Looking back on three years of Synthetic LBD Beta

Miranda, Javier
U.S. Census Bureau, Washington, D.C., USA, javier.miranda@census.gov

Lars Vilhuber
Cornell University, Ithaca, NY, USA, lars.vilhuber@cornell.edu

Distributions of business data are typically much more skewed than those for household or individual data and public knowledge of the underlying units is greater. As a results, national statistical offices (NSOs) rarely release establishment or firm-level business microdata due to the risk to respondent confidentiality. One potential approach for overcoming these risks is to release synthetic data where the establishment data are simulated from statistical models designed to mimic the distributions of the real underlying microdata. The US Census Bureau's Center for Economic Studies in collaboration with Duke University, the National Institute of Statistical Sciences, and Cornell University made available a synthetic public use file for the Longitudinal Business Database (LBD) comprising more than 20 million records for all business establishment with paid employees 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 documents the scope of projects that have requested and used the SynLBD.

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

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 , in an approach that is closely related to multiple imputation [Rubin, 1993]. If the imputation model is correct, valid inferences can be obtained from the synthetic datasets, and Kinney et al. [2011] demonstrate the dimensions across which the SynLBD Version 2 achieves analytic validity.

The SynLBD is designed to facilitate researcher access to establishment microdata in a way that preserves the confidentiality of the underlying entities' data. The SynLBD is part of a larger strategy by the Census Bureau to provide better statistics on business dynamics as recommended National Research Council's Committee on National Statistics Haltiwanger et al. [2007]. These recommendations included both the development of new public-use data products and expanded research access to business microdata.

Providing researchers access to establishment and firm microdata while protecting respondent confidentiality poses many challenges, as the data are sparse and often unique. It is easy to think of firms or establishments that dominate a specific industry or geographic location to such degree that their identification would be trivial even in aggregated data. In the past, access to North American establishment microdata, if granted at all, has been through Research Data Centers (RDC): phys-

ical locations, to which researchers travel in order to access the data. The RDC data enclaves are monitored and administered by Census Bureau Employees. No output can leave these facilities unless previously reviewed by a disclosure review officer to ensure no confidential information is released. Access to the Research Data Centers is a time consuming process not only due to possible distance considerations but because it requires the review and approval by several government agencies of a written proposal as well as the successful completion by the researcher of a background investigation. The proposal has to demonstrate scientific merit, it also has to demonstrate it will provide benefits to Census Bureau programs, and that the research poses no risk of disclosure.

In the case of data on business dynamics, the Census Bureau followed a strategy of providing multiple access modes to better meet the needs and data users (see Foster et al. [2010] for more details). These include the development of public-use tabulations in the new Business Dynamics Statistics program (see Haltiwanger et al. [2008] for a description) as well as restricted access to the gold standard Longitudinal Business Database (LBD). These products are aimed users with different requirements and skill sets. Access to confidential information is not necessary for most research question and requires advanced research skills and significant effort to obtain the necessary approvals.

The Synthetic LBD is intended to serve two purposes that can't be met with either the BDS or access to the gold standard LBD. First, the SynLBD provides much easier access to microdata to study business dynamics and in certain research circumstance provide analytically valid results. Second, the SynLBD allows those sophisticated researchers whose research question requires access to the gold standard LBD a way to explore the data and develop and test code outside the RDC environment which increases their productivity once in the lab.

Since the current version of the SynLBD is only a first attempt to generate analytically valid synthetic business microdata, a feedback loop is implemented through a validation mechanism. Researchers, who may have doubts as to the validity of results obtained through this novel method, might wish to know how accurate their model-based inferences are to an equivalent analysis that uses the full, confidential microdata. The Census Bureau and its partners want to know how to improve future versions of the data, along existing and new dimensions, by leveraging the diversity of the researchers' models. 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, such as those imposed by remote processing servers housing confidential information. Doing so on a server designed to replicate the software stack and environment used on the Census Bureau's Research Data Center servers allows replication to proceed with very little additional effort. In this paper, we describe how many projects took up the offer of using the data through the Cornell Virtual RDC, what topics they covered, and how many have progressed to the replication stage. We start by briefly describing the SynLBD.

