Applying Data Synthesis for Longitudinal Business Data across Three Countries

M. Jahangir Alam, Benoit Dostie, Jörg Drechsler, Lars Vilhuber*

January 26, 2020

Abstract

Data on businesses collected by statistical agencies are challenging to protect. Many businesses have unique characteristics, and distributions of employment, sales, and profits are highly skewed. Attackers wishing to conduct identification attacks often have access to much more information than for any individual. As a consequence, most disclosure avoidance mechanisms fail to strike an acceptable balance between usefulness and confidentiality protection. Detailed aggregate statistics by geography or detailed industry classes are rare, public-use microdata on business are virtually inexistant, and access to confidential microdata can be burdensome.

Synthetic microdata have been proposed as a secure mechanism to publish microdata, as part of a broader discussion of how to provide broader access to such datasets to researchers In this article, we document an experiment to create analytically valid synthetic data, using the exact same model and methods previously employed for the United States, for data from two different countries: Canada (Longitudinal Employment Analysis Program (LEAP)) and Germany (Establishment History Panel (BHP)). We assess utility, protection, and provide an assessment of the feasibility of extending such an approach in a cost-effective way to other data.

*Corresponding author: Lars Vilhuber, lars.vilhuber@cornell.edu. Alam was a part-time employee of Statistics Canada when this research was conducted. Drechsler is an employee of the Institut für Arbeitsmarktforschung, IAB. The opinions expressed here are those of the authors, and do not reflect the opinions of any of the statistical agencies involved. All results were reviewed for disclosure risks by their respective custodians, and released to the authors. Vilhuber acknowledges funding through NSF Grants SES-1131848 and SES-1042181, and a grant from Alfred P. Sloan Grant (G-2015-13903). Alam and Dostie acknowledge funding through SSHRC Partnership Grant “Productivity, Firms and Incomes” to McMaster University and HEC Montréal. The creation of the Synthetic LBD was funded by NSF Grant SES-0427889.

Contents

1 Introduction

Data on businesses collected by statistical agencies are challenging to protect. Many businesses have unique characteristics, and distributions of employment, sales, and profits are highly skewed. Attackers wishing to conduct identification attacks often have access to much more information than for any individual. It is easy to find examples of firms and establishments that are so dominant in their industry or location that they would be immediately identified if data were publicly released that included their survey responses or administratively collected data. Finally, there are also greater financial incentives to identifying the particulars of some firms and their competitors.

As a consequence, most disclosure avoidance mechanisms fail to strike an acceptable balance between usefulness and confidentiality protection. Detailed aggregate statistics by geography or detailed industry classes are rare, public-use microdata on business are virtually inexistant,[1] and access to confidential microdata can be burdensome. It is not uncommon that access to establishment microdata, if granted at all, is provided through data enclaves (Research Data Centers), at headquarters of statistical agencies, or some other limited means, under strict security conditions. These restrictions on data access reduce the growth of knowledge by increasing the cost to researchers of accessing the data.

Synthetic microdata have been proposed as a secure mechanism to publish microdata (Drechsler 2011b, 2012; Council 2007; Jarmin, Louis, and Miranda 2014), based on suggestions and methods first proposed by Rubin (1993) and Little (1993). Such data are part of a broader discussion of how to provide broader access to such datasets to researchers (Bender 2009; Vilhuber 2013; Abowd and Lane 2004; Abowd and Schmutte 2015).[2] For business data, synthetic business microdata were released in the United States (Kinney et al. 2011b) and in Germany (Drechsler 2011c) in 2011. The former dataset, called Synthetic Longitudinal Business Database (LBD) (SynLBD), was released to an easily web-accessible computing environment (Abowd and Vilhuber 2010), and combined with a validation mechanism. By making disclosable synthetic microdata available through a remotely accessible data server, combined with a validation server, the SynLBD approach alleviates some of the access restrictions associated with economic data. The approach is mutually beneficial to both agency and researchers. Researchers can access public use servers at little or no cost, and can later validate their model-based inferences on the full confidential microdata.

In this article, we document an experiment to create analytically valid synthetic data, using the exact same model and methods previously used to create the SynLBD and described in Kinney et al. (2011b, henceforth KRRMJA) and Kinney et al. (2011a), applied to data from two different countries: Canada (Longitudinal Employment Analysis Program (LEAP)) and Germany (Establishment History Panel (BHP)). We describe

[1]See Guzman and Stern (2016) and Guzman and Stern (2020) for an example of scraped, public-use microdata.

[2] For a recent overview of some, see Vilhuber, Abowd, and Reiter (2016). See Drechsler (2011a) for a review of the theory and applications of the synthetic data methodology. Other access methods include secure data enclaves (e.g., research data centers of the U.S. Federal Statistical System, of the German Federal Employment Agency, others), and remote submission system systems. We will comment on the latter in the conclusion.

all three countries' data in Section 2.

In Canada, the Canadian Center for Data Development and Economic Research (CDER) was created in 2011 to allow Statistics Canada to make better use of its business data holdings, without compromising security. Secure access to business microdata for approved analytical research projects is done through a physical facility located in Statistics Canada's headquarters.

CDER implements many risks mitigation measures to alleviate the security risks specific to micro-level business data including limits on tabular outputs, centralized vetting, monitoring of programs logs. Access to the data is done through a Statistics Canada designed interface in which actual observations cannot be viewed. But the most significant barrier to access is the travel cost of coming to Ottawa.

The experiment is not so much in finding the *best* synthetic data method for each file, but rather to assess the effectiveness of using a 'pre-packaged' method to cost-effectively generate synthetic data. In particular, while we could have used newer implementations of methods combined with a pre-defined or automated model (Nowok, Raab, and Dibben 2016; Raab, Nowok, and Dibben 2018), we chose to use the exact SAS code used to create the original SynLBD. A brief synopsis of the KRRMJA method, and any adjustments we made to take into account structural data differences, are described in Section refsec:methodology.

In Germany,

Discuss German case

We verify the analytical validity of synthetic data files 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.

do modify according to what was exactly done in the respective cases of Canada and Germany

In each case, 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. The probability for the first year is higher because of censoring and lack of previous information.

2 Data

In this section, we briefly describe the structure of the original US dataset, then the specific Canadian and Germany databases, that served as inputs for the algorithm.

