9-22-2017

Proceedings from the Synthetic LBD International Seminar

Lars Vilhuber
Cornell University, lv39@cornell.edu

Saki Kinney
RTI

Ian M. Schmutte
University of Georgia, schmutte@uga.edu

Proceedings from the Synthetic LBD International Seminar

Abstract

On May 9, 2017, we hosted a seminar to discuss the conditions necessary to im- plement the SynLBD approach with interested parties, with the goal of providing a straightforward toolkit to implement the same procedure on other data. The proceed- ings summarize the discussions during the workshop.

Comments

Funding for the workshop was provided by the National Science Foundation (Grants 1012593; 1131848) and the Alfred P. Sloan Foundation (G-2015-13903). Organizational support was provided by the Labor Dynamics Institute at Cornell University.

Revised Sept 25 2017 for typographical errors.

Proceedings from the Synthetic LBD International Seminar

Lars Vilhuber Saki Kinney Ian M. Schmutte [*]

September 25, 2017

Abstract

On May 9, 2017, we hosted a seminar to discuss the conditions necessary to implement the SynLBD approach with interested parties, with the goal of providing a straightforward toolkit to implement the same procedure on other data. The proceedings summarize the discussions during the workshop.

[*] Funding for the workshop was provided by the National Science Foundation (Grants 1012593; 1131848) and the Alfred P. Sloan Foundation (G-2015-13903). Organizational support was provided by the Labor Dynamics Institute at Cornell University.

1 Introduction

Since 2010, the US Census Bureau's Synthetic LBD (Kinney et al., 2011b) has been made available to researchers through Cornell University's Synthetic Data Server (SDS). The main purpose of the Synthetic LBD is to facilitate researcher access to establishment microdata in a way that preserves the confidentiality of the underlying entities' data. The main purpose of the SDS is to allow researcher access to happen, and to enable a feedback loop that leverages research use to inform and improve future releases of the Synthetic LBD, including through improvements of its methodology.

Establishment and firm microdata pose many challenges to the application of disclosure avoidance techniques and thus to public distribution, as they are sparse and often unique. It is easy to concoct examples of firms and establishments that are so dominant in their industry or location that they would be immediately identified if their data were publicly 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 of statistical agencies, or some other limited means. These restrictions on data access reduce the growth of knowledge by increasing the cost to researchers of accessing the data.

Synthetic data are created by replacing sensitive values with repeated draws from a model fit to the original data (Little, 1993; Rubin, 1993), in an approach that is closely related to multiple imputation. 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 within a few weeks of their initial application and can later validate their model-based inferences on the full confidential microdata. The statistical agency has an interest in improving future versions of the synthetic data to be more accurate and reliable. They can do so by leveraging the diversity of the researchers' models and analyzing discrepancies. The SDS at Cornell University provides the infrastructure to implement this approach for two different synthetic datasets, with funding from the National Science Foundation (NSF) and the Alfred P. Sloan Foundation.

On May 9, 2017, we hosted a seminar to discuss the conditions necessary to implement the SynLBD approach with interested parties, with the goal of providing other statistical agencies, both in the US and abroad, a straightforward 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 longterm goal is a series of country-specific datasets on establishments and/or firms available within the same computing environment. The seminar brought together academics working on cutting-edge methods for the protection of privacy in statistical databases, along with researchers at statistical agencies who have started or are interested in developing synthetic business microdata. Five sessions touched on the full life cycle of the SynLBD development and implementation. In each session, we first discussed existing implementations and experiences, and then, as a group, discussed issues as they pertain to the broader community. Participants from several US agencies, Canada, and Germany were present. We discussed 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. The discussion was conducted following the Chatham House Rule[1].

