STATISTICS IN TRANSITION new series, August 2020
Vol. XX, No. X, pp. xxx–xxx

APPLYING DATA SYNTHESIS FOR LONGITUDINAL BUSINESS DATA ACROSS THREE COUNTRIES

M. Jahangir Alam[1], **Benoit Dostie**[2] **Jörg Drechsler**[3] **Lars Vilhuber**[4]

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 businesses 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 and protection, and provide an assessment of the feasibility of extending such an approach in a cost-effective way to other data.

Keywords: synthetic data, business data, confidentiality, LBD, LEAP, BHP, synthetic.

1. Introduction

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.

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

[1]Institutional affiliation. E-mail: some@one.edu
[2]Institutional affiliation. E-mail: author2@address.com
[3]Institute for Employment Research. E-mail: joerg.drechsler@iab.de
[4]Cornell University. E-mail: lars.vilhuber@cornell.edu

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,[5] 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 et al. 2008; Drechsler 2012; National Research 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 improved access to such datasets to researchers (Bender 2009; Vilhuber 2013; Abowd and Lane 2004; Abowd and Schmutte 2015). [6] For business data, synthetic business microdata were released in the United States (Kinney et al. 2011b) and in Germany (Drechsler 2011b) 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. Details about the modeling strategies used for the SynLBDcan be found in Kinney et al. (2011b) and Kinney et al. (2011a).

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, but applied to data from two different countries: Canada (Longitudinal Employment Analysis Program (LEAP)) and Germany (Establishment History Panel (BHP)). We describe all three countries' data in Section 2.

In Canada, the Canadian Center for Data Development and Economic Research

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

[6] 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.

(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 remains the cost of traveling to Ottawa.

The Institute for Employment Research (IAB) in Germany also strictly regulates the access to its business data. All business data can be accessed exclusively onsite at the research data center (RDC) and only after the research proposal has been approved by the Federal Ministry of Labour and Social Affairs. All output is carefully checked by staff at the RDC and only cleared output can be removed from the RDC.

The experiment described in this paper aims not so much at 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 predefined 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 method, and any adjustments we made to take into account structural data differences, are described in Section 3.

We verify the analytical validity of the synthetic data files so created along a variety of measures. First, we show how well average firm characteristics (gross employment, total payroll) in the synthetic data match those from the original data. We also consider how well the synthetic data replicates various measures of firm dynamics (entry and exit rates) and job flows (job creation and destruction rate). Second, we assess whether measures of economic growth vary between both datasets using dynamic panel data models. Finally, to assess the analytical validity from a more general perspective, we compute global validity measures based on the ideas of propensity score matching as proposed by Woo et al. (2009) and Snoke et al. (2018).

To assess how protective the newly created synthetic database is, we estimate the probability that the synthetic first year equals the true first year given the synthetic fist year.

The rest of the paper is organized as follows. Section 2 describes the different data sources and summarizes which steps were taken to harmonize the datasets prior to the actual synthesis. Section 3 provides some background on the synthesis methods, limitations in the applications, and a discussion of some of the measures, which are used in Section 4 to measure the analytical validity of the generated datasets. Preliminary results regarding the achieved level of protection are included in Section 5. The paper concludes with a discussion of the implications of the study for future data synthesis projects.

2. Data

In this section, we briefly describe the structure of the three data sources.

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 (Statistics Canada 2019b) 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 (Statistics Canada 2019c), and (3) data from Statistics Canada's Survey of Employment, Payrolls and Hours (SEPH) (Statistics Canada 2019a). 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. The Business Register is Statistics Canada's central repository of baseline information on businesses and institutions operating in Canada. It is used as the survey frame for all business related data sets. 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) an industry classification, (3) payroll and (4) a measure of employment. 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. The industry classification 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 currently based on the North American Industrial Classification System (NAICS). We use the information at the NAICS four-digit (industry group) level. The firm's payroll is measured as the sum of all T4s reported to the CRA for the calendar year. Employment is measured either using Individual Labour Unit (ILU) or Average Labour Unit (ALU). ALUs are obtained by dividing the payroll 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.

For the purpose of this experiment, we exclude the public sector (NAICS 61, 62, and 91), even though they are contained in the database, because they may not be accurately captured (Statistics Canada 2019b). Statistics Canada does not publish any statistics for those sectors.

2.3. Germany: Establishment History Panel (BHP)

The core database for the Establishment History Panel 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. Aggregating this information via an establishment identifier yields the Establishment History Panel (Bundesagentur für Arbeit 2013, 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. Industries are identified according to the WZ 2003 classification system (Statistisches Bundesamt 2003) at the five digit level.[7] We aggregated the industry information for this project using the first four digits of the coding system.

2.4. Comparability and Pre-processing

2.5. Harmonizing and Preprocessing

In all countries, the underlying data provides annual measures. However, SYNLBD assumes a longitudinal (wide) structure of the dataset, with invariant industry (and location). In all cases, the modal industry is chosen to represent the entity's industrial activity. Further adjustments made to the BHP for this project include estimating full-year payroll, creating time-consistent geographic information, and applying employment flow methods (Hethey and Schmieder 2010) to adjust for spurious births and deaths in establishment identifiers. Drechsler and Vilhuber (2014b) provide a detailed description of the steps taken to harmonize the input data.

In both Canada and Germany, we encountered various technical and data-driven limitations. In all countries, data in the first year and last year are occasionally problematic, and such data were dropped. Both the German and the Canadian data experience some level of industry coding change, which may affect the classification of some entities. Furthermore, due to the nature of the underlying data, entities are establishments in Germany and the US, but employers in Canada.

After the various standardizations and choices made above, the data structure is intended to be comparable, as summarized in Table 1. The column "Nature" identifies the treatment of the variable in the synthesis process SYNLBD.

[7] The WZ 2003 classification system is compliant with the requirements of the Statistical Classification of Economic Activities in the European Community (NACE Rev. 1.1), which is based on the International Standard Industrial Classification (ISIC Rev. 3.1).

Table 1: Variable descriptions and comparison

Name	Type	Description	US	Canada	Germany	Nature
Entity Identifier	identifier		Establishment	Employer	Establishment	Created
Industry code	Categorical	Various across countries	SIC3 (3-digit)	NAICS4 (4-digit)	WZ2003 (4-digit)	Unmodified
First year	Categorical	First year entity is observed		— firstyear —		Synthesized
Last year	Categorical	Last year entity is observed		— lastyear —		Synthesized
Year	Categorical	Year dating of annual variables		— year —		Derived
Employment	Continuous	Employment measure	Count (March 15)	ALU* (annual)	Count (June 30)	Synthesized
Payroll	Continuous	Payroll (annual)	Reported	Computed	Computed, Adjusted	Synthesized

* ALU = Average Labour Unit. See text for additional explanations.

3. Methodology

To create a partially synthetic database with analytic validity from longitudinal establishment data, Kinney et al. (2011a) synthesize the life-span of establishments, as well as the evolution of their employment, conditional on industry over that synthetic lifespan. Geography is not synthesized, but is suppressed from the released file (Kinney et al. 2011a). Applying this to the LBD, Kinney et al. (2011b) created the current version of the Synthetic LBD, based on the Standard Industrial Classification (SIC) and extending through 2000. Kinney, Reiter, and Miranda (2014a) describe efforts to create a new version of the Synthetic LBD, using a longer time series (through 2010) and newer industry coding (NAICS), while also adjusting and extending the models for improved analytic validity and the imputation of additional variables. In this paper, we refer to and re-use the older methodology, which we will call SynLBD. Our emphasis is on the comparability of results obtained for a given methodology across the various applications.