2.1 United States: Longitudinal Business Database (LBD)

The LBD (U.S. Census Bureau 2015) is created from the U.S. Census Bureau's Business Register (BR) by creating longitudinal links of establishments using name and address matching. The database has information on birth, death, location, industry, firm affiliation of employer establishments, and ownership by multi-establishment firms, as well as their employment over time, for nearly all sectors of the economy from 1976 through 2015 (as of this writing). It serves as a key linkage file as well as a research dataset in its own right for numerous research articles, as well as a tabulation input to the U.S. Census Bureau's Business Dynamics Statistics (U.S. Census Bureau 2017, BDS). Other statistics created from the underlying Business Register include the County Business Patterns (U.S. Census Bureau 2016a, CBP) and the Statistics of U.S. Businesses (U.S. Census Bureau 2016b, SBUSB). For a full description, readers should consult Jarmin and Miranda (2002). The key variables of interest for this experiment are birth and death dates, payroll, employment, and the industry coding of the establishment. Kinney, Reiter, and Miranda (2014b) explore a possible expansion of the synthesis methods described later to include location and firm affiliation. Note that information on payroll and employment does not come from individual-level wage records, as is the case for both the Canadian and German datasets described below, as well as for the Quarterly Workforce Indicators (Abowd et al. 2009) derived from the Longitudinal Employer-Household Dynamics (Vilhuber 2018, LEHD) in the United States. Thus, methods that connect establishments based on labor flows (Benedetto et al. 2007; Hethey and Schmieder 2010) are not employed. We also note that payroll is the cumulative sum of wages paid over the entire calendar year, whereas employment is measured as of March 12 of each year.

2.2 Canada: Longitudinal Employment Analysis Program (LEAP)

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

To construct the LEAP, Statistics Canada uses three sources of information: (1) T4 administrative data from the Canada Revenue Agency (CRA), (2) data from Statistics Canada's Business Register, and (3) data from Statistics Canada's Survey of Employment, Payrolls and Hours (SEPH).

T4 In general, all employers in Canada provide employees with a T4 slip if they paid employment income, taxable allowances and benefits, or any other remuneration in any calendar year. The T4 information is reported to the tax agency, which in turn provides this information to Statistics Canada.

BR The Business Register is Statistics Canada's central repository of baseline information on business and institutions operating in Canada. It is used as the survey frame for all business related data sets.

SEPH The objective of the SEPH is to provide monthly information on the level of earnings, the number of jobs, and hours worked by detailed industry at the national and provincial levels. To do so, it combines a census of approximately one million payroll deductions provided by the CRA, and the Business Payrolls Survey, a sample of 15,000 establishments.

The core LEAP contains four variables (1) A Longitudinal Business Register Identifier (LBRID), (2) Industry, (3) Employment and (4) Payroll.

1. The LBRID uniquely identifies each enterprise and is derived from the Business Register. To avoid "false" deaths and births due to mergers, restructuring or changes in reporting practices, Statistics Canada uses employment flows. Similar to Benedetto et al. (2007) and Hethey and Schmieder (2010), the method compares cluster of workers in each newly identified enterprise with all the clusters of workers in firms from the previous year. This comparison yields a new identifier (LBRID) derived from those of the BR.

2. The industry information comes from the BR for single-industry firms. If a firm operates in multiple industries, information on payroll from the SEPH is used to identify the industry in which the firm pays the highest payroll. Prior to 1991, information on industry was based on the SIC but it is now based on the North American Industrial Classification System (NAICS) at the four-digits level.

3. Employment is measured either using Individual Labour Unit (ILU) or Average Labour Unit (ALU). ALUs are obtained by dividing the business annual payroll (as the sum of T4 slips income issued in the year) and diving by the average annual earnings in its industry/province/class category computed using the SEPH. ILUs are a head count of the number of T4 issued by the enterprise, with employees working for multiple employers split proportionately across firms according to their total annual payroll earned in each firm.

4. Finally, the firm's payroll comes from the sum of all T4s as reported to the CRA.

With that information, the LEAP is the only data set in Canada that allows research on a variety of themes, like employment growth, industry turnover, firm survival, job creation and job destruction, etc.

2.3 Germany: Establishment History Panel (BHP)

Because the Institute for Employment Research is affiliated with the German labor ministry, it collects very little data at the business or establishment level.[3]

[3]Exceptions are the IAB Establishment Panel (Bundesagentur für Arbeit 2013c) and the IAB Job Vacancy Survey (Bundesagentur für Arbeit 2013b).

The core database is the German Social Security Data (GSSD), which is based on the integrated notification procedure for the health, pension and unemployment insurances, introduced in 1973. Employers report information on all their employees, by establishment. Aggregating this information via an establishment identifier yields Establishment History Panel (Bundesagentur für Arbeit 2013a, German abbreviation: BHP). We used data from 1975 until 2008, which at the time this project started was the most current data available for research. Information for the former Eastern German States is limited to the years 1992-2008.

Due to the purpose and structure of the GSSD, some variables present in the LBD are not available on the BHP. Firm-level information is not captured, and it is thus not known whether establishments are part of a multi-establishment employer. In 1999, reporting requirements were extended to all establishments; prior to that date, only establishments that had at least one employee covered by social security on the reference date June 30 of each year were subject to filing requirements. Payroll and employment are both based on a reference date of June 30, and are thus consistent point-in-time measures. Drechsler and Vilhuber (2014b) describe adjustments made to the BHP for this project, including estimating full-year payroll, creating time-consistent industry identifiers, and applying employment flow methods (Hethey and Schmieder 2010) to adjust for spurious births and deaths in establishment identifiers.

LARS STOPPED HERE

3 Methodology

The Synthetic LBD is derived from the LBD as a partially synthetic database with analytic validity, by synthesizing the life-span of establishments, as well as the evolution of their employment, conditional on industry. Geography is not synthesized, but is suppressed from the released file. The current version 2.0 is based on the Standard Industrial Classification (SIC) and extends through 2000. Work currently underway using the existing methodology will extend the data through 2010, using NAICS, and newer imputation methodology (Version 3) is under development (see paper by Kinney and Reiter in this issue) to improve the analytic validity and extend the imputation to additional variables. In this paper, when we refer to the "SynLBD algorithms", we refer to Version 2.

3.1 Overview

There is growing demand for firm-level data allowing detailed studies of firm dynamics. Recent examples include Bartelsman, Haltiwanger, and Scarpetta (2009) who use cross-country firm-level data to study average post-entry behavior of young firms. Sedláček and Sterk (2017) use the Business Dynamics Statistics (BDS) to show the role of firm size in firm dynamics. 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. (2011b) 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 (2002) for details). The variables currently available in the LBD are industry, annual payroll, employment, geography (county), birth year, death year, year, and firm structure (multiunit status).

