Cross-National Longitudinal Business Database: A Synthetic Data Approach

Vilhuber, Lars[1,2], Miranda, Javier[2], Kinney, Satkartar[3], Reiter, Jerome[4], [*]

@ 2013 Comparative Analysis of Enterprise Data Conference
Date: Thursday, September 19, 2013
Location: Crabapple Room, Federal Reserve Bank of Atlanta, USA
Time: 8:20 to 9:00 a.m.

Abstract

In most countries, statistical agencies do not release establishment-level business microdata because doing so represents too large a risk to establishments' confidentiality. One potential approach for overcoming these risks is to release synthetic data. The US Census Bureau Center for Economic Studies in collaboration with Duke University, the National Institute of Statistical Sciencies, and Cornell University made available a synthetic public use file for the Longitudinal Business Database (LBD), an annual economic census of establishments in the United States comprising more than 20 million records dating back to 1976. The resulting product, dubbed the SynLBD, was released in 2010 and is the first-ever comprehensive business microdata set publicly released in the United States including data on establishments, employment and payroll, birth and death years, and industrial classification. This paper describes how to extend this approach in a cost-effective way to other countries, allowing to expand general access to business microdata, and permitting novel ways of cross-country analysis of business dynamics.

Keywords: confidentiality, comparative studies, Longitudinal Business Database, synthetic data

1 Introduction

In 2010, the Census Bureau made available the first analytically valid synthetic establishment microdata, the Synthetic LBD (SynLBD), through a restricted-access compute server at Cornell University. The data creation process is documented in Kinney et al. [2011]. Synthetic data are created by replacing sensitive values with repeated draws from a model fit to the original data [Rubin, 1993], in an approach that is closely related to multiple imputation. Kinney et al. [2011] demonstrate the dimensions across which the Synthetic LBD Version 2 achieves analytic validity.

The main purpose of the US Census Bureau's Synthetic LBD is to facilitate researcher access to establishment microdata in a way that preserves the confidentiality of the underlying entities' data. Establishment and firm microdata pose many challenges in this dimension, as they are sparse

[*] 1: Cornell University, Ithaca, NY, USA, 2: U.S. Census Bureau, Washington, D.C., USA, 3: National Institute of Statistics Sciences,Research Triangle Park, NC, USA 4: Duke University, Durham, NC, USA

and often unique. It is easy to think of firms or establishments that are dominant in a specific industry or geographic location to such degree that their identification would be trivial if their data were released. This is true for many countries. Consequently, it is not uncommon that access to establishment microdata, if granted at all, is provided through data enclaves (Research Data Centers), at headquarters, or some other limited access. If more general access has been granted to researchers (for instance, in Denmark), it is most often restricted to researchers affiliated with a particular country's universities. All of these restrictions on data access reduce the research output by increasing the cost to researchers of accessing the data.

The Synthetic LBD departs from this approach by making disclosable *synthetic* microdata available through a remotely accessible data server. The approach is mutually benefitial. Researchers can access public use servers at little or no cost within a few weeks of their initial application. Researchers, who may have doubts as to the validity of results obtained through this novel method, are offered the option to validate their model-based inferences on the full confidential microdata. They agree to this by sharing their code with employees of the agency who will then run it on the internal servers. The statistical agency has an interest in improving future versions of the synthetic data along existing and new dimensions. They can do so by leveraging the diversity of the researchers' models and analyzing discrepancies. The implementation of the Synthetic Data Server (SDS) at Cornell University (http://www.vrdc.cornell.edu/sds/) provides a streamlined method by allowing researchers to develop models using statistical software (SAS, Stata, R, Matlab) without any imposed restrictions. Upon request the statistical agency validates results against the confidential data. Given the structure of the synthetic data and of the SDS, replication on the confidential data is achieved with very little additional effort.

This paper describes the conditions necessary to implement this approach, with the goal of providing other countries' statistical agencies a straighforward toolkit to implement the same procedure on their own data. Our hope is that by implementing similar procedures on comparable business microdata, new research both within and across countries can be enabled. The ideal end result is a series of country-specific datasets on establishments and/or firms available within the same computing environment. We describe the data and software requirements for the lowest-cost approach, the disclosure protection statistics already implemented that can be used to achieve release of the data in this way, the validation procedures that an agency should agree to, and the likely cost of maintaining such procedures. We point to Miranda and Vilhuber [2013] for a description of the type of projects that have used the US version of the SynLBD, and limitations of its use. Drechsler and Vilhuber [2013] describe a first cross-country implementation project for a German version.

2 Underlying data: the Longitudinal Business Database

The key input is of course a longitudinal database on establishments and/or firms. In the United States, the creation of the Longitudinal Business Database (LBD) that underlies the SynLBD is described in detail in Jarmin and Miranda [2002]. Different countries create longitudinal linkages in different ways.

In the United States, the LBD is sourced from administrative records for all employer tax units operating in the United States with paid employees (universe coverage). These records are subsequently enhanced with Census collections, including Economic Censuses and the Company Organization Survey, to identify the establishments and firms associated with the administrative reporting unit. The Business Register is updated annually and provides an end of year snapshop of all employer business in the US.[1] The LBD is constructed from the annual snapshot by linking the business units over time using longitudinal establishment identifiers including internal census identifiers as well as name and address matching.

Most critically, the final database has information on:

- employment over time
- payroll over time
- birth, death,
- location
- industry (measured consistently over time)
- firm affiliation of employer establishments.

