Replicating the Synthetic LBD with German Establishment Data

Drechsler, Jörg
Institute for Employment Research, Nürnberg, Germany, joerg.drechsler@iab.de

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

One major criticism against the use of synthetic data has been that the efforts necessary to generate useful synthetic data are so intense that many statistical agencies cannot afford them. However, we argue in this paper that the field is still evolving and many lessons that have been learned in the early years of synthetic data generation can now be used in the development of new synthetic data products, considerably reducing the required investments. We evaluate whether synthetic data algorithms that have been developed in the U.S. to generate a synthetic version of the Longitudinal Business Database (LBD) can easily be transferred to generate a similar data product for other countries. We construct a German data product with information comparable to the LBD - the German Longitudinal Business Database (GLBD) - that is generated from different administrative sources at the Institute for Employment Research, Germany. In a second stage, the algorithms developed for the synthesis of the LBD will be applied to the GLBD. Extensive evaluations will illustrate whether the algorithms provide useful synthetic data without further adjustment. The ultimate goal of the project is to provide access to multiple synthetic datasets similar to the SynLBD at Cornell to enable comparative studies between countries. The Synthetic GLBD is a first step towards that goal.

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

Introduction

The release of synthetic data is an innovative approach to disseminate sensitive data to the public without violating the confidentiality of the survey respondents. The basic idea is to replace sensitive information by repeatedly drawing from a model fit to the original data. The approach, originally proposed by Rubin [1993], is closely related to the idea of multiple imputation for nonresponse with the key difference that not the missing values but the sensitive values are replaced by multiple imputations. If the imputation model is correct, valid inferences can be obtained from the synthetic datasets if the combining rules developed by Raghunathan et al. [2003] (for fully synthetic datasets) or Reiter [2003] (for partially synthetic datasets) are applied. In the last decade several data products based on this approach were released to the public [Abowd et al., 2009, Machanavajjhala et al., 2008, Kinney et al., 2011, Drechsler, 2012].

Despite these promising developments, many agencies complain that developing synthetic datasets for complex surveys is too labor intensive, takes too long, and requires experts who are familiar with the data on the one hand but also need detailed knowledge of Bayesian statistics and excellent modeling skills to generate synthetic data with a high level of data utility. Given that the development of the first synthetic data products was very time and cost intensive many small agencies argue that they simply cannot afford to fund research on synthetic data for several months or even years. However, this criticism of the synthetic data approach ignores that most research on synthetic data based on

real data only happened in the last couple of years and the first data products had to be developed from scratch with no experience regarding strategies how to proceed. But many lessons learned during the early years of synthetic data and also the use of mostly automated non-parametric methods such as CART models [Reiter, 2005] will simplify the development of future synthetic data products tremendously.

In this project we evaluate to what extent synthesis code developed for a specific dataset in the U.S. – the Longitudinal Business Database (LBD) – can easily be transferred to another dataset with comparable data structure. The first part of the project focuses on the construction of a dataset – the German Longitudinal Business Database (GLBD) – that contains information that is comparable to the information available in the confidential version of the LBD. Since no business register is available at the Institute for Employment Research, the GLBD will be constructed by aggregating the information on employees from the administrative sources of the German Federal Employment Agency. Once the GLBD is generated, the synthesis algorithms that were used for the SynLBD will be run on the data. We will evaluate whether the automated synthesis will generate useful datasets that offer analytical validity for a wide range of statistical analyses. If the synthesis is successful, the data will be made available at the Cornell VirtualRDC[1] to enable comparative studies between the U.S. and Germany based on the two synthetic datasets. The GLBD will be a first step: The ultimate goal of the project is to provide access to synthetic datasets from multiple different countries.

The remainder of the paper is structured as follows: In Section 2 we will briefly summarize the synthesis of the LBD and describe which variables are contained in the SynLBD. Section 3 provides a detailed overview regarding the steps that were taken to construct a German version of the LBD from our administrative sources. The paper concludes with an outlook and possible obstacles we expect when generating the synthetic version of the GLBD.

The Synthetic Longitudinal Business Database

The creation of the Longitudinal Business Database (LBD) is described in detail in Miranda and Jarmin [2002], that of the Synthetic LBD 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 by creating longitudinal links of establishments using name and address matching. The database has information on birth, death, location, industry, and firm affiliation of employer establishments, as well as their employment over time, for nearly all sectors of the economy from 1976 up through the most recent available years (as of this writing, 2011). It is used both as a key file for research applications as well as tabulation input to the U.S. Census Bureau's Business Dynamics Statistics. Other statistics created from the underlying Business Register include the County Business Patterns (CBP).