2 History and Overview of SynLBD

The US LBD was created in the early 2000s (Miranda and Jarmin, 2002), following previous research files with more restrictive coverage. At its core, it is a research database containing longitudinally linked data records from a statistical business register of establishments. Breaks of longitudinal links are resolved using probabilistic name and address matching. The variables currently in the LBD are industry, annual payroll, employment, geography, birth year, death year, and firm structure. Though it has very few variables on the database itself, it serves as a backbone for many linkages into establishment and firm surveys and censuses at the US Census Bureau.[2]

The LBD provides coverage of the entire economy and is used to study economic activities such as business dynamics and job flows. The fundamental structure of the LBD (and thus the SynLBD) is a longitudinal file on economic entities, where each entity has a start and end date and a small number of key attributes that evolve over time. Hypothetically, this structure is shared by many other longitudinal panels, such as panels of jobs or of residences. We should note that it does not apply to data structures like a linked employer-employee database, since there are no linkages between entities at a point in time. Thus, using concepts from graph theory, it is a mapping of a network that contains only nodes, and no edges. These structural characteristics are relevant for any attempt to generalize the synthesizing methodology to other contexts, such as matched employer-employee data (but see Barrientos

[1] https://www.chathamhouse.org/about/chatham-house-rule

[2] An alternate database is the Business Information Tracking Series (BITS). There are efforts underway to bring LBD and BITS closer together.

et al., 2017).

The primary goal of the SynLBD project is to create partially synthetic microdata on establishments for public release, allowing researchers easier access for the implementation of a wide (unconstrained) range of models with analytically valid inferences about the underlying population, while protecting against re-identification of any given unit or its attributes. There are multiple reasons why a public release of such data is desirable. The US LBD is one of most requested datasets in the Federal Statistical Research Data Centers (FSRDCs), but access through the FSRDCs is still subject to long approval processes. In many European countries, access to data on business registers is arduous or impossible for researchers. Access through commercial providers is possible (Bureau Van Dijk), but coverage is generally poor.

The US SynLBD was released in 2010 to the Cornell SDS (see Section 6). The Census Bureau's Disclosure Review Board (DRB), as well as the Internal Revenue Service (IRS), classified SynLBD as public-use, but access is controlled due to concerns about the quality of the data. There are no disclosure concerns but researchers are cautioned not to trust results as if they were created by a traditional public-use file without going through the validation process. For similar reasons, the preparation of tabular data based on the synthetic data is strongly discouraged, and are not validated. Nevertheless, the synthetic data are of much easier access than the confidential data.

3 Synthesizing methodology

We briefly describe the synthesizing methodology here, a more detailed description is provided elsewhere (Kinney et al., 2011a,b). Currently, two versions of the process are in use. For the US, the "Phase 2" version is currently in its final stages (Kinney et al., 2014). In Germany, work is underway implementing the earlier "Phase 1" version (Drechsler and Vilhuber, 2014).

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 (and potentially unreleased) variables. Variables are synthesized in a sequential fashion, based on the representation of the joint distribution as a product of conditional distributions. Generically, categorical variables (birth and death years, multi-unit status) are processed first, using a variant of Dirichlet-Multinomial. It was mentioned that the German data does not identify multiunit status, however, the relevant step in the process can be used to process arbitrary meaningful indicator variables. Known disadvantages to this approach are the fact that one cannot use continuous predictors (this has not been an issue for the SynLBD), and that there is some subjectivity involved in dropping predictor variables when the synthetic

X do not exist in the observed data. The current ordering, as outlined in Kinney et al. (2011b), was determined based on experimentation, and should carry over to similar data.

After categorical variables, continuous variables are synthesized. It is often difficult to achieve good analytical fit with variables such as employment and payroll, since these variables are highly skewed. In the SynLBD, these variables are imputed year by year, and within each year, first employment and then payroll. Phase 1 used a normal linear regression model with kernel density-based transformation of the response (Woodcock and Benedetto, 2009). Phase 2 shifted to a CART model with Bayesian bootstrap.

