along the 45-degrees.

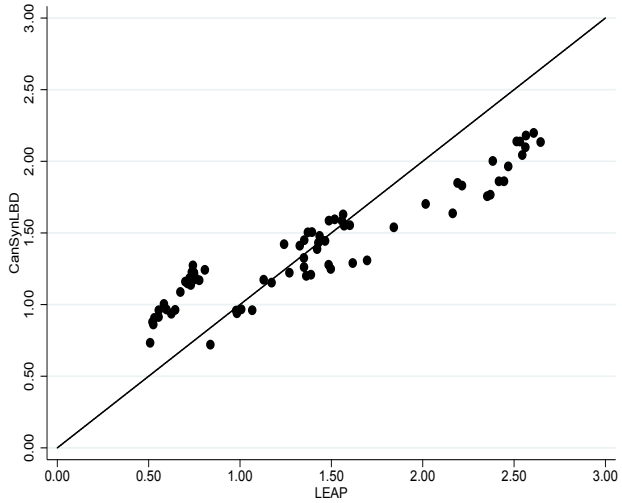
BD: Define x and X

Figure 9: Share of payroll by NAICS two-digit and year (private)



Note: *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. In this graph, we use the 2015 vintage of LEAP for the private sector and drop the last year of observation for each firm.

Figure 10: Share of payroll by NAICS two-digit and year (manufacturing)



Note: *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. In this graph, we use the 2015 vintage of LEAP for the private sector and drop the last year of observation for each firm.

4.2 Firm Dynamics

To assess how well the CanSynLBD captures firm dynamics, we also compute entry and exit rates of the private sector by year. Table 3 shows that those rates for CanSynLBD are similar to LEAP database. In addition, we compute the difference between the entry rate as the entry rate of CanSynLBD net the entry rate of LEAP and the divergence of exit rate as the exit rate of CanSynLBD net the exit rate of LEAP (see Table 3).

BD: This is not easily apparent from the graph

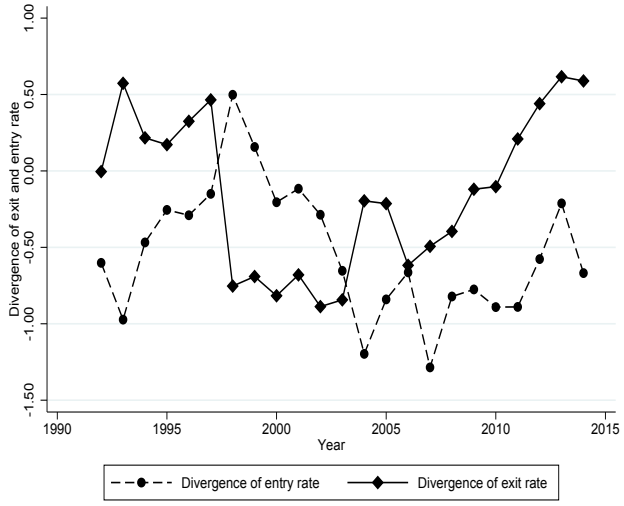
BD: This is not easily apparent from the graph

Table 3: Entry and exit rates by year

Year	LEAP		CanSynLBD		Divergence	
	Entry Rate	Exit Rate	Entry Rate	Exit Rate	Entry Rate	Exit Rate
1992	11.77	11.72	11.16	11.71	-0.60	-0.00
1993	11.81	11.61	10.84	12.18	-0.97	0.57
1994	12.04	11.79	11.57	12.01	-0.47	0.22
1995	11.94	12.09	11.69	12.26	-0.25	0.17
1996	12.91	10.31	12.62	10.64	-0.29	0.32
1997	13.18	9.75	13.03	10.21	-0.15	0.47
1998	12.48	10.89	12.97	10.13	0.50	-0.75
1999	12.00	10.66	12.16	9.97	0.16	-0.69
2000	11.80	10.51	11.59	9.70	-0.20	-0.82
2001	11.44	10.20	11.33	9.52	-0.12	-0.68
2002	11.39	9.91	11.10	9.03	-0.29	-0.89
2003	11.17	10.21	10.52	9.37	-0.65	-0.84
2004	12.13	9.76	10.94	9.57	-1.20	-0.20
2005	11.92	10.07	11.07	9.86	-0.84	-0.21
2006	11.81	9.96	11.15	9.34	-0.66	-0.62
2007	12.28	9.80	10.99	9.31	-1.29	-0.49
2008	11.60	10.14	10.78	9.75	-0.82	-0.40
2009	10.77	9.93	9.99	9.81	-0.78	-0.12
2010	10.80	9.75	9.91	9.65	-0.89	-0.10
2011	10.62	9.79	9.73	10.00	-0.89	0.21
2012	10.60	9.76	10.02	10.20	-0.58	0.44
2013	10.16	9.71	9.95	10.32	-0.21	0.62
2014	9.93	10.11	9.26	10.70	-0.67	0.59