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

[1] http://www.vrdc.cornell.edu/sds/

ables. In this paper, when we refer to the "SynLBD algorithms", we refer to Version 2.

Constructing the German Longitudinal Business Database

Except for the IAB Establishment Panel and the IAB Job Vacancy Survey no data are collected at the business or establishment level at the Institute for Employment Research. Instead, establishment level information is derived by aggregating the information contained in the German Social Security Data to the establishment level. The German Social Security Data (GSSD) is based on the integrated notification procedure for the health, pension and unemployment insurances, which was introduced in January 1973. Since then each employer is required to provide information on all employees on a regular basis. One of the data products derived from the GSSD is the Employment History Panel (BHP) which provides detailed information on all establishments covered in the GSSD by aggregating the employee level information via the establishment ID. The BHP will be the main data source for constructing a German equivalent to the synthetic version of the LBD. Currently, information from the years 1975 until 2008 are available for Western Germany. Information for the former Eastern German States is limited to the years 1992-2008. Unfortunately, although the BHP contains many more variables (about ??) than the synLBD including very detailed information on the personnel structure of the establishments, not all the variables contained in the synLBD are available in the BHP. Since the data are based on employee level information, the information whether the establishment belongs to a multi unit business cannot be obtained. Furthermore, until 1999 the BHP only contains establishments that had at least one employee covered by social security on the reference date June 30 of each year, since before that employers were only required to notify all employees covered by social security. Finally, the payroll information contained in the BHP is also based on the reference date June 30 each year. The SynLBD on the other hand contains the yearly payroll of each establishment. We will describe below, how we added yearly payroll information when generating the German Longitudinal Business Database (GLBD).
Table 1 in the online appendix lists the variables that were extracted from the BHP to form the basis for the GLBD. There are four different industry classification codes ($WZ73$, $WZ93$, $WZ03$, $WZ08$), as the classification changed four times during the reference period 1975–2008. Below we will describe how we derive a consistent industry code for all establishments in the data. The three different variables on the number of employees ($Employment_{tot}$, $Employment_{ss}$, $Employment_{me}$) are included since the notification requirements in Germany changed in 1999. Prior to 1999, establishments were only required to report employees covered by social security. Since 1999 all employees must be reported. As a consequence the total number of employees increases substantially between 1998 and 1999 and the same holds for the number of establishments covered in the BHP since establishments that only hired marginal employees are also now included in the BHP. To keep the data consistent, we subtracted the number of employees with marginal employment from the total number of employees and set the total number of employees to missing for all establishments that had zero employment after the subtraction of the marginal employment. Finally, we deleted the 967,086 establishments that never had any employees covered by social security. The final dataset consisted of 6,916,183 establishments.

Generating a unique geographic location and industry code: In the LBD geographic location and in-

dustry code are constant over the entire lifespan of an establishment. This is not true for the BHP for which it is possible that both variables can change from year to year. To maintain the consistency with the LBD, we apply the methodology of the LBD to create time-invariant geography and industry by selecting the mode of the reported geographic location and industry over the lifespan for each establishment. If several modes existed, we selected the mode that appeared first. This approach could be improved by selecting the mode randomly or by selecting the mode for which the average number of employees was highest. However, since only a small number of records contained a change in location or industry code and even less would have two or more modes, we did not implement this. Table 2 in the online appendix provides an overview of the number of establishments that reported a change over their lifespan for the different variables. The difference in the number of changes between $WZ73$, $WZ93$, and $WZ03$ is due to the fact that the variables differ in the number of years and the number of records for which they are collected. The variable $WZ08$ is not included since this variable was only collected in 2008 and thus no changes are possible.

Updating the information on establishment births and deaths: The information on the first year and last year an establishment is observed in the BHP is not necessarily equivalent with the birth and the death of the establishment for two reasons. First, the data in the BHP are necessarily left censored at 1975 (1991 for establishments from former Eastern Germany) and right censored at 2008. Second, new establishments appear in the BHP whenever a new establishment ID is generated. However, this doesn't necessarily mean that the new establishment ID belongs to a new establishment. Several scenarios are possible under which the occurrence of a new establishment and vice versa the disappearance of an establishment are not equivalent to the birth of a new establishment or death of an existing one. New establishment IDs are also assigned to an existing establishment if the ownership or the declared industry classification changes. In general, these establishments should not be treated as different entities when studying births and deaths of establishments. To distinguish actual births and deaths from these spurious ones, we rely on flow-based link files developed by Hethey and Schmieder [2010], following the methodology of Benedetto et al. [2007], and merge them to the BHP via the establishment ID. Hethey and Schmieder [2010] develop several different exit and entry classification. The basic idea is that if all (or most of) the employees from an exiting establishment work in a new establishment in the following year, it is prudent to assume that this is actually the same establishment and the occurrence of a new establishment and disappearance of an existing one are only a result of a change in ownership or declared industry classification. Likewise, if all (or most of) the employees of a new establishment worked in the same establishment in the year before but this establishment still exists in the current year, the new establishment is most likely a spin-off. For the GLBD, we treat all birth and death categories as actual births and deaths. In the absence of other information, we also treat all establishments with unknown status as actual births and deaths acknowledging that this might lead to a slight overestimation of the real numbers. Only a very small number of establishments actually fall into this category, and we don't expect major effects from this decision. For those establishments that are identified as ID changers we join the two separate records into a single establishment record that combines the information from the two. All employment and payroll related information is generated by aggregating the information from the two records. For variables that should not change over the lifespan of the establishment such as region and industry, the value for the combined record was taken from the establishment that was observed for a longer period in