The key unsynthesized but released variable is industry. Industry code was chosen for conditioning and release because it is considered public information. The U.S. Census Bureau considers both location and industry (activity) of an establishment as public knowledge about an establishment, though other attributes of the establishment thus identified are considered confidential. All synthesizing occurs within industry groups (though some collapsing of industry codes is done for efficiency reasons). It was also noted that Phase 1 of the SynLBD processing used Standard Industry Classification (SIC),[3] whereas Phase 2 development was done using North American Industry Classification System (NAICS)-based industry coding. This distinction is important since the distributions of multi-unit and single unit firms and establishments may look very different under different coding systems.

Job creation and destruction are biased in the Phase 1 approach (Drechsler and Vilhuber, 2014; Kinney et al., 2011b). It is a feature of the economic activity of most establishments that most employment levels do not change at all from year to year. The Phase 1 model for employment, using a kernel density estimator (KDE) transform combined with a linear regression approach, predicted too many changes, both positive and negative, at each point in time, resulting in biased estimates of job creation and destruction even though the net job flows were unbiased. The transition in Phase 2 to a Classification and Regression Trees (CART) approach helped fix this issue.

In general, tails of the employment and payroll distributions are relatively well preserved, though this turned out to be a disclosure concern in Phase 2. The CART synthesis approach tended to preserve extreme outliers excessively well, so that it was necessary to add an additional disclosure avoidance measure (in this case, multiplicative noise, prior to imputation). Another option to counter this issue, being currently considered, is the use of quantile regression.

While only one implicate has been released for SynLBD, multiple implicates are created. Even in Phase 2, there is not a lot of between-implicate variability for aggregate estimates.

[3] Note that the LBD-based statistics called Business Dynamics Statistics (BDS) are still only published by SIC.

If multiple implicates were to be released, between-implicate variability could be increased by adding a bootstrap step to the CART synthesis as a proxy for parameter draws.

3.1 Additional features

The group discussed additional options and features that are either being developed as part of Phase 2, or are of interest in alternate applications. Some relate to the structure and attributes of the synthetic data, others to attributes of the underlying population. For instance, one participant noted the absence of non-employers. This is a feature of the underlying confidential data – the LBD only covers an employer universe, a feature of many such registers – and would require some thought if it were to be expanded to non-employers, as the economic behavior of the self-employed differs substantially from that of even the smallest employers.

Phase 1 modeling did not account for firm structure, other than through an categorical variable summarizing multi-unit status over an establishment's lifetime. To properly account for this is a non-trivial extension of the original model. Phase 2 includes modeling of the firm structure (Kinney et al., 2014).

3.2 Lessons learned

Some comments were made on lessons learned over the course of model development. Modeling did not start by addressing everything at once, but leveraged the sequential structure to start with a few variables, and build up, obtaining feedback along the way. A participant noted, however, that it is helpful if the the order in which variables are synthesized can be specified early on. Changing the order of predictors later on creates some computational burden since the code is highly customized to the LBD. For example, the "Inactive status" indicators were the last set of variables to be added to the SynLBD synthesis; however, it made sense to synthesize it before multiunit, payroll, employment, and firm ID, since none of these are synthesized for inactive establishments. Consequently, the synthesis code for these variables needed to be updated to account for the new variable.

One of the participants noted that, within the overall process, the collaboration between subject matter experts (here, economists) and statisticians was found to be immensely valuable. The statisticians refined data building processes to highlight key business features identified by economists.

4 SynLBD Inputs

The group subsequently turned their attention to the process of defining and preparing the input data that is to be synthesized. In creating a synthetic database like SynLBD, one must consider the nature of the inputs. A discussion in the group centered around whether the data input should be unmodified ("raw"), preserving idiosyncracies in the data, or whether the data should be cleaned. The current approach, both for U.S. and German data, uses cleaned data as input to the synthesizing process. Implausible observations are edited, and missing data are imputed. The process explicitly precludes imputing missing data patterns. One reason for imputing missing values prior to synthesis is so that the completed data can be used to validate the synthesis process itself. Another alternative is to synthesize missingness. Imputing missing values simultaneously with synthesis is also possible but would complicate validation of researcher code.