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. The synthetic data are generated by sampling new values from this distribution. In SynLBD, variables are synthesized in a sequential fashion, with categorical variables being generally processed first using a variant of Dirichlet-Multinomial models. Continuous variables are then synthesized using a normal linear regression model with kernel density-based transformation (Woodcock and Benedetto 2009).[8] The synthesis models are run independently for each industry. SynLBD is implemented in SAS™, which is frequently used in national statistical offices.

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 SynLBD to the German Longitudinal Business Database

[8]Kinney, Reiter, and Miranda (2014a) shift to a Classification and Regression Trees (CART) model with Bayesian bootstrap.

(GLBD). In this paper, we extend the analysis from the earlier paper, and extend the application to the Canadian context (SynLEAP).

3.1. Limitations

In all countries, the synthesis of certain industries failed to complete. In both Canada and the US, this number is less than 10. In Canada, they account for about 7 percent of the total number of observations (see Table 13 in the Appendix).

In the German case, our experiments were limited to only a handful of industries, due to a combination of time and software availability factors. The results should still be considered preliminary. In both countries, as outlined in Section 2, there are subtle but potentially important differences in the various variable definitions. Industry coding differs across all three countries, and the level of detail in each of the industry codings may affect the success and precision of the synthesis.[9]

As noted in Section 2, entities are establishments in Germany and the US, but employers in Canada. SynLBD should work on any level of entity aggregation (see Kinney, Reiter, and Miranda (2014a) for an application to hierarchical firm data with both firm/employer and establishment level imputation). However, these differences may affect the observed density of the data within industry-year categories, and therefore the overall comparability.

Finally, due to a feature of SynLBD that we did not fully explore, synthesis of data in the last year of the data generally was of poor quality. For some industry-country pairs, this also happened in the first year. We dropped those observations.

3.2. Measuring outcomes

In order to assess the outcomes of the experiment, we inspect analytical validity by various measures and also evaluate the extent of confidentiality protection. To check analytical validity, we compare basic univariate time series between the synthetic and confidential data (employment, entity entry and exit rates, job creation and destruction rates), and the distribution of entities (firms and establishment, depending on country), employment, and payroll across time by industry. For a more complex assessment, we compute a dynamic panel data model of economic (employment) growth on each dataset. We computed, but do not report here the confidence interval overlap measure (CIO) proposed by Karr et al. (2006) and Woo et al. (2009) in all these evaluations.[10] The CIO is a popular measure when evaluating the validity for specific analyses. It evaluates how much the confidence intervals of the original data and the protected data overlap. We did not find this measure to be useful in our context. Most of our analyses are based on millions of records,

[9] Statistics Canada and Bureau of the Census (1991), when comparing the 1987 US Standard Industrial Classification (SIC) to the 1980 Canadian SIC, already pointed out that the degree of specialization, the organization of production, and the size of the respective markets differed. Thus, the density of establishments within each of the chosen categories is likely to affect the quality of the synthesis.

[10] The full parameter estimates and the computed CIO are available in our online repository at DOI TBD.

and observed confidence intervals were so small that confidence intervals (almost) never overlap even when the estimates between the original data and the synthetic data are quite close.

To provide a more comprehensive measure of quality of the synthetic data relative to the confidential data, we compute the $pMSE$ (propensity score mean-squared error, Woo et al. 2009; Snoke and Slavkovic 2018; Snoke et al. 2018): 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 Woo et al. (2009) and Snoke and Slavkovic (2018) to calculate the $pMSE$, using the following algorithm:

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, where both X and X^s are in the long format.

2. Create a variable I_{et} denoting membership of an observation for entity e, $e = 1,\ldots,E$, at time point t, $t = 1,\ldots,T$, in the component databases, $I_{et} = \{1 : X^{comb}_{et} \in X^s\}$. I_{et} takes on values of 1 for the synthetic database and 0 for the confidential database.

3. Fit the following generalised linear model to predict I

$$P(I_{et} = 1) \quad = \quad g^{-1}(\beta_0 + \beta_1 Emp_{et} + \beta_2 Pay_{et} + Age^T_{et}\beta_3 + \lambda_t + \gamma_i), \qquad (1)$$

   where Emp_{et} is log employment of entity e in year t, Pay_{et} is log payroll of entity e in year t, Age_{et} is a vector of age classes of entity e in year t, λ_t is a year fixed effect, γ_i is an time-invariant industry-specific effect, and g is an appropriate link function (in this case, the logit link).

4. Calculate the predicted probabilities, $\hat{p}_{et}$.

5. Compute $pMSE = \frac{1}{N}\sum_{t=1}^{T}\sum_{e=1}^{E}(\hat{p}_{et} - c)^2$, where $c = n_2/N$.

If $n_1 = n_2$, $pMSE$ = 0 means every $\hat{p}_i = 0.5$, and the two databases are distributionally indistinguishable, suggesting high analytical validity. While the number of records in the protected data typically matches the number of records in the original data, i.e., $n_1 = n_2$, this does not necessarily hold in our application. While the synthesis process ensures that the total number of entities is the same in both datasets, the years in which the entities are observed will differ between the original data and the synthetic data and thus the number of records in the long format will not necessarily match between the two datasets. For this reason we follow Woo et al. (2009) and Snoke et al. (2018) and use $c = n_2/N$. Using this more general definition, c will always be the mean of the predicted propensity scores so that the $pMSE$ measures the sum of the squared deviations from the mean, as intended.

Since the $pMSE$ depends on the number of predictors included in the propensity score model, Snoke et al. (2018) derived the expected value and standard deviation

for the $pMSE$ under the null hypothesis ($pMSE_0$) that the synthesis model is correct, i.e, it matches the true data generating process (Snoke et al. 2018, Equation 1):

$$E[pMSE_0] = (k-1)(1-c)^2 \frac{c}{N}$$

and

$$StDev[pMSE_0] = \sqrt{2(k-1)}(1-c)^2 \frac{c}{N}$$

where k is the number of synthesized variables used in the propensity model. To measure the analytical validity of the synthetic data, they suggest to look at the *pMSE ratio*

$$pMSEratio = \frac{\widehat{pMSE}}{E[pMSE_0]}$$

and the *standardized pMSE*

$$pMSE_s = \frac{\widehat{pMSE} - E[pMSE_0]}{StDev[pMSE_0]},$$

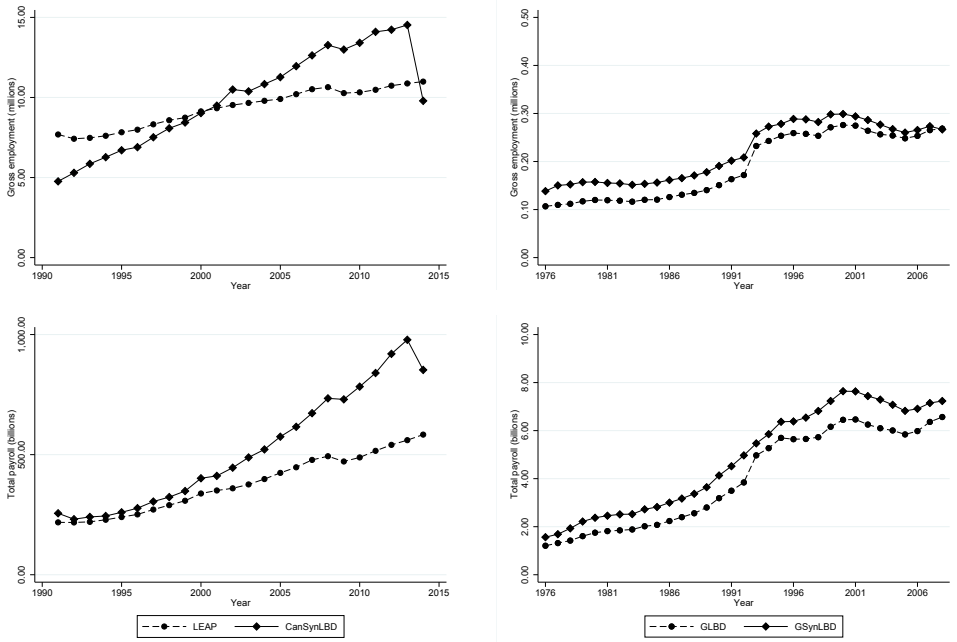
where $\widehat{pMSE}$ is the estimated pMSE based on the data at hand. Under the null hypothesis, the $pMSE$ ratio has an expectation of 1 and the expectation of the standardized $pMSE_s$ is zero.