The Synthetic Longitudinal Business Database

The creation of the Longitudinal Business Database (LBD) that underlies the SynLBD is described in detail in Miranda and Jarmin [2002], that of the SynLBD in Kinney et al. [2011]; we briefly summarize the key characteristics of both here. The LBD is created from the U.S. Census Bureau's Business Register files. At its core are administrative records for all employer tax units operating in the United States with paid employees. 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./footnoteFor 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. 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. The database has information on birth, death, location, industry, and firm affiliation of employer establishments, as well as their employment and payroll over time, for the private non farm economy from 1976 up through the most recent available years (as of this writing, 2011). Due to its unique features the LBD has become the most requested dataset for research applications inside the RDCs. It supports an active research agenda on business entry and exit, gross employment flows, employment volatility, industrial organization and other topics that cannot be adequately addressed without establishment-level data.It is also the tabulation input to the U.S. Census Bureau's Business Dynamics Statistics.[1] Other statistics created from the underlying Business Register include the County Business Patterns (CBP).

The SynLBD is derived from the LBD as a partially synthetic database with analytic validity, by synthesizing the life-span of 21 million establishments, as well as the evolution of their employment, conditional on industry. [2] 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 of nearly 500 3-digit industry subgroups. The actual values are then replaced with data simulated from these models. The synthetic data released to the public protects confidentiality because re-identification of actual data is made difficult when the released data are not actual, collected values. The current version 2.0 is based on the Standard Industrial Classification (SIC) and extends through 2000.

Project access and diversity

In order to gain access to the data (and the server), researchers must provide an abstract of a project, and a description of the variables needed for their analysis to the Census Bureau officials in charge of the project.[3] Application decisions are based solely on feasibility, by evaluating whether the data necessary to conduct the analysis are included on the SynLBD Beta file. Decisions have generally occurred within 10 business days at which point researchers are directed to the Cornell Virtual RDC server where they can access their assigned protected project space via the use of a user name and password./footnoteNote the SynLBD cannot be downloaded from this site onto a private computer. Use of the Cornell VRDC server is a requirement at this point.

Validation instructions are provided to the researchers, and are simple: if the analysis runs error-free on the SDS, then researchers can request that programs be run against the confidential data. All such analyses are reviewed by Census Bureau Disclosure Review Officers, and approved output is provided to both the researchers as well as to the Statistics of Income Program at the United

[1]The Business Dynamics Statistics (http://www.census.gov/ces) was developed with partial funding from the Kauffman Foundation to make aggregate public use statistics from the LBD available to the public and researchers at large.

[2]The initial development of the SynLBD by researchers from Cornell, Duke, NISS and the Census Bureau was supported by NSF Grant SES - 0427889

[3]Detailed access protocols can be found at http://www.census.gov/ces/dataproducts/synlbd/index.html

States Internal Revenue Service (IRS).[4] 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.

Since the release of SynLBD version 2, 25 researchers from 21 different institutions in 3 different countries have requested access. Most projects focus on the lifecycle dynamics of businesses – age and size of establishments – and although the information provided at application does not allow a detailed view into each project, projects cover both empirical descriptions of the firm growth dynamics and employment development over the business cycle to data to estimate structural models of the economy. Some of the applicants are established (university-based) researchers, others are doctoral students searching for a thesis topic. Three of the researchers (12%) have so far requested validation, but nearly a quarter of the projects had only been approved in the month preceding this article. While it is not known precisely how many researchers have subsequently applied for access to the confidential LBD or other data through the Census Bureau's Research Data Center Network, quite a few are clearly exploratory projects. Finally, among the rejected projects, quite a few have been rejected because they requested data not currently available, such as firm identifiers, NAICS codes, or a longer time series, suggesting that there might be some pent-up demand for these features.