Note: *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. In this graph, we use 2015 vintage of LEAP for the manufacturing sector and drop last year observation of each firm. We calculate the divergence of entry rate as the entry rate of CanSynLBD net the entry rate of LEAP and the divergence of exit rate as the exit rate of CanSynLBD net the exit rate of LEAP.

Figure 11: Divergence of exit and entry rate between LEAP and CanSynLBD



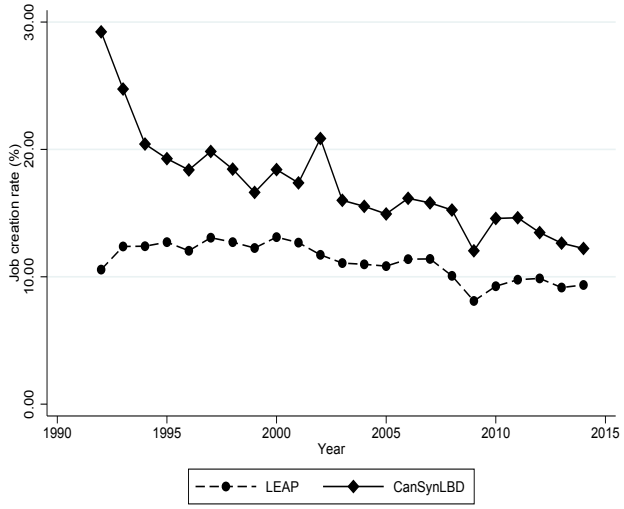
Note: *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. In this graph, we use 2015 vintage of LEAP for private sector and drop last year observation of each firm. We calculate the divergence of entry rate as the entry rate of CanSynLBD net the entry rate of LEAP and the divergence of exit rate as the exit rate of CanSynLBD net the exit rate of LEAP.

4.3 Dynamics of Job Flows

One of the most important applications of LEAP is to generate statistics that describe job flows. Following [5], the job creation is defined as the sum of all employment gains from expanding firms from year $t-1$ to year t including entry firms. The job destruction rate is defined as the sum of all employment losses from contracted firms from year $t-1$ to year t including exiting firms. Net job creation is the job creation rate minus the job destruction rate. Figures 12 and 13 show the job creation rates from the CanSynLBD compared againg those of the LEAP. These figures show that the manufacturing sector has closer pattern than the private sector. We find a similar patterns for net job creation rates (Figures 14 and 15).

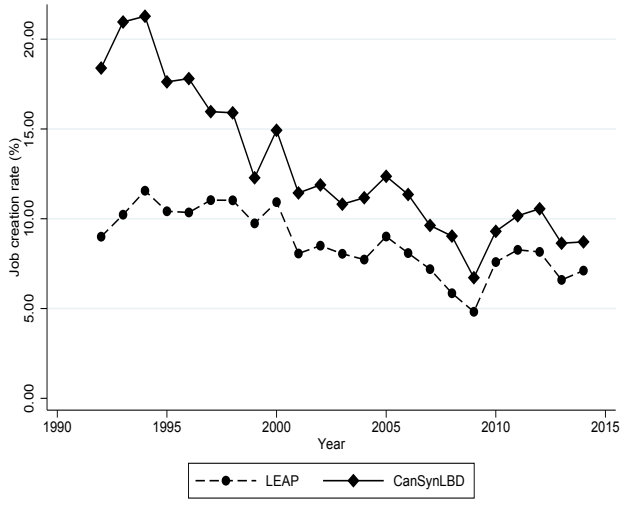
BD: I don't understand what was done here

Figure 12: Job creation rate by year (private)



Note: *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. In this graph, we use 2015 vintage of LEAP for the private sector and drop last year observation of each firm.