4. Analytical validity

In the following figures, the data for the Canadian data are shown in the left panels, and the German data in the right panels. In all cases, the Canadian data are reported for the entire private sector, including the manufacturing sector but excluding the public sector industries (NAICS 61, 62, and 91). German results are for two WZ2003 industries.

4.1. Entity Characteristics

Figure 1 shows a comparison between the synthetic data and the original data for gross employment level (upper panels) and total payroll (lower panels) by year. While the general trends are preserved for both data sources, the results for the German synthetic data resemble the trends from the original data more closely. For the Canadian data the positive trends over time are generally overestimated. However, in both cases, levels are mostly overestimated. These patterns are not robust. When considering the Canadian manufacturing sector in Canada (Figure 8 in the Appendix), trends are better matched, but a significant *negative* bias is present in levels.



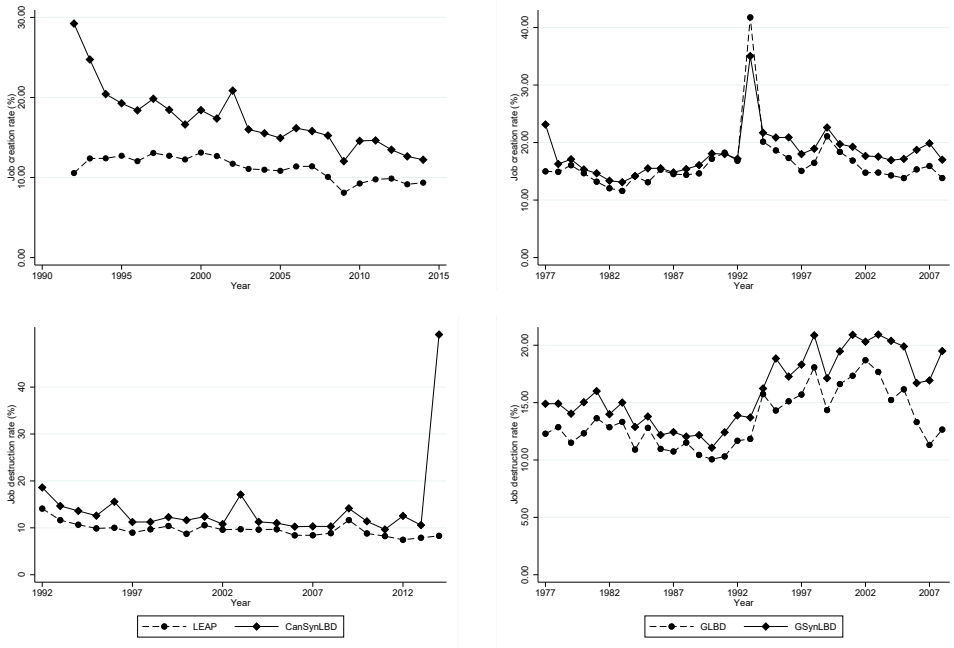
(a) CanSynLBD (b) GSynLBD

Figure 1: Gross employment level (upper panels) and total payroll (lower panels) by year.

4.2. Dynamics of Job Flows

Key statistics commonly computed from business registers such as the LEAP or the BHP include job flows over time. Following Davis, Haltiwanger, and Schuh (1996), 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. Figure 2 depicts job creation rates (upper panels) and destruction rates (lower panels). The general levels and trends are preserved for both data sources, but the time-series align more closely for the German data. Even the substantial increase in job creations in 1993, which can be attributed to the integration of the data from Eastern Germany after reunification, is remarkably well preserved in the synthetic data. Still, there seems to be a small but systematic overestimation of job creation and destruction rates in both synthetic data sources. The substantial deviation in the job destruction rate in the last year of CanSynLBD is an artefact requiring further investigation.[11]

[11] The results for the Canadian manufacturing sector are included in Figure 9 in the Appendix, and are comparable to the results for the entire private sector.



(a) CanSynLBD (b) GSynLBD

Figure 2: Job creation rates (upper panels) and job destruction rates (lower panels) by year.

4.3. Entity Dynamics

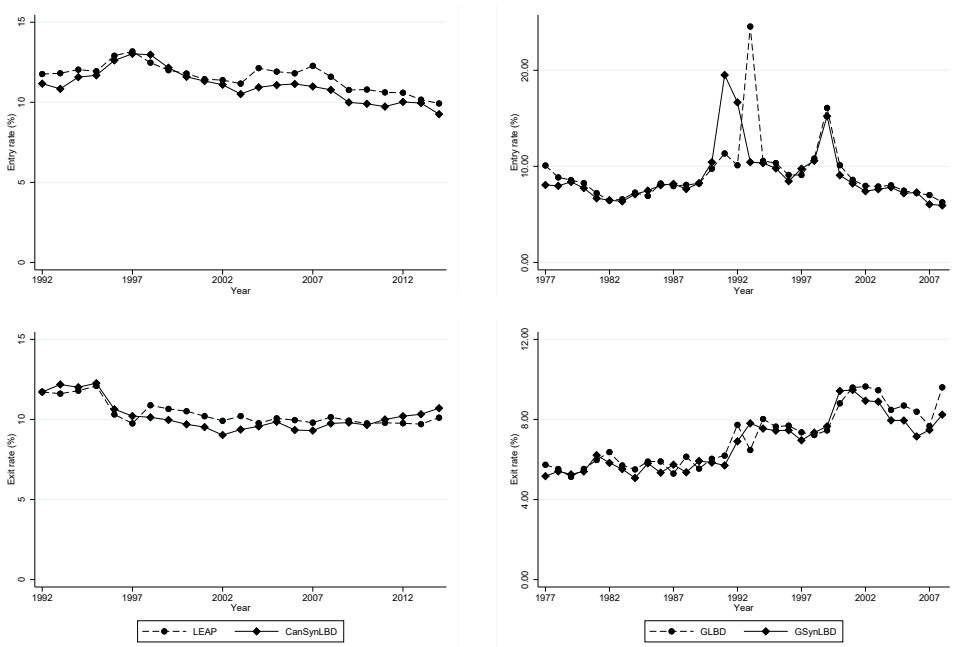
To assess how well the synthetic data capture entity dynamics, we also compute entry and exit rates, i.e. how many new entities appear in the data and how many cease to exist relative to the population of entities in a specific year.[12] Figure 3 shows that those rates are very well preserved for both data sources.

Only the (delayed) re-unification spike in the entry rates in the German data is not preserved correctly. The confidential data show a large spike in entry rates in 1993. In that year, detailed information about Eastern German establishments was integrated for the first time. However, the synthetic data shows increased entry rates in the two previous years. We speculate that this occurs due to incomplete data in the confidential data: Establishments were successively integrated into the data starting in 1991, but many East German establishments did not report payroll and number of employees in the first two years. Thus, records existed in the original data, but the establishment size is reported as missing. Such a combination is not possible in the synthetic data. The synthesis models are constructed to ensure that whenever an establishment exists, it has to have a positive number of

[12]As described in Section 2, for both countries' data, corrections based on worker flows have been applied, correcting for any bias due to legal reconfiguration of economic entities.

employees. Since entry rates are computed by looking at whether the employment information changed from missing to a positive value, most of the Eastern German establishments only exist from 1993 on-wards in the original data, but from 1991 in the synthetic data.