Data cleaning simplifies the synthesizing process but there is some value lost as a consequence. Cleaned research data is an externality of the data synthesizing process. However, one must take care not to overdo the cleaning. The intent is to restrict the modifications to the minimum necessary, as one would for the actual analytical use of the data. For example, establishments may become "inactive" at certain times. This is a feature of economic activity that one would not wish the model to remove. Similarly, while some outliers might truly be aberrant, others are again intrinsic to the data generating process. The right amount of data preparation will depend on the particular context, and users will need to strike an appropriate balance between data cleaning and complexifying the data synthesis models.

A question arose regarding what information about cleaning is shared with the users. In the U.S. context, users in the FSRDC are already provided with suggested data cleaning code. In fact, the LBD itself is the product of data cleaning, with published general procedures (Miranda and Jarmin, 2002), though some parameters are not public knowledge. Cleaning German data is still under way, but the authors plan to document and publish the process, and make the cleaned data available in the German Institute for Employment Research (IAB) Research Data Centers (RDCs). Given that different researchers faced with raw data will make different decisions regarding data cleaning, another question was asked about the possibility of creating multiple versions of SynLBD using various versions of cleaned input. The discussion noted that there might be privacy issues associated with that process, as releasing multiple versions of the same input data will inevitably create better inferences about the confidential data.[4]. It was noted that users who would like to investigate different

[4] Nowok et al. (2016) propose a toolkit to generate custom synthetic data, by extension multiple independent synthetic data releases from the same input data, using different models.

methods of cleaning data can always request access to the confidential data through the FSRDC system.

Other issues related to data input and cleaning concern measurement of key concepts, such as employment, and classifications that change over time. For instance, in Germany, coverage of the underlying administrative data has expanded over time, so that "employment" includes part-time workers in later years that were out of scope in an earlier period. All countries face the regular updating of their industry classification systems - in the U.S., the SIC gave way to NAICS, which itself has evolved over time. German industry coding has gone through similar changes. Even though currently geography is not among the released variables, U.S. county changes over time would need to be taken into account. These changes lead to both conceptual and empirical challenges in preparing a time-consistent longitudinal database, and need to be adressed prior to synthesizing. The release of an updated U.S. SynLBD has been delayed in part in order to address the transition to NAICS. The work in Germany on developing an intertemporal industry classification crosswalk for the confidential data is of independent value, and is expected to be released to IAB RDC users soon.

5 Confidentiality protection

Confidentiality protection is, of course, the core purpose. The group discussed measures of how protective the data are, and how effective the approach might be in other legal and institutional contexts.

For the U.S. SynLBD, attribute disclosure was of greater concern than identity disclosures (Kinney et al., 2011b). The underlying universe data in the LBD are basically the same as those in the County Business Patterns (CBP) program, and the Census Bureau does not consider the existence of an establishment to be confidential. Nevertheless, re-identification based on attributes was a potential concern. The key protection comes from the fact that any observable characteristics used for re-identification are themselves synthetic. Preventing re-identification based on birth and death dates was the key criterion during the development of SynLBD. An establishment's birth and death years are synthesized such that with high probability they differ from the actual birth and death dates for a given establishment (Kinney et al., 2011b, Table 2). Attribute disclosure (about payroll and employment) was the other key concern. (Kinney et al., 2011b, Figure 13) showed that observed and synthesized employment differed in almost every year for establishments. These two outcomes are inherent to the synthesizing methodology, and should be transferable to other attributes as well. It was noted that the high protectiveness of Phase 1 data in terms of employment came at the cost of some lower analytic validity. Preliminary Phase 2 results based on the use of the CART model (Kinney et al., 2014) led to too strong correlation between synthesized and actual payroll by this criterion in the upper tails of the distribution, and necessitated the addition of other protection mechanisms, as noted earlier.

Although the U.S. SynLBD has only released a single implicate, it is in principle feasible to release multiple implicates. The specific establishment ID is randomly generated, and thus each implicate has the same weak correlation in attributes between each other as they do with the confidential data, providing another source of protection, should a data provider choose to release additional implicates.