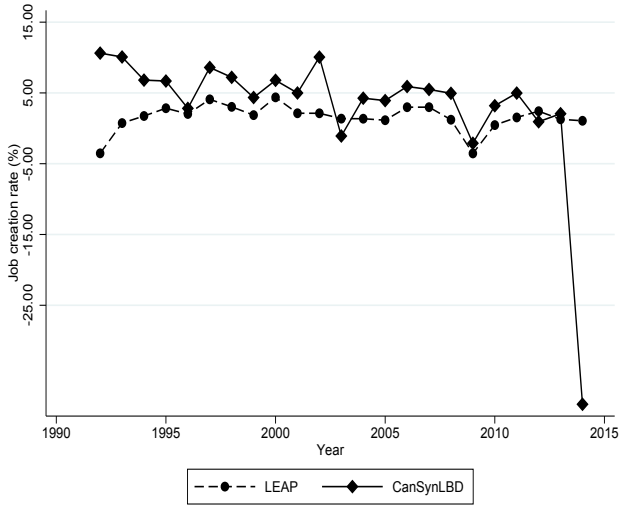
Figure 13: Job creation rate by year (manufacturing)



Note: *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. In this graph, we use 2015 vintage of LEAP for the manufacturing sector and drop last year observation of each firm.

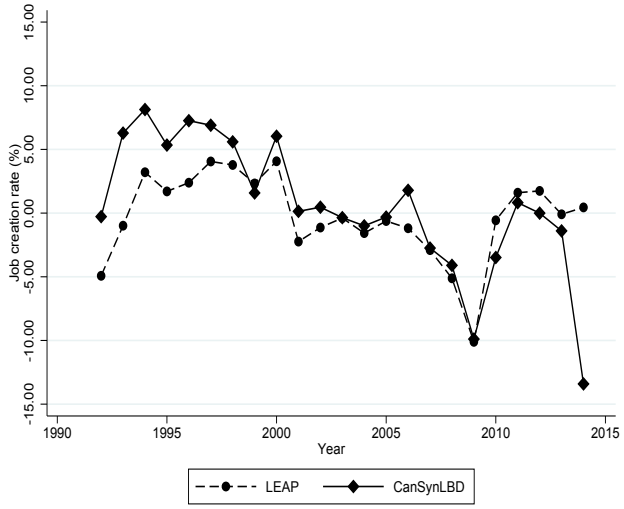
LV Reformat table, compute divergence

Figure 14: Net job creation rate by year (private)



Note: *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. In this graph, we use 2015 vintage of LEAP for the private sector and drop last year observation of each firm.

Figure 15: Net job creation rate by year (manufacturing)



Note: *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. In this graph, we use 2015 vintage of LEAP for the manufacturing sector and drop last year observation of each firm.

4.4 pMSE

To compare the quality of the synthetic data relative to the confidential data, we compute $pMSE$, which is the mean-squared error of the predicted probabilities (i.e., propensity scores) for those two databases. Specifically, $pMSE$ is a metric to assess how well we are able to discern between synthetic data and confidential data.

We follow the method by Snoke and Slavkovic [12] to calculate the $pMSE$. This method involved the following steps:

1. Append the n_1 rows of the confidential database X to the n_2 rows of the synthetic database X^s to create X^{comb} with $N = n_1 + n_2$ rows.

2. Create an indicator variable, I, to X^{comb} subject to $I = \{1 : X^{comb} \in X^s\}$. This means that we create an indicator variable of 1 for the synthetic database and 0 for the confidential database.

3. Fit the following model to predict I

$$I = \alpha + ALU_{it} + \lambda Pay_{it} + Age_{it}^{T}\beta + \lambda_t + \alpha_s + \epsilon_{it} \quad (1)$$

where ALU_{it} is the logarithm of average labour unit (ALU) of firm i in year t, Pay_{it} is the logarithm of payroll of firm i in year t, Age_{it} is a vector of dummy variables for age of firm i in year t, λ_t is the year fixed effect, α_s is an unobserved time-invariant industry-specific effect, and ϵ_{it} is the disturbance term of firm i in year t.

4. calculate the predicted probabilities, $\hat{p}_i$ for each row of X^{comb}

5. Compute the $pMSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - 0.5)^2$

A $pMSE = 0$ means every $\hat{p}_i = 0.5$.