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 (2009)).

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 (2014a)).

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 (2014a) 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?

JA: I think we could use it. However, I did use a few times.

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 less than ten observations in a given year.

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.

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 less than ten observations in a given year. We do not synthesize the public sector (NAICS 61, 62, and 91).

4 Analytical validity

4.1 Firm Characteristics

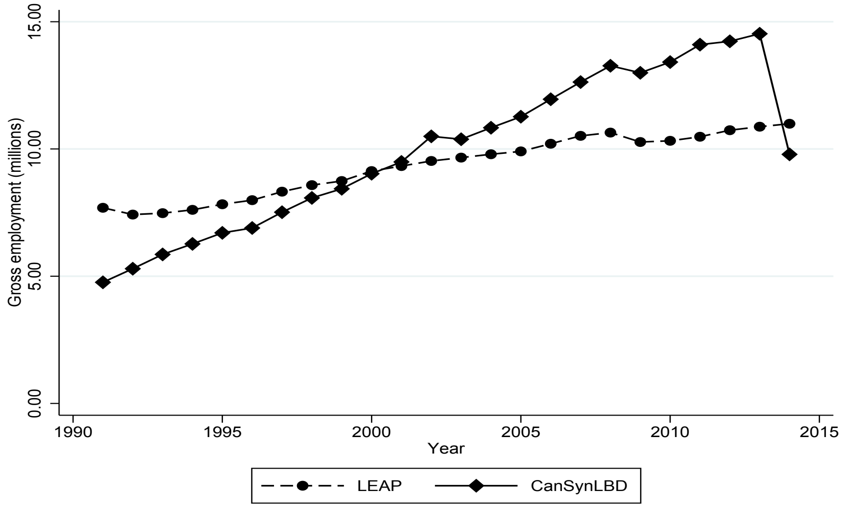
The CanSynLBD and LEAP generally provide comparable inferences on aggregate means and correlations. For example, Figures 1 and 2 show that gross employment levels for each year in the CanSynLBD are very close to those in the LEAP. However, the manufacturing sector shows closer patterns than the private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

[4] The private sector comprises all industries including the manufacturing sector except the public sector (NAICS 61, 62, and 91)

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

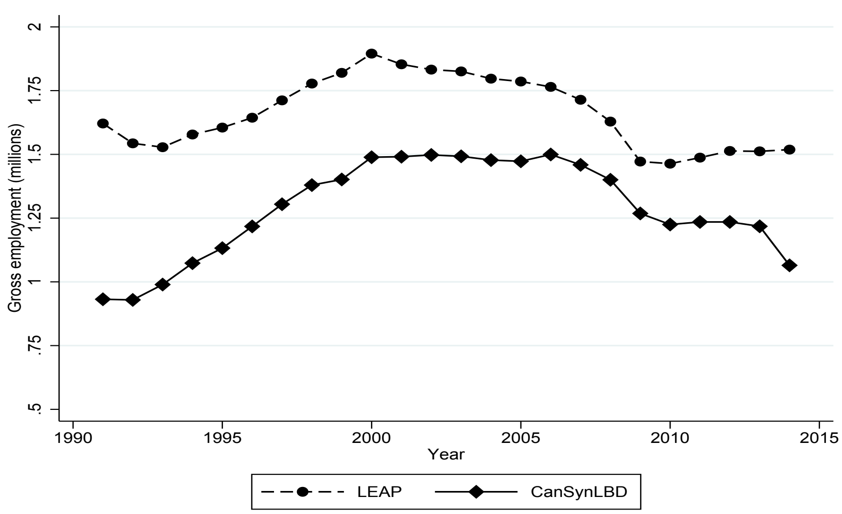
JA: I checked be-

Figure 1: Gross employment level by year (private)



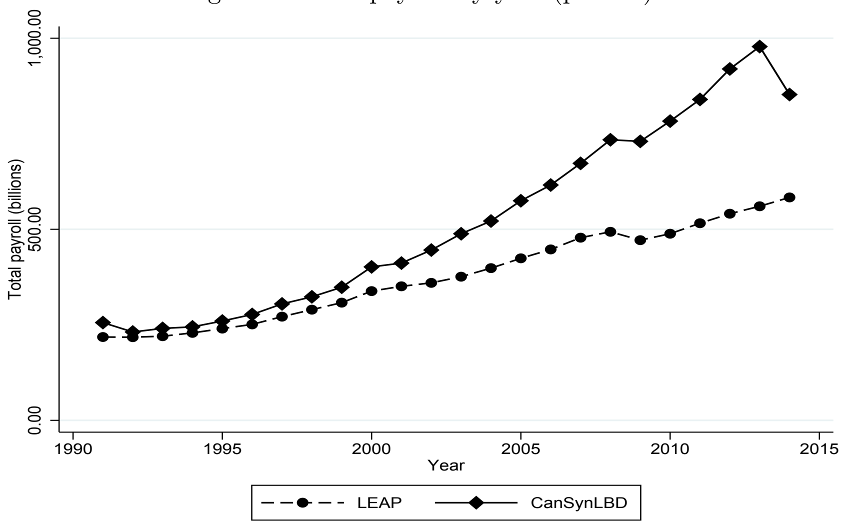
Note: *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. Here, we use the 2015 vintage of LEAP and drop the last year observations.

Figure 2: Gross employment level by year (manufacturing)



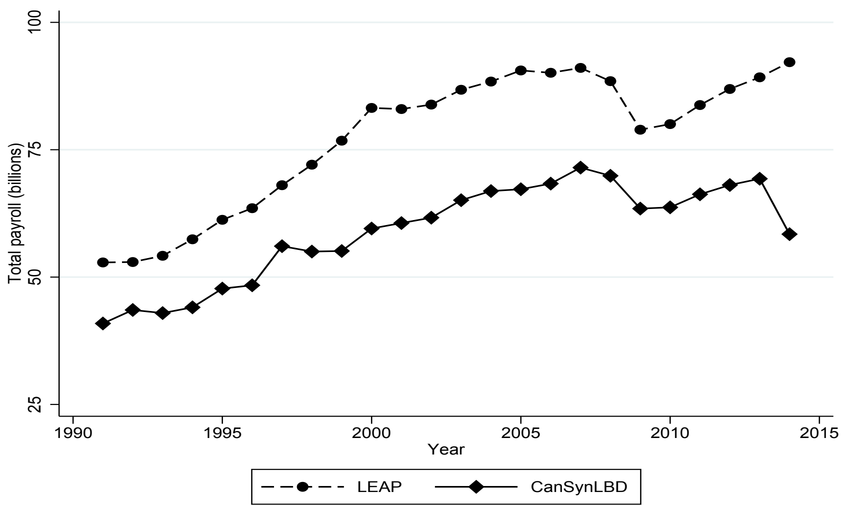
Note: *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. Here, we use the 2015 vintage of LEAP and drop the last year observations.