For the U.S. Census Bureau, the criteria and metrics used for data releases such as the SynLBD are determined by the Census Bureau's DRB, in conjunction with the IRS. The group discussed various other confidentiality criteria in use by other agencies, such as Statistics Canada and the U.S. IRS, and how the SynLBD might conform to these.

The current Canadian criteria to assess confidentiality are governed by systems designed to evaluate output. For example, tools at Statistics Canada for tabulations might indicate if it satisfies rules like industry dominance. It seemed unclear what criterion might be used for a synthetic database such as the SynLBD. Any deviation from the current state of affairs would be contingent upon convincing policymakers of the validity of new protection measures.

In Germany, a distinction is made between public use data, scientific use data, and confidential data. Scientific use files require that recipients of the data sign agreements and provide some security for the data, though not at the level of a RDC. In turn, their disclosure avoidance requirements are based on the concept of "de facto anonymization" - the effort necessary to achieve reidentification is deemed to be higher than the "benefit" an attacker might receive from the reidentification or attribute disclosure. There is some experience with the release of partially synthetic data as scientific-use data (Drechsler, 2012). Disclosure avoidance criteria there were based on the risk of reidentification (taking into account underlying sampling uncertainty). The type of release (public-use or scientific use) of a more comprehensive longitudinal German SynLBD still remains to be assessed. The criteria for release are reviewed by lawyers at the level of the supervising government department (in the case of the IAB, the *Bundesarbeitsministerium*).

It was noted that the IRS might have interest in releasing longitudinal files, such as of tax filings. Participants with prior experience with the IRS noted that the criteria that have been used in the past have relied on the Euclidean distance between tax data sets, ensuring that they are not "too close". It should be noted that the release of the SynLBD required not only Census Bureau DRB approval, but also IRS approval, so in principle, that agency has experience with the criteria used for the SynLBD.

The question arose regarding how often to re-release synthetic data. For one, the linkage process used to construct the confidential data may use all available data to identify the most likely link, and as new years of data become available, the "best" link may change over time. On the other hand, providing additional years of data creates confidentiality issues. One of the fundamental confidentiality protection measures is the deviation of the synthetic birth year from the actual birth year. In the original release, the range of possible values is the entire year range, providing strong protection. However, this breaks down if adding only a single year, as it reveals the exact year of an establishment's birth or death, and is thus not recommended. Releases of additional data would need to rely on a full new synthesis (and the weak cross-implicate correlation).

The group also discussed whether a reframing risk in a differential privacy framework might help in this regard, but came to no conclusion. It should be noted that Kinney et al. (2011b) provided an ex-post measure of differential privacy.

6 Validation

From the start, users of the synthetic data were warned that, as a preliminary (beta) product, inferences might not always be valid. A mechanism was thus put in place to allow users of both the SynLBD and the SIPP Synthetic Beta (SSB) a means to obtain valid inferences, while at the same time contributing to improvements of the synthetic data generating process. This mechanism – called validation, had multiple components. At the end of the process, the Census Bureau would take researcher-provided programming code, run it against the confidential data, apply classical disclosure avoidance mechanisms to the model output, and provide the protected model output to the researcher. The rules applied, both in terms of documentation and disclosure avoidance checks, are the same as for FSRDC-based research.

Because of limited resources, researchers were required to provide code that ran error-free on the synthetic data, and to ensure that the transferability of the code was as smooth as possible, researchers were provided access to a remote computing environment modeled on that of the Census Bureau. This server, called the Synthetic Data Server (SDS), was funded by NSF and subsequently the Alfred P. Sloan Foundation and hosted by Cornell University (Vilhuber and Abowd, 2016).Users access the server remotely, but cannot remove data from the system. They are free to use the model output from the synthetic data, but are not allowed to create custom tabulations. Since the synthetic data are structured the same way the confidential data is (the database schema are the same), users can create the required documentation for validation requests quite easily: it is a verbose and technical description of the table created from the synthetic data, plus auxiliary documentation, such as dominance criteria and cell size counts.