The second, smaller spike in the entry rate in the German data occurs in 1999. In that year, employers were required to report marginally employed workers for the first time. Some establishments exclusively employ marginally employed workers, and will thus appear for the first time in the data after 1999. The synthetic data preserves this pattern.



(a) CanSynLBD (b) GSynLBD

Figure 3: Entry rates (upper panels) and exit rates (lower panels) by year.

4.4. Distribution of variables across time and industry

The SYNLBD code ensures that the total number of entities that ever exist within the considered time frame match exactly between the original data and the synthetic data. But each entity's entry and exit date are synthesized, and the total number of entities at any particular point in time may differ, and with it employment and payroll. To investigate how well the information is preserved at any given point in time, we compute the following statistic:

$$x_{its} = X_{its} / \sum_i \sum_t X_{its}, \tag{2}$$

where i is the index for the industry (aggregated to the two digit level for the Canadian data), t is the index for the year and s denotes the data source (original or synthetic). $X_{its} = \sum_j X_{itsj},\ j = 1, \ldots, n_{its}$ is the variable of interest aggregated at the industry level and n_{its} is the number of entities in industry i at time point t in data source s. To compute the statistic provided in Equation (2), this number is then divided by the total payroll aggregated across all industries and years. Figure 4 plots the results from the original data against the results from the synthetic data for the number of entities, employment, and payroll. If the information is well preserved, all points should be close to the 45 degree line.

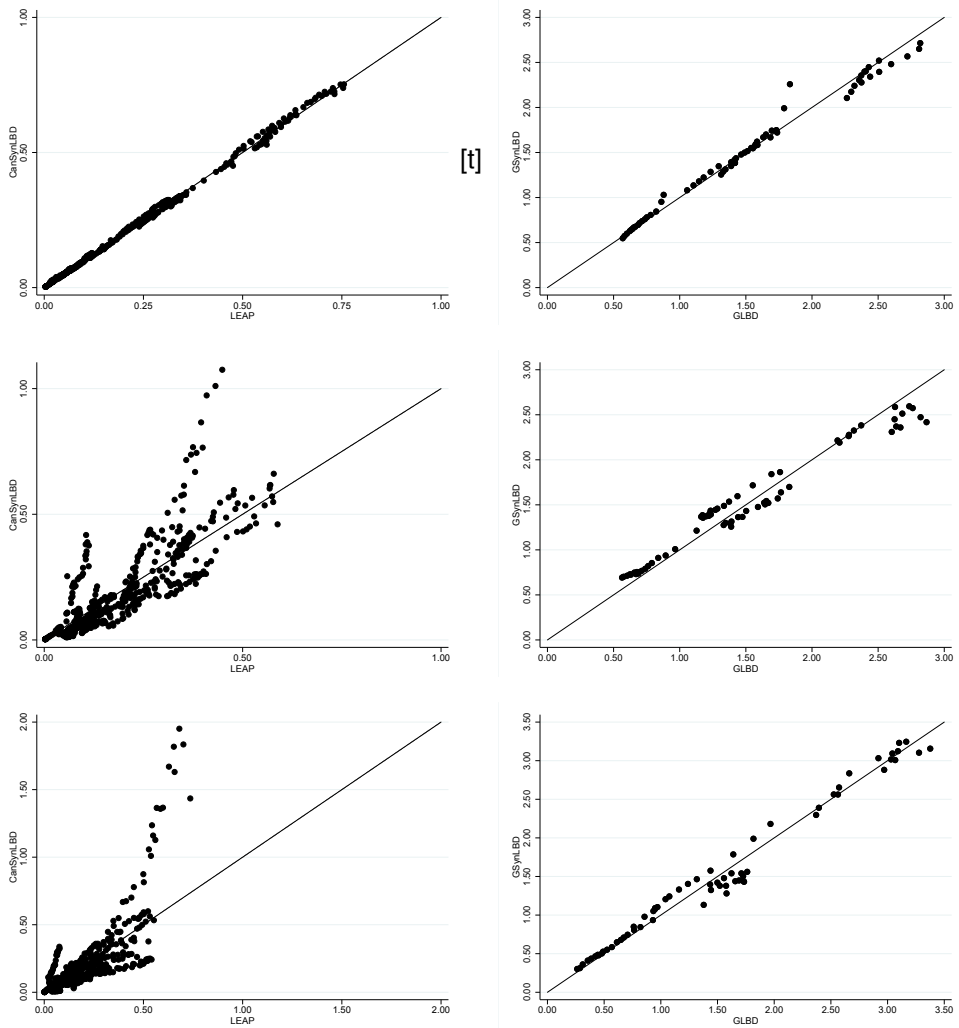
We find that the share of entities is well preserved for both data sources, but share of employment and share of payroll vary more in the Canadian data with an upward bias for the larger shares. It should be noted that the German data shown here and elsewhere in this paper only contain data from two industries, whereas the Canadian data contains nearly all available industry codes at the two digit level. Thus, results from Canada are expected to be more diverse. When only considering the Canadian manufacturing sector (see Figure 10 in the Online Appendix), less bias is present.

4.5. Modelling strategy

To assess how well the synthetic data perform in a more complex model and in the context of an analyst's modelling strategy, we simulate how a macroeconomist (the typical user of these data) might approach the problem of estimating a model for the evolution of employment if only the synthetic data are available. The analyst will consider both the literature and the data to propose a meaningful model. In doing so, a sequence of models will be proposed, and tests or theory brought to bear on their merits, potentially rejecting their appropriateness. In doing so, the outcome that the analyst obtains from following that strategy using the synthetic data should not diverge substantially from the outcome they would obtain when using the (inaccessible) confidential data. The specific parameter estimates obtained, and the actual model retained, are not the goal of this exercise — the focus is on the process.

To do so, our analyst would start by using a base model (typically OLS), and then let economic and statistical theory suggest more appropriate models. In this case, we will estimate variants of a dynamic panel data model for the evolution of employment. For each model, tests can be specified to test whether the model is an appropriate fit under a certain hypothesis.[13] The outcome of this exercise, illustrated by Figure 5, allow us to assess whether the synthetic data capture variability in economic growth due to industry, firm age and payroll — the key variables in the data — and whether the analyst might reasonable choose the same, or a closely related modelling strategy.

[13]We do not describe these models in more detail here, referring the reader to the literature instead, in particular Arellano and Bover (1995) and Blundell and Bond (1998).



[t]

(a) CanSynLBD (b) GSynLBD

Figure 4: Share of entities (upper panels), share of employment (middle panels), and share of payroll (lower panels) by year and industry.