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. Here, we use the 2015 vintage of LEAP and drop the last year observations.

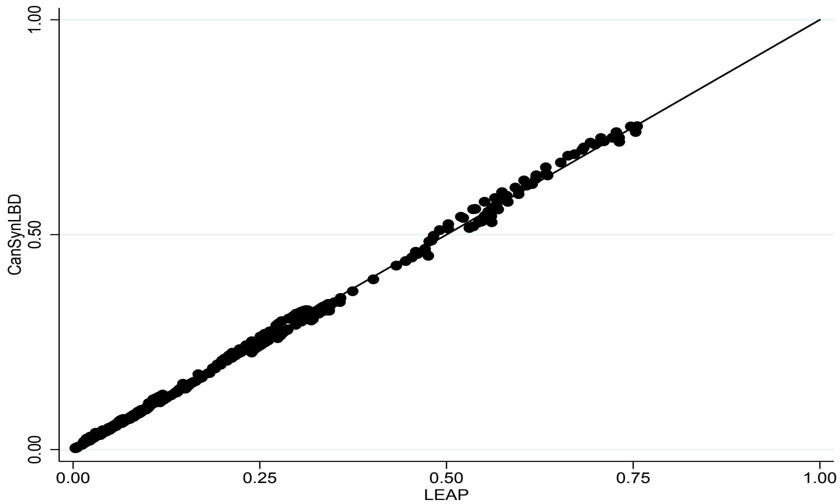
Figure 4: Total payroll by year (manufacturing)



Note: $LEAP$ is the Longitudinal Employment Analysis Program and $CanSynLBD$ is the Canadian synthetic database based on LEAP. Here, we use the 2015 vintage of LEAP and drop the last year observations.

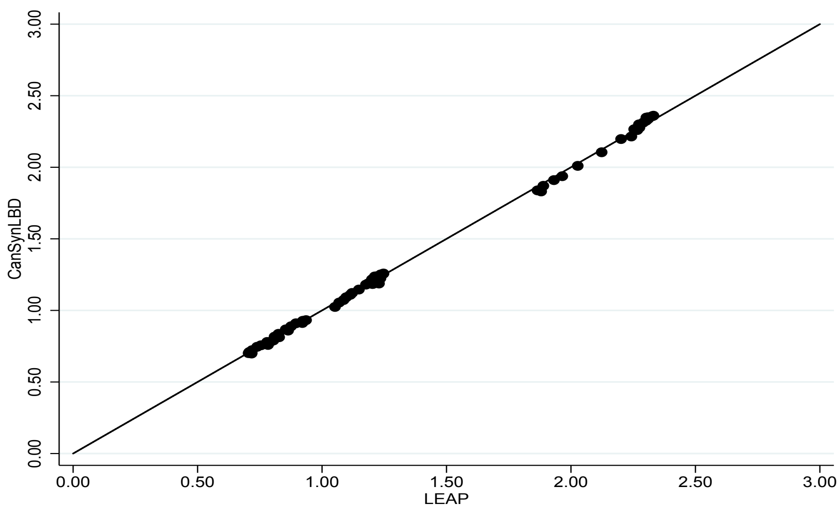
Figures 5 and 6 plot the share of firms by two-digit industry and year for both the CanSynLBD and the LEAP databaseand show that those shares clustering along the 45-degree line.

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. Here, we use the 2015 vintage of LEAP and drop the last year observations.

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

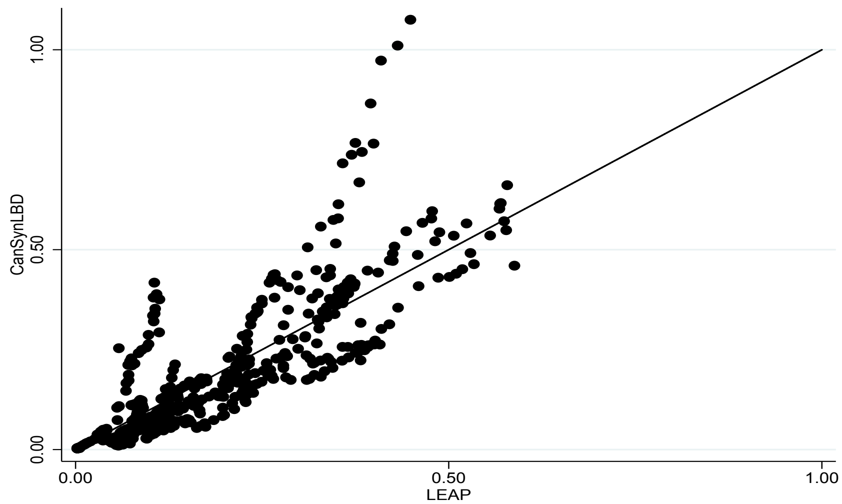


Note: *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. Here, we use the 2015 vintage of LEAP and drop the last year observations.

Figures 7 and 8 plot the share of employment by two-digit industry and year for both the CanSynLBD and the LEAP database [5] 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 more clustering along the 45-degrees.