To compute the $pMSE$, we estimate equation 1 using both the logit and probit models. Table 4 shows the calculated value of $pMSE$, which is lower for the manufacturing sector than the public sector in both regressions. This is because, as we explained before, the synthetic data mirrors the original data more closely in the case of the manufacturing sector.

BD: Contracted?

Table 4: pMSE estimates

Independent Variables	**Logistic Regression**		**Probit Regression**	
	Manufacturing	Private	Manufacturing	Private
Ln ALU	0.1580***	0.7138***	0.1003***	0.4390***
	(0.0039)	(0.0010)	(0.0024)	(0.0006)
Ln Pay	0.0039	-0.4426***	0.0012	-0.2691***
	(0.0037)	(0.0010)	(0.0023)	(0.0006)
Age 3-4	0.0392***	0.0972***	0.0252***	0.0618***
	(0.0078)	(0.0017)	(0.0049)	(0.0010)
Age 5-7	-0.0382***	0.0477***	-0.0233***	0.0309***
	(0.0073)	(0.0016)	(0.0045)	(0.0010)
Age 8-12	-0.1258***	-0.0263***	-0.0781***	-0.0152***
	(0.0071)	(0.0015)	(0.0044)	(0.0009)
Age 13 or more	-0.2190***	-0.1024***	-0.1365***	-0.0627***
	(0.0074)	(0.0016)	(0.0046)	(0.0010)
N	2243011	34638723	2243011	34638723
pseudo R^2	0.0112	0.0318	0.0112	0.0320
pMSE	0.0041	0.0121	0.0041	0.0124

Note: An observation is a firm and a year of both synthetic and original databases. In all specifications, we include both time and industry fixed effects. Standard errors are in parentheses. In this table, we use 2015 vintage of LEAP to create the synthetic database and drop last year observation of each firm. ***, **, and * indicate statistically significant coefficients at 1%, 5%, and 10% percent levels, respectively.

4.5 Regression Analysis

To assess how well the CanSynLBD captures variability in economic growth due to industry and firm age, we estimate the following dynamic panel data model:

$$ALU_{it} = \alpha + \theta ALU_{i,t-1} + \lambda Pay_{it} + Age_{it}^{T}\beta + \lambda_t + \alpha_s + \epsilon_{it} \qquad (2)$$

where ALU_{it} is the logarithm of average labour unit (ALU) of firm i in year t, $ALU_{i,t-1}$ is the logarithm of last year's average labour unit (ALU) of firm i, Pay_{it} is the logarithm of payroll of firm i in year t, Age_{it} is a vector of dummy variables for age of firm i in year t, λ_t is the year fixed effect, α_s is an unobserved

time-invariant industry-specific effect, and ϵ_{it} is the disturbance term of firm i in year t.

Table 5: Regression coefficients (OLS)

Independent Variables	**LEAP**		**CanSynLBD**	
	Private	Manufacturing	Private	Manufacturing
AR(1) Coefficient	0.2031***	0.2481***	0.3970***	0.4405***
	(0.0001)	(0.0005)	(0.0002)	(0.0007)
Ln Pay	0.7847***	0.7300***	0.5481***	0.5228***
	(0.0001)	(0.0005)	(0.0002)	(0.0006)
Age 3-4	-0.1202***	-0.1717***	-0.1223***	-0.2340***
	(0.0003)	(0.0014)	(0.0004)	(0.0016)
Age 5-7	-0.1260***	-0.1891***	-0.1235***	-0.2507***
	(0.0003)	(0.0014)	(0.0004)	(0.0016)
Age 8-12	-0.1268***	-0.1973***	-0.1169***	-0.2551***
	(0.0003)	(0.0013)	(0.0004)	(0.0016)
Age 13 or more	-0.1246***	-0.1992***	-0.1101***	-0.2577***
	(0.0003)	(0.0014)	(0.0004)	(0.0017)
N	15708195	1015293	13573225	959764
R^2	0.9696	0.9743	0.9444	0.9523

Note: In all specifications, we include both year and industry fixed effects. Standard errors are in parentheses. *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. In this table, we use the 2015 vintage of LEAP and drop last year observation of each firm. ***, **, and * indicate statistically significant coefficients at 1%, 5%, and 10% percent levels, respectively.

We estimate the model separately on LEAP and CanSynLBD data for the private and manufacturing sectors and find that the CansynLBD data provides similar predictions to LEAP data (Tables 5).