The base model is an OLS specification:

$$Emp_{et} = \beta_0 + \theta Emp_{e,t-1} + \eta Pay_{et} + Age_{et}^T \beta + \gamma_i + \lambda_t + \varepsilon_{et} \quad (3)$$

where Emp_{et} is log employment of entity e in year t, $Emp_{e,t-1}$ is its one year lag, Pay_{et} is the logarithm of payroll of entity e in year t, Age_{et} is a vector of dummy variables for age of entity e in year t, λ_t is a year effect, γ_i is a time-invariant industry-specific effect for each industry i, and ε_{et} is the disturbance term of entity e in year t. As

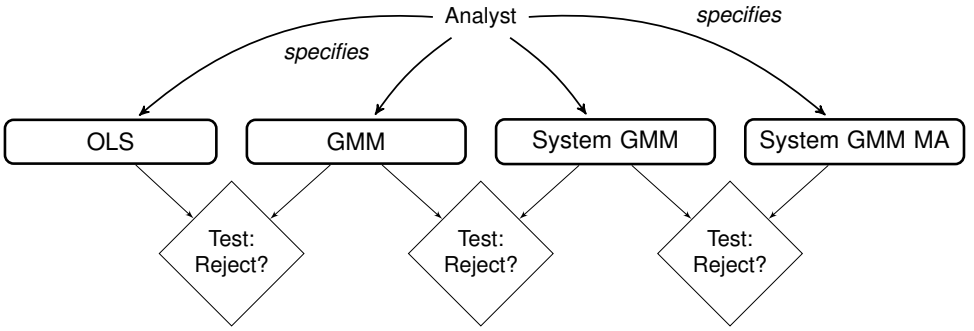


Figure 5: Modelling strategy of a hypothetical analyst

$Emp_{e,t-1}$ is correlated with γ_i because $Emp_{e,t-1}$ is itself determined by time-invariant γ_i, OLS estimators are biased and inconsistent. To obtain consistent estimates of the parameters in the model, Arellano and Bond (1991) suggest using generalized method of moments (GMM) estimation methods, as well as associated tests to assess the validity of the model. We also estimate the model using system GMM methods proposed by Arellano and Bover (1995) and Blundell and Bond (1998) (System GMM), as well as a variant of equation (3) that includes a first-order moving average in the error term ε_{et} (System GMM MA):

$$Emp_{et} = \beta_0 + \theta Emp_{e,t-1} + \eta Pay_{et} + Age_{et}^T\beta + \lambda_t + \alpha_e + \varepsilon_{et} + \varepsilon_{e,t-1} \quad (4)$$

where α_e is a time-invariant entity effect, which includes any time-invariant industry effects.

The Sargan test (Hansen 1982; Arellano and Bond 1991; Blundell, Bond, and Windmeijer 2001) is used to assess the validity of the over-identifying restrictions. We also compute the z-score for the $m2$ test for zero autocorrelation in the first-differenced errors of order two (Arellano and Bond 1991).

An interesting derived effect is to consider the long-run effect of (log) payroll on (log) employment, or the elasticity of employment with respect to payroll. This can be estimated as

$$\eta^\star = \frac{\hat{\eta}}{1-\hat{\theta}}.$$

It is important that this model is close, but not identical to the model used to *synthesize* the data. In SynLBD, Emp_{et} is synthesized as $f(Emp_{e,t-1}, X_{et})$ (where X_{et} does not contain Pay_{et}), and $Pay_{et} = f(Pay_{e,t-1}, Emp_{et}, X_{et})$ (Kinney et al. 2011b, pg. 366). Thus, the model we chose is purposefully not (completely) congenial with the synthesis model, but the synthesis process SynLBD should preserve sufficient serial correlation in the data to be able to estimate these models.

We estimate each model and test statistics separately on confidential and synthetic data for the private sector (and for Canada, for the manufacturing sector). Detailed estimation results are reported in the Online Appendix. Here we focus

on the two regression coefficients of major interest: θ and η, the coefficients for lagged employment and payroll, as well as the elasticity $\eta^{\star}$. Figure 6 plots the bias in the synthetic coefficients, i.e., $\theta_{synth} - \theta_{conf}$ and $\eta_{synth} - \eta_{conf}$, for all four models. While the detailed results in the Online Appendix confirm that all regression coefficients still have the same sign, all estimates plotted in Figure 6 show substantial bias in all models in all datasets (the OLS model for the German data being the only exception). However, the computed elasticity $\eta^{\star}$ has very little bias in most models.

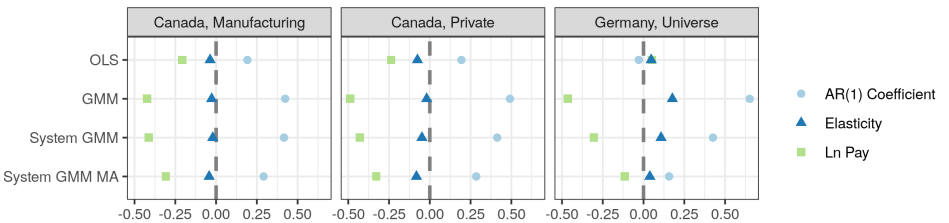


Figure 6: Bias in estimates of coefficients on pay and lagged employment

Note: For details on the estimated coefficients, see the Online Appendix.

However, we observe a striking pattern: The biases of the two regression coefficients are always symmetric, i.e. the sum of the biases of θ_{synth} and η_{synth} is close to zero in all models (and cancel mostly out in the computation of $\eta^{\star}$). This may simply be a feature of the modeling strategy pointed out earlier, which generates serial correlation with a slightly different structure. Another possible explanation could be that the model is poorly identified because of multicollinearity generating a ridge for the estimated coefficients. The estimated coefficients would be highly unstable in this case even in the original data and thus it would not be surprising to find substantial differences between the coefficients from the original data and the coefficients from the synthetic data. Better understanding this phenomenon will be an interesting area of future research.

Table 2: m2 and Sargan tests by country

		Canada		Germany	
Model	Test	Confidential	Synthetic	Confidential	Synthetic
GMM	m2	-14.5	-27.54	-2.51	-4.13
	Sargan test	69000	15000	3600	2000
System GMM	m2	-11.43	-41.6	19.49	-8.83
	Sargan test	77000	18000	4500	2800
System GMM MA	m2	8.2	-40.03	19.03	-11.69
	Sargan test	28000	17000	3100	2500

Note: The Sargan test (Blundell, Bond, and Windmeijer 2001; Arellano and Bond 1991) is used to assess the validity of the over-identifying restrictions. The z-score for the $m2$ tests for zero autocorrelation in the first-differenced errors of order two (Arellano and Bond 1991). See text for additional information.

Whereas the bias in coefficients is quite consistent across countries and models, specification tests such as the $m2$ test for autocorrelation and the Sargan test paint a slightly less consistent picture. Table 2 shows the two tests for each of the models estimated by country, synthetic status, and model. The Sargan test rejects the null in both countries and for all models, consistently for confidential and synthetic data. But the $m2$ test is of opposite signs for half of the comparisons.

4.6. pMSE

To compute the $pMSE$, we estimate Equation (1) using logit models. The estimated $pMSE$ is 0.0121 for the Canadian data (0.0041 for the manufacturing sector) and 0.0013 for the German data (see Table 3). While these numbers may seem small, the $pMSE$ ratio and the standardized $pMSE$ are large, indicating that the null hypothesis that the synthetic data and the original data stem from the same data generating process should be rejected. The expected $pMSE$ is quite sensitive to sample size N. Even small differences between the original and synthetic data will lead to large values for this test statistic. In both countries, the confidential data files are quite large (about 2 million cases for Germany and the manufacturing sector in Canada and about 34.5 million cases for the full Canadian dataset). In practice, therefore, it is quite likely to reject the null of equivalence given this test's very high power.

Table 3: pMSE by sector and country

Country	Sector	pMSE	pMSE ratio	standardized pMSE
Canada	Manufacturing	0.0041	656.88	4908.17
Canada	Private	0.0121	10957.61	135525.77
Germany	Universe	0.0013	725.21	2896.85

5. Confidentiality protection

To assess the risk of disclosure, we use a measure proposed by Kinney et al. (2011b): For each industry, we estimate the fraction of entities by industry for which the synthetic birth year equals the true birth year, conditional on the synthetic birth year, and interpret it as a probability. Tables 14 and 15 in the Online Appendix show the minimum, maximum, and mean of these probabilities, by year. Figure 7 shows the maximum and average values across time, for each country.[14] The figure shows that these probabilities are quite low except for the first year. As Figure 3 showed, entry rates in the first year are much larger than in any other year due to censoring. It is therefore quite likely that the (left-censored) entry year of the synthetic record

[14]The Canadian manufacturing sector is not shown. In the German case, we only use two industries, but we show the average of the two, rather than the values for both industries, to maintain comparability with the Canadian plot.

matches that of the (left-censored) original record if the synthetic entry year is the first year observed in the data. A somewhat more muted version of this effect can be seen for Germany in the years 1991 and 1992, when the lower panel of Figure 7 shows another spike. These are the years in which data from Eastern Germany were added to the database successively, leading to new sets of (left-censored) entities.