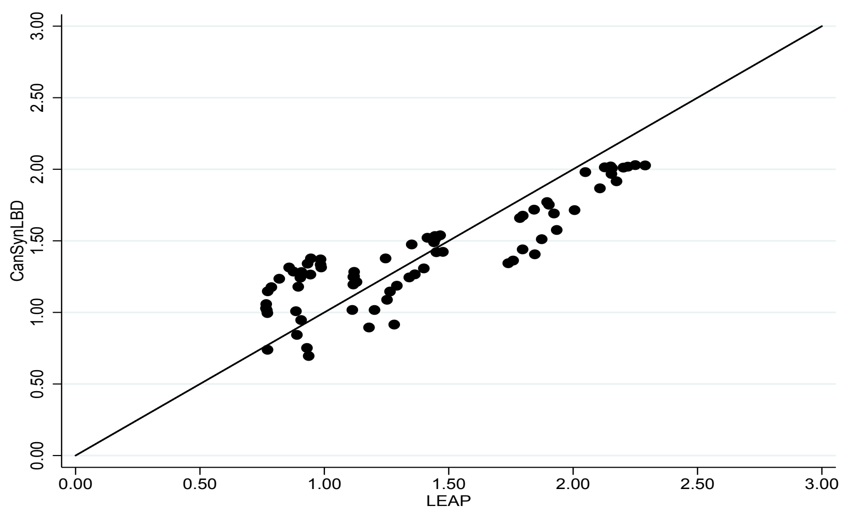
[5]We define the share of employment as $x_{its} = X_{its}/\sum_i \sum_t X_{its}$, where i are two-digit NAICS industries, t are the years in-sample, s indicates whether it is in the synthetic or confidential data, and X_{its} is the total employment for industry i and year t for either the synthetic or confidential data s.

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. Here, we use the 2015 vintage of LEAP and drop the last year observations.

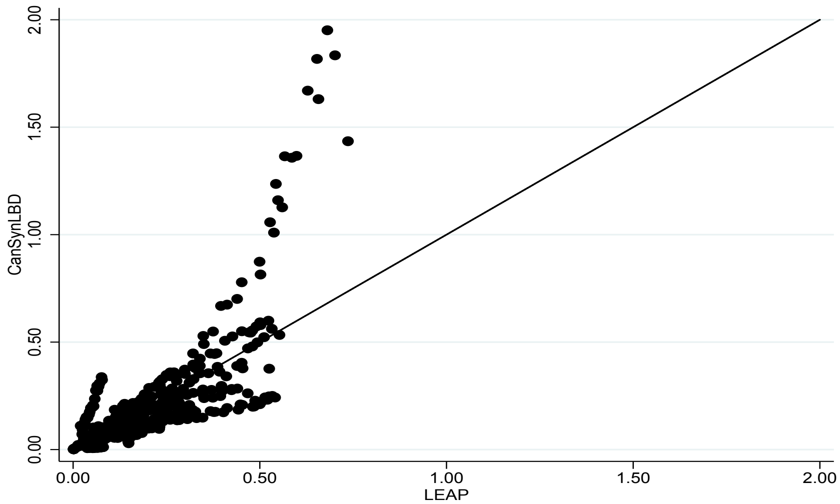
Figure 8: Share of employment by NAICS two-digit and year (manufacturing)



Note: $LEAP$ is the Longitudinal Employment Analysis Program and $CanSynLBD$ is the Canadian synthetic database based on LEAP. Here, we use the 2015 vintage of LEAP and drop the last year observations.

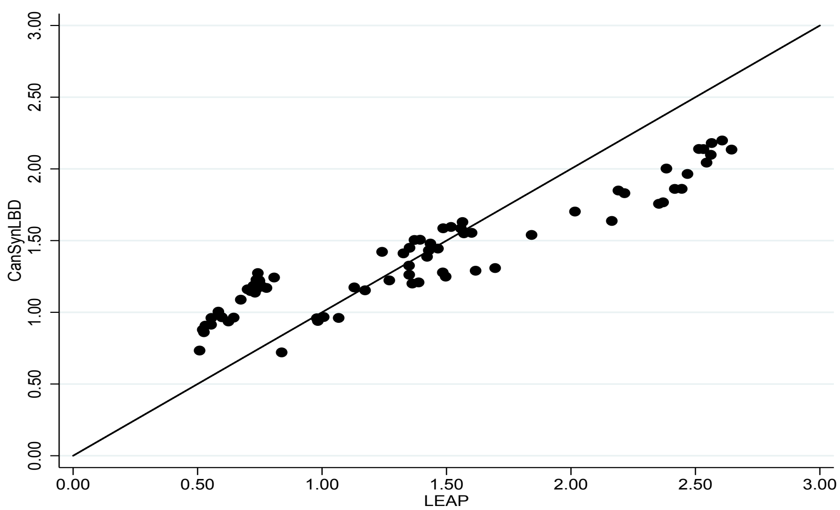
Figures 9 and 10 plot the share of payroll by two-digit industry and year for both CanSynLBD and LEAP database and show that those shares do not cluster along the 45-degree line. Again, we do notice that for the share of employment for the manufacturing sector, we do observe more clustering along the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

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. Here, we use the 2015 vintage of LEAP and drop the last year observations.

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. Here, we use the 2015 vintage of LEAP and drop the last year observations.

4.2 Firm Dynamics

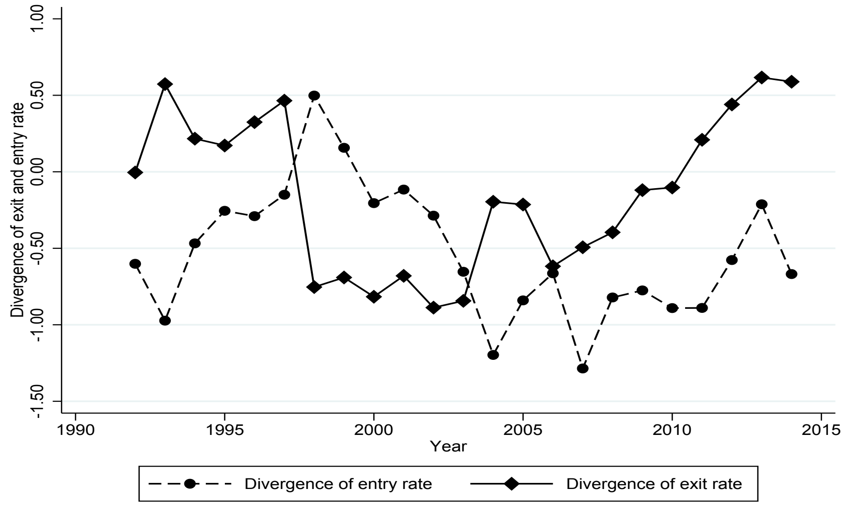
To assess how well the CanSynLBD captures firm dynamics, we also compute entry and exit rates of the private sector by year. Table 3 shows that those rates for CanSynLBD are similar to LEAP database. To show further those rates are similar, we compute the divergence of entry rate as the entry rate of CanSynLBD net the entry rate of LEAP as well as the divergence of exit rate as the exit rate of CanSynLBD net the exit rate of LEAP (see Figure 11).