the BHP, i.e. we set the value equal to the mode of that variable. We discard links classified as spin-offs.

Adding payroll information: The BHP contains information on the payroll of each establishment only on a reference date basis. Several quantiles of the payroll at June 30 of each year are included in the dataset. However, the LBD contains information on the yearly payroll of each establishment. Thus we needed to add this information to the original BHP data. While it would have been possible to obtain this information from the underlying administrative data directly by aggregating the income information for each employee that was ever employed at each establishment in a given year, we used similar information from a data product that was produced for a different project at the IAB. This data product contains the yearly payroll for all full time employees for all establishments that had at least one employee in the given year whose wage is above the limit for marginal employment for that year. Because of this additional filtering, the payroll information is not available for all establishments in the BHP. For almost 230,000 (3.3%) of the 6,916,183 records in our dataset no payroll information was available. Additionally 12 establishments couldn't be linked to the BHP because no establishment ID was available. Furthermore, the reported payroll is based on full time employees only, whereas the payroll in the LBD is based on all employees. Payroll information for all establishments in the BHP based on all employees from the underlying administrative data could be incorporated in the future.

Plans for the imputation of the industry classification: As a next step we plan to impute the industry classification whenever it is missing due to the changes in the reporting system (for example the WZ03 classification is never observed before 2003). For this we will use a simple probabilistic crosswalk based on the methodology used for the LEHD ECF [Abowd et al., 2009]. The methodology relies on double-coding for at least some periods, and uses $P(WZ08|WZ03 = wz03)$ to impute $WZ08$ codes even in periods where the $WZ08$ system did not yet exist, and similarly for othe periods. Subsequent to the imputation, the modal industry across all years for each establishment is computed. We will assess the sensitivity of the procedure to the use of employment weights.

Next Steps

Once the core GLBD is created, we will apply the SynLBD data synthesizing algorithms developed by Kinney et al. [2011] and extended by us to the confidential German data. Disclosure avoidance analysis will be performed, and the appropriate release protocols at IAB will be addressed. The data will be made available on Cornell University's Synthetic Data Server using access and replication protocols that will be posted at http://www.vrdc.cornell.edu/sds/. We expect that these protocols will mirror those currently in effect for the American SynLBD, the most current version of which can be accessed at http://www.census.gov/ces/dataproducts/synlbd/index.html.

References

John M. Abowd, Bryce E. Stephens, Lars Vilhuber, Fredrik Andersson, Kevin L. McKinney, Marc Roemer, and Simon D. Woodcock. The LEHD infrastructure files and the creation of the Quarterly Workforce Indicators. In Timothy Dunne, J. Brad Jensen, and Mark J. Roberts, editors, *Timothy Dunne, J. Bradford Jensen, and Mark J. Roberts*. University of Chicago Press, 2009.

Gary Benedetto, John Haltiwanger, Julia Lane, and Kevin McKinney. Using worker flows in the analysis of the

firm. *Journal of Business and Economic Statistics*, 25(3):299–313, July 2007.

Jörg Drechsler. New data dissemination approaches in old Europe – synthetic datasets for a German establishment survey. *Journal of Applied Statistics*, 39(2):243–265, April 2012. URL `http://ideas.repec.org/a/taf/japsta/v39y2012i2p243-265.html`.

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

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`.

Ashwin Machanavajjhala, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. *International Conference on Data Engineering (ICDE)*, 2008.

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

T.E. Raghunathan, J.P. Reiter, and D.B. Rubin. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19:1–16, 2003.

J.P. Reiter. Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29(2):181–188, 2003.

J.P. Reiter. Using cart to generate partially synthetic public use microdata. *Journal of Official Statistics*, pages 441–462, 2005.

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