Challenges can arise defining all of the above, but in particular the linkage will differ across countries. In the United States,

- linkage is by administrative identifiers and probabilistic matches of name and addresses
- metrics are updated annually
- employment is a point in time measure capturing all employment for the payroll period covered by March 12
- payroll is an annual measure
- industry as reported can vary over time
- consistency of industry coding systems across time is achieved through probabilistic coding methods [see Fort, 2013].
- firm structure is partially derived from survey data.

In other countries, all of these may be different, and may not be adjustable by the data synthesizing team. Significant differences are likely to arise in the longitudinal linkage process. The US process appears to be fairly unique in adjusting longitudinal linkages using name and address matching; most other countries use worker-flow links, leading to conceptually somewhat different linkages. However, the differences themselves may not be critical for the synthesizing process.

[1] For this reason the Business Register is the frame for all of the Census Bureau's business surveys and servers as the repository of administrative business data.

3 Synthetic data generating process

The synthetic data generating process (version 2) relies on all of the above, except for the firm affiliation information. Industry can be an arbitrary hierarchical classification, allowing for at least 3 levels of classification (the US methodology has used SIC2/3/4 [Kinney et al., 2011] and currently it is using NAICS3/4/5; NACE coding is the basis of the German project). Location is not actively used, and is not disclosed in the final output product. The process synthesizes the life-span of all establishments, as well as the evolution of their employment, conditional on industry. Geography is not synthesized, but is suppressed from the released file. The data synthesis process involves fitting models for the sensitive information in the confidential data including birth and death year, employment, and payroll separately for each industry subgroup. The actual values are then replaced with data simulated from these models. The synthetic data released to the public protects confidentiality because reidentification of actual data is made difficult when the released data are not actual, collected values. The current US version 2.0 data is based on the Standard Industrial Classification (SIC) and extends through 2000; an updated file with data through 2011 based on NAICS is expected to be released soon.

Variations may include the release of the higher hierarchies of the industry code instead of the exact industry code, and the release of exact or coarsened geography.

Kinney and Reiter [2013] describes the development of version 3 of the synthesizing process, which leverages and synthesizes the firm linkage information. It is expected that later versions of the cross-national implementation can leverage robust versions of the newer estimation system; however, initial implementation is assumed to rely on version 2 of the code.

4 Software requirements

The existing software is implemented in SAS, with a particular dependency on SAS/IML. For convenience, Unix systems can leverage a custom job scheduler, written in shell script and relying only on Sqlite3; however, this is not critical. Estimation is in parallelizable chunks (one per second-level industry, e.g., SIC3 or NAICS4), where the memory and time requirements depend on each industry. Using 20 parallel threads, the US system has been run through in about 20 hours for a single implicate, with about 100GB of data generated (of which half are detailed processing logs which can be reduced).

The version 3 code currently under development leverages R statistical software, in addition to SAS, and may not rely on SAS/IML in its final version.

5 Releasing synthetic data: Disclosure avoidance analysis

The current code base contains statistics on the extent of the protection provided by the synthetic data. They have not been generalized to be a part of the software package, but work is under way to do so.

6 Validation server and replication support

It is impossible to model all possible relationships in the data (the nature of research is often to discover hitherto unexplored or unknown relations) and undoubtedly the synthetic data will only be as good as the models that we use to synthesize them. In this regard its use as a stand alone analytical tool should be discouraged unless validated results are provided.

The current methodology at the Census Bureau is to validate the results on the confidential data if the analysis runs error-free on the SDS. All such analyses are still reviewed through the standard disclosure avoidance review process at the Census Bureau. Restrictions as to the type of output that can be disclosed are standard. Regression output is typically disclosed without problems unless it is a simple dummy model producing cell means. In such cases the cells are subject to disclosure techniques and the researcher is asked to provide additional statistics. Tabular output beyond simple means are not validated due to the time intensive nature of disclosing these tables.

The validation process itself should be replicated with other synthetic implementations, as it is critical to acceptance by researchers. A central contact point to access the synthetic data of multiple countries is desirable, and a streamlined disclosure avoidance process will allow researchers to do cross-national analyses easily.

References

Jörg Drechsler and Lars Vilhuber. Replicating the Synthetic LBD with German establishment data. Presentation, World Statistics Conference, 2013.

Teresa Fort. Applying a consistent industry classification across time. Technical report, Center for Economic Studies, 2013.

Ron Jarmin and Javier Miranda. The Longitudinal Business Database. Discussion Paper CES-WP-02-17, U.S. Census Bureau, Center for Economic Studies, 2002.

S. K. Kinney and J. Reiter. SynLBD: providing firm characteristics on synthetic establishment data. Presentation, World Statistics Conference, 2013.

Satkartar K. Kinney, Jerome P. Reiter, Arnold P. Reznek, Javier Miranda, Ron S. Jarmin, and John M. Abowd. Towards unrestricted public use business microdata: The Synthetic Longitudinal Business Database. *International Statistical Review*, 79(3):362–384, December 2011. URL http://ideas.repec.org/a/bla/istatr/v79y2011i3p362-384.html.

Javier Miranda and Lars Vilhuber. Looking back on three years of Synthetic LBD Beta. Presentation, World Statistics Conference, 2013.

Donald B. Rubin. Discussion of statistical disclosure limitation. *Journal of Official Statistics*, 9(2): 461–468, 1993.