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. Here, we use the 2015 vintage of LEAP and drop the last year observations.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

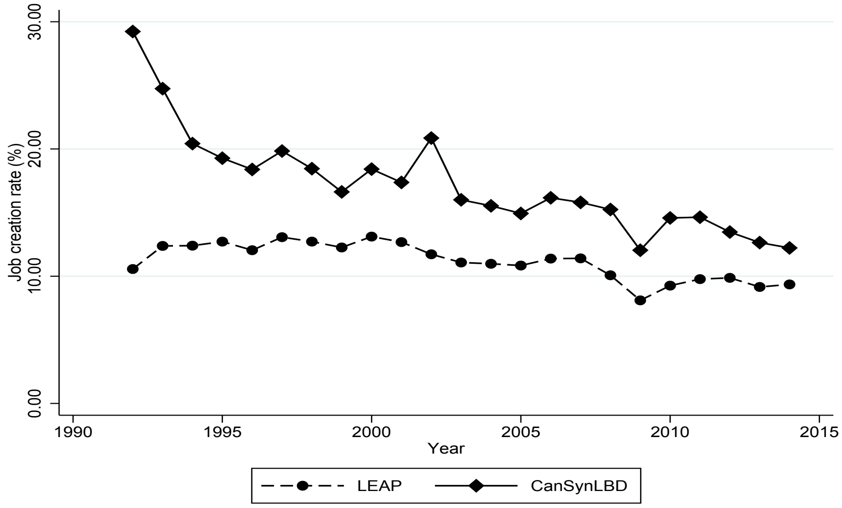


Note: *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. Here, we use the 2015 vintage of LEAP and drop the last year observations.We calculate the divergence of entry rate as the entry rate of CanSynLBD net the entry rate of LEAP and the divergence of exit rate as the exit rate of CanSynLBD net the exit rate of LEAP.

4.3 Dynamics of Job Flows

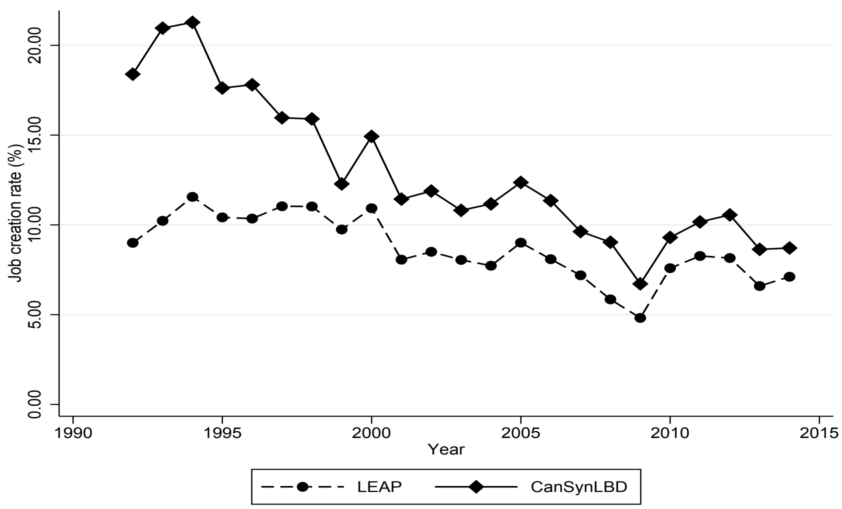
One of the most important applications of LEAP is to generate statistics that describe job flows. Following Davis, Haltiwanger, and Schuh 1996, the job creation is defined as the sum of all employment gains from expanding firms from year $t-1$ to year t including entry firms. The job destruction rate is defined as the sum of all employment losses from contracting 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).

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. Here, we use the 2015 vintage of LEAP and drop the last year observations.

Figure 13: Job creation rate by year (manufacturing)

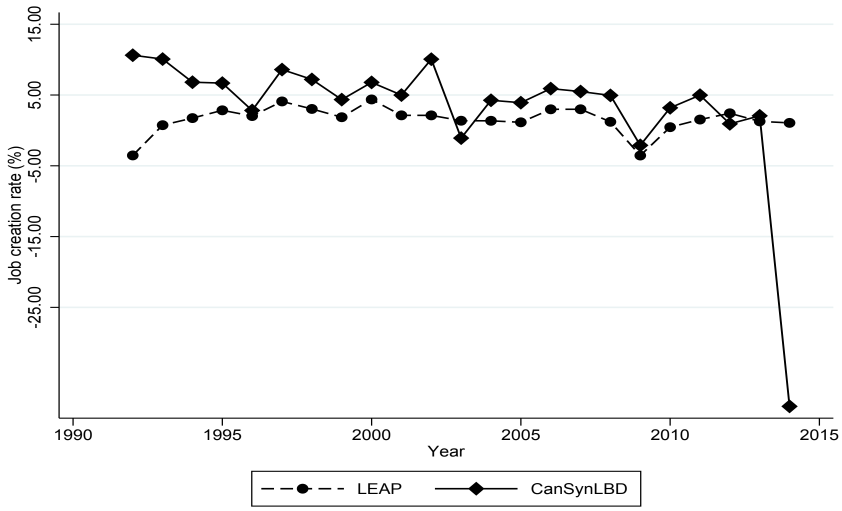


Note: *LEAP* is the Longitudinal Employment Analysis Program and *CanSynLBD* is the Canadian synthetic database based on LEAP. Here, we use the 2015 vintage of LEAP and drop the last year observations.

LV regraph, dropping last year

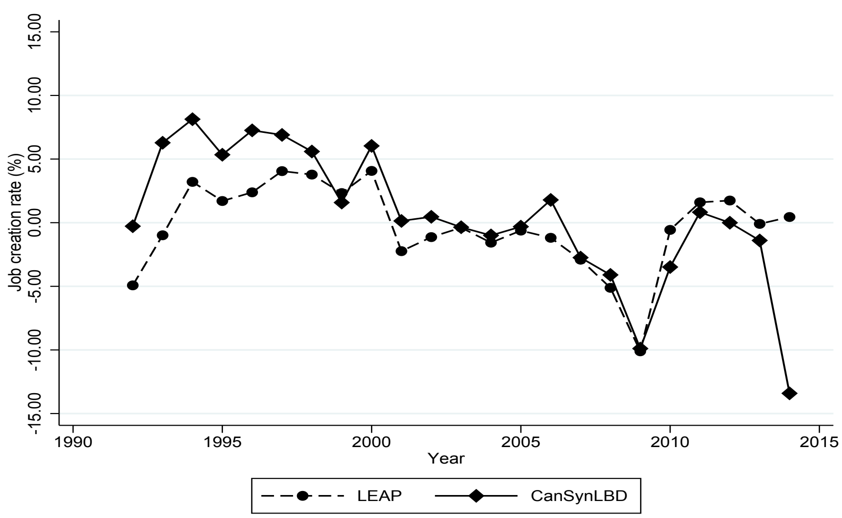
JA: Should we mention this in the text including reasons if we drop the last year here?