BD: Make sure we use the same table footnote when appropriate; this should be coded.

Table 6: Regression coefficients (Dynamic)

Independent Variables	**LEAP**		**CanSynLBD**	
	Private	Manufacturing	Private	Manufacturing
AR(1) Coefficient	0.0805***	0.1189***	0.5722***	0.5425***
	(0.0003)	(0.0018)	(0.0024)	(0.0084)
Ln Pay	0.8991***	0.8523***	0.4101***	0.4302***
	(0.0002)	(0.0015)	(0.0018)	(0.0067)
Age 3-4	-0.0450***	-0.0797***	-0.2075***	-0.2972***
	(0.0002)	(0.0014)	(0.0010)	(0.0051)
Age 5-7	-0.0438***	-0.0860***	-0.2129***	-0.3162***
	(0.0002)	(0.0015)	(0.0011)	(0.0059)
Age 8-12	-0.0418***	-0.0923***	-0.2187***	-0.3294***
	(0.0003)	(0.0017)	(0.0013)	(0.0070)
Age 13 or more	-0.0379***	-0.0898***	-0.2318***	-0.3414***
	(0.0003)	(0.0019)	(0.0015)	(0.0080)
N	15708195	1015293	13573225	959764
m2	-14.5000	-2.2200	-27.5400	-9.4400
Sargan test	6.9e+04	4.6e+03	1.5e+04	1.5e+03
df of Sargan Test	252.0000	252.0000	252.0000	252.0000
P value of Sargan test	0.0000	0.0000	0.0000	0.0000

Note: In this table, $m2$ is the Arellano-Bond test for zero autocorrelation in first-differenced errors for order two. *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. In this graph, we use the 2015 vintage of LEAP and drop last year observation of each firm. Standard errors are in parentheses. ***, **, and * indicate statistically significant coefficients at 1%, 5%, and 10% percent levels, respectively.

As ALU_{st-1} is correlated with α_s because ALU_{st-1} is a function of α_s, OLS estimators are biased and inconsistent. To take this endogeneity bias into account, we use the estimation method from Arellano and Bond [1] and find similar predictions (Table 6). To check the validity of the model, we use two tests. First, to test for autocorrelation, we use the test $m2$ by Arellano and Bond [1]. In the table, we report the z test statistic for $m2$ test for zero autocorrelation in the first-differenced errors of order two. Second, we use the Sargan test to verify the validity of instrument subsets (showned in the last three rows in the

table).

We furthermore estimate the model using the system GMM method proposed by Arellano and Bover [2] and Blundell and Bond [4] and find similar predictions as before (Table 7).

Table 7: Regression coefficients (Dynamic - system GMM)

Independent Variables	**LEAP**		**CanSynLBD**	
	Private	Manufacturing	Private	Manufacturing
AR(1) Coefficient	0.0978***	0.1614***	0.5111***	0.5780***
	(0.0002)	(0.0014)	(0.0008)	(0.0041)
Ln Pay	0.8854***	0.8161***	0.4562***	0.4022***
	(0.0002)	(0.0012)	(0.0006)	(0.0033)
Age 3-4	-0.0555***	-0.1097***	-0.1828***	-0.3177***
	(0.0002)	(0.0012)	(0.0004)	(0.0028)
Age 5-7	-0.0558***	-0.1201***	-0.1860***	-0.3408***
	(0.0002)	(0.0013)	(0.0005)	(0.0031)
Age 8-12	-0.0548***	-0.1298***	-0.1875***	-0.3583***
	(0.0002)	(0.0014)	(0.0005)	(0.0036)
Age 13 or more	-0.0524***	-0.1317***	-0.1943***	-0.3747***
	(0.0002)	(0.0016)	(0.0006)	(0.0041)
N	15708195	1015293	13573225	959764
m2	-11.4300	1.3900	-41.6000	-7.6700
Sargan test	7.7e+04	6.3e+03	1.8e+04	1.7e+03
df of Sargan Test	274.0000	274.0000	274.0000	274.0000
P value of Sargan test	0.0000	0.0000	0.0000	0.0000

Note: An observation is a firm and a year. In this table, $m2$ is the Arellano-Bond test for zero autocorrelation in first-differenced errors for order two. $LEAP$ is the Longitudinal Employment Analysis Program and $CanSynLBD$ is the Canadian synthetic database based on LEAP. In this table, we use 2015 vintage of LEAP and drop last year observation of each firm. Standard errors are in parentheses. ***, **, and * indicate statistically significant coefficients at 1%, 5%, and 10% percent levels, respectively.