With the exception of the first year in the data, the average rate of concordance between synthetic and observed birth year of an establishment in the Canadian data is below 5%, and the maximum is never above 50%. The German data reflect results from a smaller set of industries, and while the average concordance is higher (never above 10%), the maximum is never above 6% other than during the noted entry spikes. This suggests that the synthetic lifespan of any given entity is highly unlikely to be matched to its confidential real lifespan. This is generally considered to be a high degree of confidentiality.

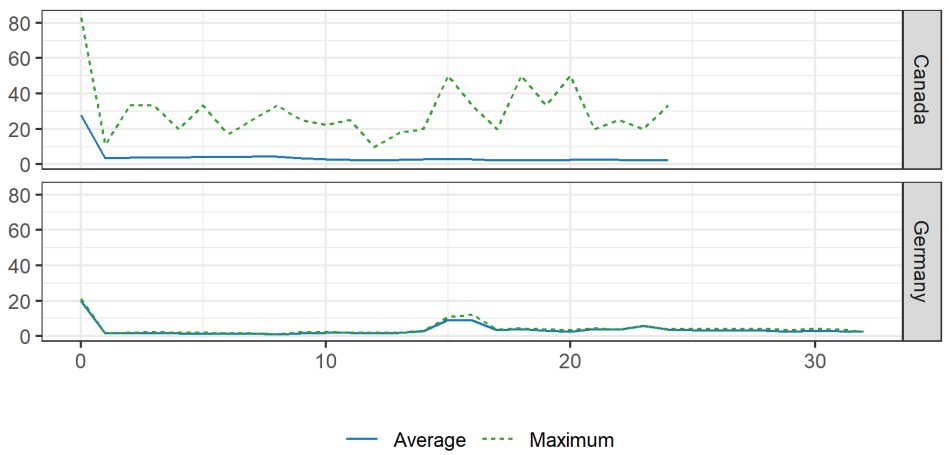


Figure 7: Average and maximum likelihood that synthetic birthyear matches actual birthyear

Note: Plot shows fraction of entities by industry for which the synthetic birth year equals the true birth year, conditional on the synthetic birth year, and interpret it as a probability. Plot has been rescaled to be relative to the first year observed in the data.

6. Conclusion

In this paper, we presented results from two projects that evaluated whether code developed to synthesize the U.S. LBD can easily be adapted to create synthetic versions of similar data from Canada and Germany. We considered both univariate time-series comparisons as well as model-based comparisons of coefficients and

model fit. In general, utility evaluations show significant differences between each country's synthetic and confidential data. Frequently-used measures such as confidence interval overlap and $pMSE$ suggest that the synthetic data are an unreliable image of the confidential data. Less formal comparisons of specification test scores suggest that the synthetic data do not reliably lead to the same modeling decisions.

Interestingly, the utility of the German synthetic data was higher than the utility of the Canadian data in almost all dimensions evaluated. At this point we can only speculate about potential reasons. The most important difference between the two data sources is that the German data comprises only a handful of industries while almost all industries have been included in the Canadian evaluation. Given that the industries included in the German data were rather large, and synthesis models are run independently for each industry, it might have been easier to preserve the industry level statistics for the German data. We cannot exclude the possibility that the structure of the German data aligns more closely with the LBD and thus the synthesis models tuned on the LBD data provide better results on the (adjusted) BHP than on the LEAP. We note that both the LBD and the BHP are establishment-level datasets, whereas the LEAP is a employer-level dataset.

We emphasize that adjustments to the original synthesis code were explicitly limited to ensuring that the code runs on the new input data. The validity of the synthetic data could possibly be improved by tuning the synthesis models to the particularities of the data at hand, such as the non-standard dynamics introduced into the German data by reunification. However, the aim of this project was to illustrate that the high investments necessary for developing the synthesis code for the LBD offered additional payoffs as the re-use of the code substantially reduced the amount of work required to generate decent synthetic data products for other business data. One of the major criticisms of the synthetic data approach has been that investments necessary to develop useful synthesizers are substantial. This project illustrated that substantial gains can be achieved when exploiting knowledge from previous projects. With the advent of tailor-made software such as the *synthpop* package in R (Nowok, Raab, and Dibben 2016), the investments for generating useful synthetic data might be further reduced in the future.

However, even without fine-tuning or customization of models, the current synthetic data have, in fact, proven useful. De facto, many deployments of synthetic data, including the Synthetic LBD in the US, have been used for model preparation by researchers in a public or lower-security environment, with subsequent remote submission of prepared code for validation against the confidential data. When viewed through the lens of such a validation system, the synthetic data prepared here would seem to have reasonable utility. While time series dynamics are not the same, they are broadly similar. Models converged in similar fashions, and while coefficients were strictly different, they were broadly similar and plausible. Specification tests did not lead to the same conclusions, but they also did not collapse or yield meaningless conclusions. Thus, we believe that the synthetic data, despite being different, have the potential to be a useful tool for analysts to prepare models without direct access to the confidential data. Vilhuber and Abowd (2016) and Vil-

huber (2019) come to a similar conclusion when evaluating usage of the synthetic datasets available through the Synthetic Data Server (Abowd and Vilhuber 2010), including the Synthetic LBD. A more thorough evaluation would need to explicitly measure the investment in synthetic data generation, the cost of setting up a validation structure, and the number of studies enabled through such a setup. We note that such an evaluation is non-trivial: the counter-factual in many circumstances is that no access is allowed to sensitive business microdata, or that access occurs through a secure research data system that is also costly to maintain. This study has contributed to such a future evaluation by showing that plausible results can be achieved with relatively low up-front investments.

The use of synthetic datasets to broaden access to confidential microdata is likely to increase in the near future, with increasing concerns by statistical agencies regarding the disclosure risks of releasing microdata. The resulting reduction in access to scientific microdata is overwhelmingly seen as problematic. Broadly "plausible" if not analytically valid synthetic datasets such as those described in this paper, combined with scalable remote submission systems that integrate modern disclosure avoidance mechanisms, may be a feasible mitigation strategy.

Acknowledgements

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. Alam was a part-time employee of Statistics Canada when this research was conducted. Alam thanks Claudiu Motoc and Danny Leung for help with the Canadian data. 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". The creation of the Synthetic LBD was funded by NSF Grant SES-0427889.

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". *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., Bryce E. Stephens, Lars Vilhuber, Fredrik Andersson, Kevin L. McKinney, Marc Roemer, and Simon D. Woodcock. 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.

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

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, John Haltiwanger, Julia Lane, and Kevin McKinney. 2007. "Using Worker Flows in the Analysis of the Firm". *Journal of Business and Economic Statistics* 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.

Blundell, Richard, Stephen Bond, and Frank Windmeijer. 2001. "Estimation in dynamic panel data models: Improving on the performance of the standard GMM estimator". In *Nonstationary Panels, Panel Cointegration, and Dynamic Panels*, ed. by Badi H. Baltagi, Thomas B. Fomby, and R. Carter Hill, 15:53–91. Advances in Econometrics. Emerald Group Publishing Limited. ISBN: 9781849500654 9780762306886, visited on 04/30/2020. doi:10.1016/S0731-9053(00)15003-0. https://doi.org/10.1016/S0731-9053(00)15003-0.

Bundesagentur für Arbeit. 2013. *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].