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. Here, we use the 2015 vintage of LEAP and drop the last year observations.

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 the high distributional similarity between synthetic data and confidential data.

We follow the method by Snoke and Slavkovic (2018) 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: I don't understand why the dependant variable has no indices?

JA: This is an indicator variable of 1 for the synthetic database and 0 for the confidential database. In this case, we need to add one more index in all variables as well.

BD: Are those coefficients or marginal effects? Does it make sense to

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:

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

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).

compute overlap interval

JA: @Lars, I think you mentioned once that you would like to calculate this. I could calculate using the method explained in the appendix.

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 (1991) 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 (1991). 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 (1995) and Blundell and Bond (1998) 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 (1995) and Blundell and Bond (1998):

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

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.

5 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. The probability for the first year is higher because of censoring and lack of previous information.

BD: That is some strange wording

JA: I got this definition from the U.S. SynLBD paper.

BD: Are we worried about this?

JA: Somebody asked

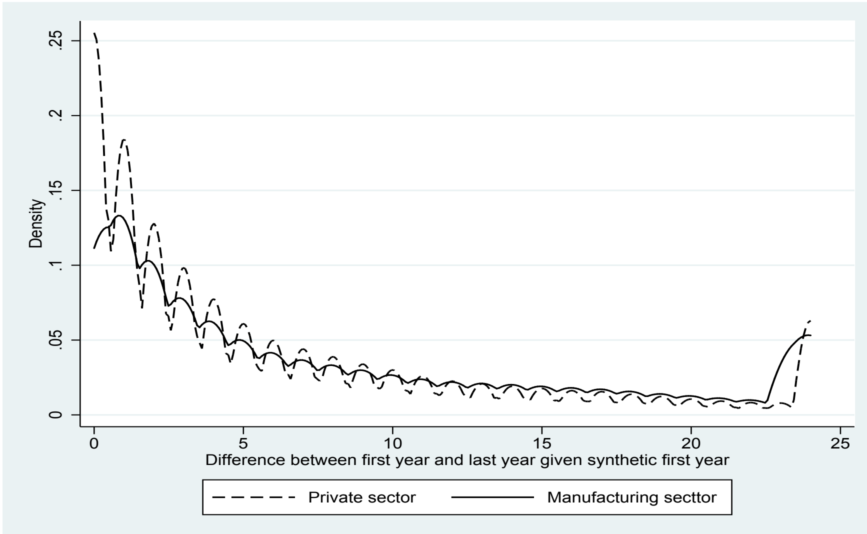
Table 9: Observed firm births given synthetic births (private)

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

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



6 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.

Supplementary Graphs

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 Karr et al. 2006. 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. Results are pending.

BD: ?

References

Abowd, John M., and Julia I. Lane. 2004. "New Approaches to Confidentiality Protection Synthetic Data, Remote Access and Research Data Centers". In *Privacy in Statistical Databases*, ed. by Josep Domingo-Ferrer and Vicenc Torra, 3050:282–289. Lecture Notes in Computer Science. Springer. ISBN: 978-3-540-22118-0. http://www.springer.com/la/book/9783540221180.

Abowd, John M., and Ian Schmutte. 2015. "Economic analysis and statistical disclosure limitation" [**inlang**English]. *Brookings Papers on Economic Activity* Fall 2015. ISSN: 00072303. http://www.brookings.edu/about/projects/bpea/papers/2015/economic-analysis-statistical-disclosure-limitation.

Abowd, John M., and Lars Vilhuber. 2010. "VirtualRDC - Synthetic Data Server". Cornell University, Labor Dynamics Institute. http://www.vrdc.cornell.edu/sds/.

Abowd, John M., et al. 2009. "The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators". In *Producer Dynamics: New Evidence from Micro Data*, ed. by Timothy Dunne, J. Bradford Jensen, and Mark J. Roberts. University of Chicago Press. ISBN: 978-0-226-17256-9. http://www.nber.org/chapters/c0485.

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

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

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

Bender, Stefan. 2009. "The RDC of the Federal Employment Agency as a part of the German RDC Movement". In *Comparative Analysis of Enterprise Data, 2009 Conference.* Visited on 05/05/2014. http://gcoe.ier.hit-u.ac.jp/CAED/index.html.

Benedetto, Gary, et al. 2007. "Using Worker Flows in the Analysis of the Firm". 25, no. 3 (): 299–313.

Blundell, Richard, and Stephen Bond. 1998. "Initial conditions and moment restrictions in dynamic panel data models". *Journal of Econometrics* 87, no. 1 (): 115–143. https://ideas.repec.org/a/eee/econom/v87y1998i1p115-143.html.

Bundesagentur für Arbeit. 2013a. *Establishment History Panel (BHP).* [Computer file]. Nürnberg, Germany: Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) [distributor].

— . 2013b. *German Job Vacancy Survey of the IAB.* [Computer file]. Nürnberg, Germany: Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) [distributor].

— . 2013c. *IAB Establishment Panel.* [Computer file]. Nürnberg, Germany: Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) [distributor].

Council, National Research. 2007. *Understanding Business Dynamics: An Integrated Data System for America's Future.* Ed. by John Haltiwanger, Lisa M. Lynch, and Christopher Mackie. Washington, DC: The National Academies Press. ISBN: 978-0-309-10492-0. doi:10.17226/11844. https://www.nap.edu/catalog/11844/understanding-business-dynamics-an-integrated-data-system-for-americas-future.

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

Drechsler, J. 2011a. *Synthetic Datasets for Statistical Disclosure Control–Theory and Implementation.* New York: Springer. doi:10.1007/978-1-4614-0326-5.

Drechsler, Jörg. 2011b. "Multiple imputation in practice: a case study using a complex German establishment survey". *AStA Advances in Statistical Analysis* 95, no. 1 (): 1–26. http://ideas.repec.org/a/spr/alstar/v95y2011i1p1-26.html.