We also estimate above dynamic panel data model with a first-order moving average using appropriate instruments for both level and difference equation as

proposed by Arellano and Bover [2] and Blundell and Bond [4]:

$$ALU_{it} = \alpha + \theta ALU_{i,t-1} + \lambda Pay_{it} + Age_{it}^{T}\beta + \lambda_t + \alpha_s + \epsilon_{it} + \gamma\epsilon_{it-1} \quad (3)$$

Table 8 shows that the CansynLBD provides similar predictions to the LEAP.

Table 8: Regression coefficients (Dynamic - system GMM with MA(1))

Independent Variables	**LEAP**		**CanSynLBD**	
	Private	Manufacturing	Private	Manufacturing
AR(1) Coefficient	0.2005***	0.2821***	0.4850***	0.5737***
	(0.0007)	(0.0040)	(0.0012)	(0.0059)
Ln Pay	0.8044***	0.7135***	0.4760***	0.4056***
	(0.0005)	(0.0034)	(0.0009)	(0.0046)
Age 3-4	-0.1245***	-0.2033***	-0.1716***	-0.3158***
	(0.0005)	(0.0032)	(0.0006)	(0.0037)
Age 5-7	-0.1328***	-0.2264***	-0.1733***	-0.3389***
	(0.0005)	(0.0035)	(0.0006)	(0.0043)
Age 8-12	-0.1383***	-0.2454***	-0.1731***	-0.3560***
	(0.0006)	(0.0039)	(0.0007)	(0.0051)
Age 13 or more	-0.1441***	-0.2586***	-0.1774***	-0.3717***
	(0.0006)	(0.0042)	(0.0008)	(0.0058)
N	15708195	1015293	13573225	959764
m2	8.2000	7.0600	-40.0300	-6.6400
Sargan test	2.8e+04	2.3e+03	1.7e+04	1.3e+03
df of Sargan Test	251.0000	251.0000	251.0000	251.0000
P value of Sargan test	0.0000	0.0000	0.0000	0.0000

Note: An observation is a firm and a year. In this table, $m2$ is the Arellano-Bond test for zero autocorrelation in first-differenced errors for order two. $LEAP$ is the Longitudinal Employment Analysis Program and $CanSynLBD$ is the Canadian synthetic database based on LEAP. In this table, we use 2015 vintage of LEAP and drop last year observation of each firm. Standard errors are in parentheses. ***, **, and * indicate statistically significant coefficients at 1%, 5%, and 10% percent levels, respectively.

4.6 Confidentiality protection

In this section, we estimate the probability that the synthetic first year equals the true first year, given the synthetic first year.Tables 9 and 10 show that these probabilities are quite low except for the first year.This is because of censoring and lack of previous information.

BD: I don't understand why the dependant variable has no indices?

Table 9: Observed firm births given synthetic births (private)

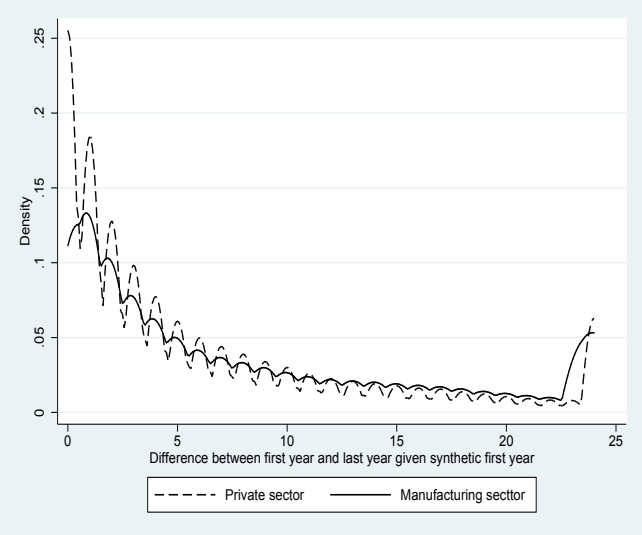
First (Birth) Year		% of Births over NAICS		
Synthetic	Actual	Minimum	Mean	Maximum
1991	1991	0.00	27.69	83.02
1992	1992	0.00	3.37	11.11
1993	1993	0.00	3.79	33.33
1994	1994	0.00	3.73	33.33
1995	1995	0.00	3.86	20.00
1996	1996	0.00	4.25	33.33
1997	1997	0.00	4.10	16.94
1998	1998	0.00	4.41	25.00
1999	1999	0.00	4.23	33.33
2000	2000	0.00	3.41	25.00
2001	2001	0.00	2.73	22.22
2002	2002	0.00	2.65	25.00
2003	2003	0.00	2.22	10.00
2004	2004	0.00	2.60	17.86
2005	2005	0.00	2.71	20.00
2006	2006	0.00	2.83	50.00
2007	2007	0.00	2.90	33.33
2008	2008	0.00	2.38	20.00
2009	2009	0.00	2.47	50.00
2010	2010	0.00	2.12	33.33
2011	2011	0.00	2.65	50.00
2012	2012	0.00	2.41	20.00
2013	2013	0.00	2.48	25.00
2014	2014	0.00	2.23	20.00
2015	2015	0.00	2.15	33.33