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. 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.

— . 2011b. *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, Jörg, Agnes Dundler, Stefan Bender, Susanne Rässler, and Thomas Zwick. 2008. "A new approach for disclosure control in the IAB establishment panel—multiple imputation for a better data access". *AStA Advances in Statistical Analysis* 92 (4): 439–458.

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.

Hansen, Lars Peter. 1982. "Large Sample Properties of Generalized Method of Moments Estimators". *Econometrica* 50, no. 4 (): 1029. ISSN: 00129682, visited on 04/30/2020. doi:10.2307/1912775. https://www.jstor.org/stable/1912775?origin=crossref.

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., C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil. 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., Jerome P. Reiter, Arnold P. Reznek, Javier Miranda, Ron S. Jarmin, and John M. Abowd. 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.

National Research Council. 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.

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, Gillian M. Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. 2018. "General and specific utility measures for synthetic data". *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181 (3): 663–688. doi:10.1111/rssa.12358. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssa.12358. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12358.

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.

Statistics Canada. 2019a. *Business Register (BR)*. Visited on 01/30/2020. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey%5C&SDDS=1105.

— . 2019b. *Longitudinal Employment Analysis Program (LEAP)*. Visited on 01/30/2020. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey%5C&SDDS=8013.

— . 2019c. *Survey of Employment, Payrolls and Hours (SEPH)*. Visited on 01/30/2020. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey%5C&SDDS=2612.

Statistics Canada and Bureau of the Census. 1991. *Concordance between the Standard Industrial Classifications of Canada and the United States, 1980 Canadian SIC - 1987 United States SIC*. Catalogue No. 12-574E. Statistics Canada. Visited on 01/30/2020. http://publications.gc.ca/site/eng/9.847987/publication.html.

Statistisches Bundesamt. 2003. "Classification of Economic Activities, issue 2003 (WZ 2003)". Statistisches Bundesamt (Federal Statistical Office) of Germany. Visited on 02/02/2020. https://www.klassifikationsserver.de/klassService/index.jsp?variant=wz2003.

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 1975-2015 [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/.

— . 2019. *Utility of two synthetic data sets mediated through a validation server: Experience with the Cornell Synthetic Data Server*. Presentation. Conference on Current Trends in Survey Statistics. https://hdl.handle.net/1813/43883.

Vilhuber, Lars, and John M. Abowd. 2016. *Usage and outcomes of the Synthetic Data Server*. Presentation. Meetings of the Society of Labor Economists. https://hdl.handle.net/.

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.

Woo, Mi-Ja, Jerome P. Reiter, Anna Oganian, and Alan F. Karr. 2009. "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation". *Journal of Privacy and Confidentiality* 1, no. 1 (). doi:10.29012/jpc.v1i1.568. https://journalprivacyconfidentiality.org/index.php/jpc/article/view/568.

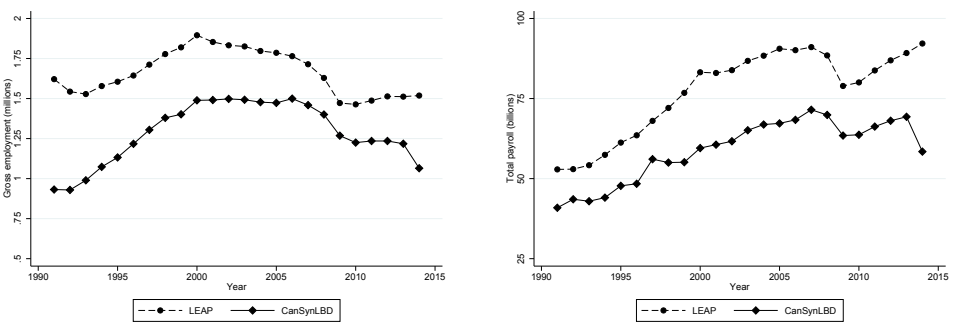
Woodcock, Simon D., and Gary Benedetto. 2009. "Distribution-preserving statistical disclosure limitation". *Computational Statistics & Data Analysis* 53 (12): 4228–4242. ISSN: 0167-9473. doi:https://doi.org/10.1016/j.csda.2009.05.020. http://www.sciencedirect.com/science/article/pii/S0167947309002011.

Online Appendix

"Applying Data Synthesis for Longitudinal Business Data across Three Countries"

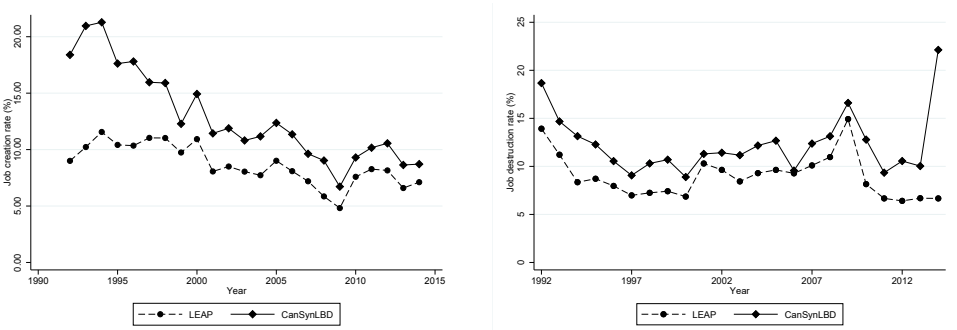
M. Jahangir Alam, Benoit Dostie, Jörg Drechsler, Lars Vilhuber

A. Figures for the Manufacturing Sector in Canada



(a) Gross employment level by year (b) Total payroll

Figure 8: Entity characteristics for the manufacturing sector in Canada by year.



(a) Job creation rates (b) Job destruction rates

Figure 9: Dynamics of job flows for the manufacturing sector in Canada by year.

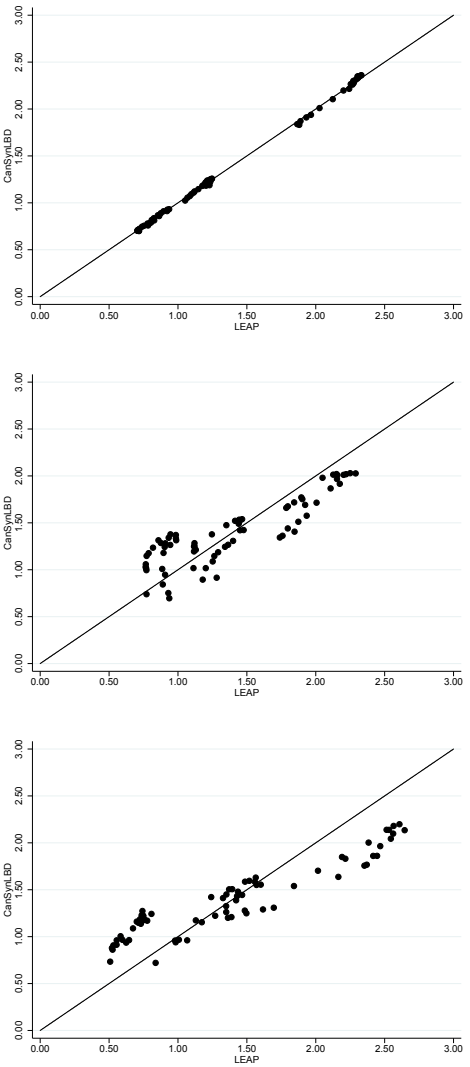


Figure 10: Share of entities (upper panel), share of employment (middle panel), and share of payroll (lower panel) by year and industry for the Canadian manufacturing sector.