— . 2012. "New data dissemination approaches in old Europe – synthetic datasets for a German establishment survey". *Journal of Applied Statistics* 39, no. 2 (): 243–265. http://ideas.repec.org/a/taf/japsta/v39y2012i2p243-265.html.

— . 2011c. *Synthetische Scientific-Use-Files der Welle 2007 des IAB-Betriebspanels.* FDZ Methodenreport 201101_de. Institute for Employment Research, Nuremberg, Germany. http://ideas.repec.org/p/iab/iabfme/201101_de.html.

Drechsler, Jorg, and Lars Vilhuber. 2014a. *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. https://ideas.repec.org/p/cen/wpaper/14-13.html.

Drechsler, Jörg, and Lars Vilhuber. 2014b. "A First Step Towards A German SynLBD: Constructing A German Longitudinal Business Database". *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics* 30 (2). doi:10.3233/SJI-140812. http://iospress.metapress.com/content/X415V18331Q33150.

Guzman, Jorge, and Scott Stern. 2020. "Startup Cartography". Visited on 01/26/2020. https://www.startupcartography.com/.

— . 2016. *The State of American Entrepreneurship: New Estimates of the Quality and Quantity of Entrepreneurship for 32 US States, 1988-2014.* Working Paper, Working Paper Series 22095. National Bureau of Economic Research. doi:10.3386/w22095. http://www.nber.org/papers/w22095.

Hethey, Tanja, and Johannes F. Schmieder. 2010. *Using worker flows in the analysis of establishment turnover: Evidence from German administrative data*. FDZ Methodenreport 201006_en. Institute for Employment Research, Nuremberg, Germany. http://ideas.repec.org/p/iab/iabfme/201006_en.html.

Jarmin, Ron S., Thomas A. Louis, and Javier Miranda. 2014. "Expanding The Role Of Synthetic Data At The U.S. Census Bureau". *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics* 30 (2). doi:10.3233/SJI-140813. http://iospress.metapress.com/content/fl8434n4v38m4347/?p=00c99b98bf2f4701ae806ee638594915&pi=0.

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

Karr, A. F., et al. 2006. "A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality". *The American Statistician* 60 (3): 1–9. doi:10.1198/000313006X124640.

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

— . 2014b. "Improving The Synthetic Longitudinal Business Database". *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics* 30 (2). doi:10.3233/SJI-140808.

Kinney, Satkartar K., et al. 2011a. *LBD Synthesis Procedures*. CES Technical Notes Series 11-01. Center for Economic Studies, U.S. Census Bureau. https://ideas.repec.org/p/cen/tnotes/11-01.html.

— . 2011b. "Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database". *International Statistical Review* 79, no. 3 (): 362–384. doi:j.1751-5823.2011.00152.x. https://ideas.repec.org/a/bla/istatr/v79y2011i3p362-384.html.

Little, Roderick J.A. 1993. "Statistical Analysis of Masked Data". *Journal of Official Statistics* 9 (2): 407–426.

Nowok, Beata, Gillian Raab, and Chris Dibben. 2016. "synthpop: Bespoke Creation of Synthetic Data in R". *Journal of Statistical Software, Articles* 74 (11): 1–26. ISSN: 1548-7660. doi:10.18637/jss.v074.i11. https://www.jstatsoft.org/v074/i11.

Raab, Gillian M, Beata Nowok, and Chris Dibben. 2018. "Practical Data Synthesis for Large Samples". *Journal of Privacy and Confidentiality* 7, no. 3 (): 67–97. doi:10.29012/jpc.v7i3.407. https://journalprivacyconfidentiality.org/index.php/jpc/article/view/407.

Rubin, Donald B. 1993. "Discussion of Statistical Disclosure Limitation". *Journal of Official Statistics* 9 (2): 461–468.

Sedláček, Petr, and Vincent Sterk. 2017. "The Growth Potential of Startups over the Business Cycle". *American Economic Review* 107, no. 10 (): 3182–3210. doi:10.1257/aer.20141280. http://www.aeaweb.org/articles?id=10.1257/aer.20141280.

Snoke, Joshua, and Aleksandra Slavkovic. 2018. "pMSE Mechanism: Differentially Private Synthetic Data with Maximal Distributional Similarity: UNESCO Chair in Data Privacy, International Conference, PSD 2018, Valencia, Spain, September 26-28, 2018, Proceedings", 138–159. ISBN: 978-3-319-99770-4. doi:10.1007/978-3-319-99771-1_10.

U.S. Census Bureau. 2017. "Business Dynamics Statistics (BDS)". U.S. Census Bureau. Visited on 01/26/2020. https://www.census.gov/programs-surveys/bds.html.

— . 2016a. "County Business Patterns (CBP)". U.S. Census Bureau. Visited on 01/26/2020. https://www.census.gov/programs-surveys/cbp.html.

— . 2015. *Longitudinal Business Database YYYY-YYZZ [Data file]*. Tech. rep. Visited on 01/26/2020. https://www.census.gov/programs-surveys/ces/data/restricted-use-data/longitudinal-business-database.html.

— . 2016b. "Statistics of U.S. Businesses (SUSB)". U.S. Census Bureau. Visited on 01/26/2020. https://www.census.gov/programs-surveys/susb.html.

Vilhuber, Lars. 2018. *LEHD Infrastructure S2014 files in the FSRDC*. Working Papers 18-27. Center for Economic Studies, U.S. Census Bureau. https://ideas.repec.org/p/cen/wpaper/18-27.html.

— . 2013. *Methods for Protecting the Confidentiality of Firm-Level Data: Issues and Solutions*. Document 19. Labor Dynamics Institute. http://digitalcommons.ilr.cornell.edu/ldi/19/.

Vilhuber, Lars, John M. Abowd, and Jerome P. Reiter. 2016. "Synthetic establishment microdata around the world". *Statistical Journal of the International Association for Official Statistics* 32 (1): 65–68. doi:10.3233/SJI-160964.

Woodcock, Simon D., and Gary Benedetto. 2009. "Distribution-preserving statistical disclosure limitation". *Computational Statistics & Data Analysis* 53 (12): 4228–4242. ISSN: 0167-9473. doi:https://doi.org/10.1016/j.csda.2009.05.020. http://www.sciencedirect.com/science/article/pii/S0167947309002011.