BD: Are those coefficients or marginal effects? Does it make sense to show coefficients? Or are we interested only in the last row? If we are interested in the coefficient, why is there no discussion of those?

Table 10: Observed firm births given synthetic births (manufacturing)

First (Birth) Year		% of Births over NAICS		
Synthetic	Actual	Minimum	Mean	Maximum
1991	1991	4.76	31.64	52.03
1992	1992	0.00	3.32	10.53
1993	1993	0.00	3.97	33.33
1994	1994	0.00	4.21	33.33
1995	1995	0.00	4.41	20.00
1996	1996	0.00	5.36	33.33
1997	1997	0.00	4.09	16.94
1998	1998	0.00	5.46	25.00
1999	1999	0.00	5.27	33.33
2000	2000	0.00	3.39	25.00
2001	2001	0.00	2.19	10.00
2002	2002	0.00	2.45	25.00
2003	2003	0.00	1.71	10.00
2004	2004	0.00	2.07	17.86
2005	2005	0.00	1.92	16.67
2006	2006	0.00	2.49	50.00
2007	2007	0.00	1.74	14.29
2008	2008	0.00	1.60	20.00
2009	2009	0.00	1.60	20.00
2010	2010	0.00	1.34	33.33
2011	2011	0.00	2.43	50.00
2012	2012	0.00	1.93	20.00
2013	2013	0.00	1.61	20.00
2014	2014	0.00	1.71	14.29
2015	2015	0.00	1.41	14.29

Figure 16: The difference between first and last year given synthetic first year



5 Conclusion and Extensions

In this paper, we adapt and implement algorithms used to create the U.S. synthetic data for LBD to create Canadian synthetic data for LEAP. We show the newly created data set is analytically valid for a wide range of statistical analyses, as well as provide evidence on confidentiality properties.

5.1 Addition of variables that are not analytically valid

5.2 Addition of analytically valid variables

Capital stock or revenue for incorporated

compute overlap interval

BD: That is some strange wording

BD: Are we worried about this?

A Analytical validity

A.1 Confidence interval for gross employment and other measures

We compute the standard error for gross employment as follows. We consider gross employment E to be the sum of firm employments E_j:

$$E = \sum_j E_j \tag{4}$$

Average firm employment $\bar{E} = \frac{E}{N_j}$ is assumed to be normally distributed, with standard deviation $\sigma_{\bar{E}}$. We compare the synthetic and the confidential data for gross employment, including error bands.

A.2 Confidence interval overlap measures

More generally, the question as to the statistical precision of the results obtained from the synthetic data can be assessed. For this purpose, we computed the overlap of parameter estimates as suggested by [8]. We compute the *interval overlap measure* $J_{k,m}$ for parameter k in model m. Consider the overlap of confidence intervals (L, U) for $\beta_{k,m}$ (estimated from the confidential data) and (L^*, U^*) for $\beta^*_{k,m}$ (from the synthetic data). Let $L^{over} = \max(L, L^*)$ and $U^{over} = \min(U, U^*)$. Then the average overlap in confidence intervals is

$$J^*_{k,m} = \frac{1}{2}\left[\frac{U^{over} - L^{over}}{U - L} + \frac{U^{over} - L^{over}}{U^* - L^*}\right]$$

We then average $J^*_{k,m}$ over all estimated models and parameters, by validation request. The correct counterfactual involved running these validation requests against synthetic data that does not claim analytical validity, such as synthetic data generated from uni-dimensional distributions of variables. Re-

sults are pending.