Table 4: Detailed results for pMSE estimation by sector and country

Independent Variables	**Canada**		**Germany**
Sector:	Manufacturing	Private	All
Ln ALU	0.158	0.7138	-0.2895
	(0.0039)	(0.001)	(0.0033)
Ln Pay	0.0039	-0.4426	0.2584
	(0.0037)	(0.001)	(0.0028)
Age 3-4	0.0392	0.0972	-0.0987
	(0.0078)	(0.0017)	(0.007)
Age 5-7	-0.0382	0.0477	-0.0973
	(0.0073)	(0.0016)	(0.0066)
Age 8-12	-0.1258	-0.0263	-0.1172
	(0.0071)	(0.0015)	(0.0063)
Age 13 or more	-0.219	-0.1024	-0.1487
	(0.0074)	(0.0016)	(0.0059)
N	2243011	34638723	2121956
pseudo R-sq	0.0112	0.0318	0.0038
pMSE	0.0041	0.0121	0.0013

Note: See Equation 1 for estimation method. An observation is a entity-year in the combined database of each country-sector combination. All specifications include time and industry fixed effects. Standard errors are in parentheses.

B. Appendix Tables

B.1. pMSE

B.2. Regression analysis tables

Table 5: Regression coefficients (OLS) for LEAP

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. ***, **, and * indicate statistically significant coefficients at 1%, 5%, and 10% percent levels, respectively.

Table 6: Regression coefficients (OLS) for GLBD

Independent Variables	**GLBD**	**GSynLBD**
AR(1) Coefficient	0.4430***	0.4143***
	(0.0007)	(0.0008)
Ln Pay	0.4629***	0.5143***
	(0.0006)	(0.0007)
Age 3-4	-0.0695***	-0.0642***
	(0.0017)	(0.0016)
Age 5-7	-0.1066***	-0.0891***
	(0.0017)	(0.0016)
Age 8-12	-0.1324***	-0.1109***
	(0.0017)	(0.0016)
Age 13 or more	-0.1880***	-0.1600***
	(0.0016)	(0.0015)
N	848871	966084
R^2	0.9167	0.8968

Note: In all specifications, we include both year and industry fixed effects. Standard errors are in parentheses. ***, **, and * indicate statistically significant coefficients at 1%, 5%, and 10% percent levels, respectively.

Table 7: Regression coefficients (Dynamic) for LEAP

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. Standard errors are in parentheses. ***, **, and * indicate statistically significant coefficients at 1%, 5%, and 10% percent levels, respectively.

Table 8: Regression coefficients (Dynamic) for GLBD

Independent Variables	**GLBD**	**GSynLBD**
AR(1) Coefficient	0.0489***	0.6999***
	(0.0051)	(0.0057)
Ln Pay	0.7559***	0.2916***
	(0.0035)	(0.0042)
Age 3-4	-0.0070***	-0.1026***
	(0.0012)	(0.0015)
Age 5-7	-0.0233***	-0.1386***
	(0.0014)	(0.0017)
Age 8-12	-0.0473***	-0.1694***
	(0.0015)	(0.0018)
Age 13 or more	-0.1084***	-0.2183***
	(0.0015)	(0.0018)
N	848871	966084
m2	-2.5100	-4.1300
Sargan test	3.6e+03	2.0e+03
df of Sargan Test	495.0000	495.0000
P value of Sargan test	0.0000	0.0000

Note: In this table, $m2$ is the Arellano-Bond test for zero autocorrelation in first-differenced errors for order two. Standard errors are in parentheses. ***, **, and * indicate statistically significant coefficients at 1%, 5%, and 10% percent levels, respectively.

Table 9: Regression coefficients (Dynamic - system GMM) for LEAP

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 an entity-year. In this table, $m2$ is the Arellano-Bond test for zero autocorrelation in first-differenced errors for order two. Standard errors are in parentheses. ***, **, and * indicate statistically significant coefficients at 1%, 5%, and 10% percent levels, respectively.

Table 10: Regression coefficients (Dynamic - system GMM) for GLBD

Independent Variables	**GLBD**	**GSynLBD**
AR(1) Coefficient	0.1883***	0.6140***
	(0.0021)	(0.0027)
Ln Pay	0.6599***	0.3553***
	(0.0014)	(0.0020)
Age 3-4	-0.0292***	-0.0934***
	(0.0011)	(0.0013)
Age 5-7	-0.0512***	-0.1266***
	(0.0011)	(0.0014)
Age 8-12	-0.0791***	-0.1545***
	(0.0011)	(0.0015)
Age 13 or more	-0.1400***	-0.2012***
	(0.0011)	(0.0015)
N	848871	966084
m2	19.4900	-8.8300
Sargan test	4.5e+03	2.8e+03
df of Sargan Test	526.0000	526.0000
P value of Sargan test	0.0000	0.0000

Note: An observation is an entity-year. In this table, $m2$ is the Arellano-Bond test for zero autocorrelation in first-differenced errors for order two. Standard errors are in parentheses. ***, **, and * indicate statistically significant coefficients at 1%, 5%, and 10% percent levels, respectively.

Table 11: Regression coefficients (Dynamic - system GMM with MA(1)) for LEAP

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.

Table 12: Regression coefficients (Dynamic - system GMM with MA(1)) for GLBD

Independent Variables	**GLBD**	**GSynLBD**
AR(1) Coefficient	0.3701***	0.5268***
	(0.0060)	(0.0048)
Ln Pay	0.5349***	0.4202***
	(0.0041)	(0.0036)
Age 3-4	-0.0594***	-0.0831***
	(0.0015)	(0.0013)
Age 5-7	-0.0922***	-0.1105***
	(0.0018)	(0.0015)
Age 8-12	-0.1252***	-0.1351***
	(0.0019)	(0.0016)
Age 13 or more	-0.1850***	-0.1802***
	(0.0019)	(0.0017)
N	848871	966084
m2	19.0300	-11.6900
Sargan test	3.1e+03	2.5e+03
df of Sargan Test	494.0000	494.0000
P value of Sargan test	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. Standard errors are in parentheses. ***, **, and * indicate statistically significant coefficients at 1%, 5%, and 10% percent levels, respectively.

C. Canada: Synthesized Observations

Table 13: 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).

D. Confidentiality assessment

Table 14: Observed entity births given synthetic births for LEAP.

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 15: Observed entity births given synthetic births (GLBD)

Birth Year		% of Births over NAICS		
Synthetic	Actual	Minimum	Mean	Maximum
1976	1976	18.34	19.77	21.20
1977	1977	1.35	1.55	1.75
1978	1978	0.97	1.50	2.02
1979	1979	1.99	2.05	2.11
1980	1980	1.15	1.61	2.07
1981	1981	0.76	1.28	1.80
1982	1982	1.29	1.39	1.48
1983	1983	1.54	1.57	1.61
1984	1984	0.99	1.03	1.07
1985	1985	0.83	1.56	2.28
1986	1986	1.36	1.79	2.21
1987	1987	1.99	2.00	2.02
1988	1988	1.18	1.49	1.81
1989	1989	1.65	1.84	2.03
1990	1990	2.44	2.79	3.14
1991	1991	7.59	9.17	10.75
1992	1992	5.19	8.81	12.42
1993	1993	3.20	3.40	3.60
1994	1994	3.50	3.93	4.35
1995	1995	2.86	3.26	3.65
1996	1996	1.89	2.62	3.35
1997	1997	3.46	3.96	4.45
1998	1998	3.58	3.68	3.78
1999	1999	5.56	5.78	6.00
2000	2000	3.19	3.64	4.10
2001	2001	3.26	3.59	3.93
2002	2002	2.04	3.00	3.97
2003	2003	2.13	3.17	4.20
2004	2004	2.57	3.24	3.91
2005	2005	1.66	2.54	3.41
2006	2006	2.15	3.06	3.97
2007	2007	2.17	2.90	3.62
2008	2008	2.37	2.42	2.47