Role and Limitations of Synthetic Data

A statistical agency such as the Census Bureau collects vast amounts of information with the goal of serving as the source of quality data about the nation's people and economy. They can fulfill their mission only in as much as they can make the data easily and readily accessible to individuals and organizations that need it to make informed decisions. Synthetic data products such as the SynLBD are a new way to make information at the micro level accessible to an increasing number of data users while protecting the confidentiality of information that is entrusted to them. The SynLBD is able to replicate means of the population of businesses along certain dimensions such as establishment age, size, and detailed industry beyond what has been possible in the past. Its use as an analytical tool is limited by the number and complexity of the variables included for synthesis. Currently the SynLBD does not synthesize firm characteristics or their structure so the analysis are by necessity limited to the characteristics and evolution of the population of establishments in the US. However, the technology to create synthetic datasets continues to improve and only time will tell how far it goes. As regards the SynLBD the Census Bureau in collaboration with Duke University and the National Institute of Statistical is now developing a version of that will also reproduce firm characteristics. This effort is well on its way and results might be available over the next year.

Synthetic datsets open up new opportunities for access to microdata that were not available before. However, as the previous discussion should make clear this is still early in the development of these experimental datasets. It is difficult at best to fully understand the basic properties of any given synthetic data set vis-a-vis the gold standard data they're designed to mimic (e.g. the moments of key distributions of interest as well as moments of their joint distributions). 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

[4] Access to the confidential LBD requires approval by IRS.

use to synthesize them. In this regard its use as a stand alone analytical tool should be discouraged unless validated results are provided.

Despite possible limitations as an analytical tool synthetic datasets are opening up new access modes. For example, the Census Bureau currently validates results for researchers wishing to work with the SynLBD as part of its development. This can clearly form the bases of a remote access mode where researchers wishing to conduct research on confidential microdata develop their code and models using synthetic data only to later submit for replication on the confidential microdata. This form of access should be particularly appealing to statistical agencies facing legal constrains in the number and types of researchers that can access their confidential data directly (e.g. if only government employees can access the data and only for approved projects). Remote access modes can minimize the risk of any breech of confidential information by limiting the number of people that ever get to work with confidential information. Remote access modes need not be costly to maintain if properly set up. The initial investment should ensure that the synthetic and real environments mirror each other as much as possible. The resources needed to execute code and review output by authorized employees is limited [5].

There will be times when researchers ultimately need to access confidential data to conduct their research. In these cases Synthetic data are still useful in reducing costs to researchers. Researchers can access and become familiar with a synthetic versions of the files from their homes or universities. This is useful for researchers that would otherwise have to relocate to Research Data Centers for lengthy periods of time. Researchers can save the time they would need to be spent learning the basic features of the data and computing environments.

Synthetic datasets such as the SynLBD have brought additional benefits to the Census Bureau as well as partner institutions. For example, the SynLBD is currently used as a training tool by universities wishing to introduce their students to the use of large scale business microdatasets as well as disclosure techniques. Students can not only use these data to replicate existing studies but they can also investigate and examine the validity of their own research ideas.

Combined with actual data synthetic micro data sets can be used to create public use aggregated data products that do away with supressed cells while maximizing the number of analytically valid cells.

Next Steps

Work currently underway using the existing methodology will extend the data through 2010, using NAICS, and newer imputation methodology (version 3) is under development (see paper by Kinney and Reiter in this same session) to improve the analytic validity and extend the imputation to additional variables. The data will be made available on Cornell University's Synthetic Data Server under the current access and replication protocols posted at http://www.census.gov/ces/dataproducts/synlbd/index.html, and all current users of the SynLBD will be able to access the newer data.

References

Lucia Foster, Ron Jarmin, and Lynn Riggs. Resolving the tension between access and confidentiality: Past experience and future plans at the U.S. Census Bureau. *Statistical Journal of the IAOS*, 26:113–122, 2010.

John Haltiwanger, Lisa Lynch, and Christopher Mackie, editors. *Understanding Business Dynamics: An Inte-*

[5] Restrictions on the types of output allowed further limit the possibility of incurring large costs

grated Data System for America's Future. The National Academies Press for the National Research Council, 2007. ISBN 978-0-309-10492-0.

John Haltiwanger, Ron Jarmin, and Javier Miranda. Business Dynamic Statistics: An overview. Business Dynamic Statistics Briefing, Kauffman Foundation, 2008. URL `http://www.census.gov/ces/pdf/BDS_Overview_2009.pdf`.

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

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