BD: ?

A.3 Other models

Possible papers:

-
- Bartelsman, Haltiwanger, and Scarpetta [3] use a cross-country dataset to study average post-entry behavior of young firms.

References

[1] Manuel Arellano and Stephen Bond. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations". In: *Review of Economic Studies* 58.2 (1991), pp. 277–297. URL: https://EconPapers.repec.org/RePEc:oup:restud:v:58:y:1991:i:2:p:277-297..

[2] Manuel Arellano and Olympia Bover. "Another look at the instrumental variable estimation of error-components models". In: *Journal of Econometrics* 68.1 (1995), pp. 29–51. URL: https://EconPapers.repec.org/RePEc:eee:econom:v:68:y:1995:i:1:p:29-51.

[3] Eric Bartelsman, John Haltiwanger, and Stefano Scarpetta. "Measuring and Analyzing Cross-country Differences in Firm Dynamics". In: Timothy Dunne, J. Bradford Jensen, and Mark J. Roberts. *Producer Dynamics: New Evidence from Micro Data*. University of Chicago Press, Jan. 2009, pp. 15–76. URL: http://www.nber.org/chapters/c0480.

[4] Richard Blundell and Stephen Bond. "Initial conditions and moment restrictions in dynamic panel data models". In: *Journal of Econometrics* 87.1 (Aug. 1998), pp. 115–143. URL: https://ideas.repec.org/a/eee/econom/v87y1998i1p115-143.html.

[5] Steven J. Davis, John C. Haltiwanger, and Scott Schuh. *Job creation and destruction*. Cambridge, MA: MIT Press, 1996.

[6] Jorg Drechsler and Lars Vilhuber. *A First Step Towards A German Synlbd: Constructing A German Longitudinal Business Database*. Working Papers 14-13. Center for Economic Studies, U.S. Census Bureau, Feb. 2014. URL: https://ideas.repec.org/p/cen/wpaper/14-13.html.

[7] Ron S Jarmin and Javier Miranda. *The Longitudinal Business Database*. Working Papers 02-17. Center for Economic Studies, U.S. Census Bureau, July 2002. URL: https://ideas.repec.org/p/cen/wpaper/02-17.html.

[8] A. F. Karr et al. "A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality". In: *The American Statistician* 60.3 (2006), pp. 1–9. DOI: 10.1198/000313006X124640.

[9] Satkartar K. Kinney, Jerome P. Reiter, and Javier Miranda. *Improving The Synthetic Longitudinal Business Database*. Working Papers 14-12. Center for Economic Studies, U.S. Census Bureau, Feb. 2014. URL: https://ideas.repec.org/p/cen/wpaper/14-12.html.

[10] Satkartar K. Kinney et al. "Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database". In: *International Statistical Review* 79.3 (Dec. 2011), pp. 362–384. DOI: j.1751-5823.2011.00152.x. URL: https://ideas.repec.org/a/bla/istatr/v79y2011i3p362-384.html.

[11] Petr Sedláček and Vincent Sterk. "The Growth Potential of Startups over the Business Cycle". In: *American Economic Review* 107.10 (Oct. 2017), pp. 3182–3210. DOI: 10.1257/aer.20141280. URL: http://www.aeaweb.org/articles?id=10.1257/aer.20141280.

[12] Joshua Snoke and Aleksandra Slavkovic. "pMSE Mechanism: Differentially Private Synthetic Data with Maximal Distributional Similarity: UNESCO Chair in Data Privacy, International Conference, PSD 2018, Valen-

cia, Spain, September 26-28, 2018, Proceedings". In: Jan. 2018, pp. 138–159. ISBN: 978-3-319-99770-4. DOI: 10.1007/978-3-319-99771-1_10.

[13] Simon D. Woodcock and Gary Benedetto. "Distribution-preserving statistical disclosure limitation". In: *Computational Statistics & Data Analysis* 53.12 (2009), pp. 4228–4242. ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2009.05.020. URL: http://www.sciencedirect.com/science/article/pii/S0167947309002011.