For the Cornell SDS, about 10% of data access requests lead to validation requests. Vilhuber and Abowd (2016) describe preliminary results from several years of availability, a forthcoming document from the same authors will provide more complete results.

Thus, validation has two components: a front-end coding environment (essentially a type of integrated development environment that facilitates testing for compliance with validation requirements), and a back-end validation system which runs programs against confidential data. The validation mechanism is not real-time, but has quite short median turnaround times. The current mechanism is mostly manual - users interactively develop their programs on the SDS, and validation occurs by the same research staff that also develop the algorithms. Support on neither end is a full-time job, but is also not trivial. Improvements can occur along multiple lines. Regular user training helps users get acquainted with the (for them, unusual) system. This has been successfully done for the SSB, leading to user-organized conferences (Carr and Workers, 2016; Rutledge et al., 2016; Shore-Sheppard, 2016; Wicks-Lim, 2016) and accelerated turn-around time. Self-paced training can also help, reducing the learning curve for most users. On the back-end, automated submission and output checking mechanisms, possibly re-using systems already in place at statistical agencies, could further reduce turn-around time. The back-end system has strong similarities with remote-submission systems (Statistics Canada's Real Time Remote Access system[5] or the Luxemburg Income Study's LISSY system[6]). Additional improvements such as the use of verification servers (Barrientos et al., 2017) can further reduce the burden without (much) additional disclosure risk.

7 Next steps

One of the purposes of the workshop was to report on existing efforts to apply the U.S. SynLBD codebase to other data, whether other countries' longitudinal business data, or other data entirely. Multiple times over the course of the workshop, the question arose concerning the expandability of the overall mechanism, in particular the ability to add additional variables. Multiple (potential) users of the U.S. SynLBD have requested variables such as revenue, capital stock, or profits, and as noted earlier, geographic indicators. Because the basic mechanism relies on sequential conditional imputation, it is in principle feasible to either add additional variables at the end of the current process, or insert the creation of new variables into the process at a judiciously selected location. In general, experience suggests processing categorical variables before continuous variables. In Canada and Germany, efforts

[5] http://www.statcan.gc.ca/eng/rtra/rtra

[6] http://www.lisdatacenter.org/data-access/

are currently underway to apply the existing programs to national data to create a country-specific SynLBD. In the United States, efforts are underway to explore the feasibility of a synthetic panel file for tax returns.

Workshop participants expressed an interest in working groups and updates regarding the project, and a Github project will contain code and instructions for interested parties to leverage the existing codebase, tentatively at `https://github.com/labordynamicsinstitute/isynlbd`.

8 Acknowledgements

The organizers and the authors of these proceedings wish to express their thanks to all participants of the workshop. While many people contributed to the summary, all remaining errors and omissions are ours. We also wish to thank the National Academies' Committee on National Statistics for hosting the seminar in their Keck Center. We thank William Sexton (Cornell University) for excellent note taking, but all remaining errors are ours. This workshop could not have occurred without the pioneering work of Jerry Reiter, Arnie Reznek, Javier Miranda, Ron Jarmin, and John Abowd, without whose contributions the production and release of the original U.S. SynLBD would not have taken place.

References

Abowd, J. M., Gehrke, J., and Vilhuber, L. (2012). TC:Large: Practical privacy. Grant 1012593, National Science Foundation.

Abowd, J. M., Schmutte, I. M., and Vilhuber, L. (2015). The economics of socially-efficient privacy and confidentiality management for statistical agencies. Grant G-2015-13903, Alfred P. Sloan Foundation.

Abowd, J. M., Vilhuber, L., Li, P., and Block, W. (2011). NCRN-MN: Cornell Census-NSF Research Node: Integrated research support, training and data documentation. Grant 1131848, National Science Foundation.

Barrientos, A. F., Bolton, A., Balmat, T., Reiter, J. P., de Figueiredo, J. M., Machanavajjhala, A., Chen, Y., Kneifel, C., and DeLong, M. (2017). A framework for sharing confidential research data, applied to investigating differential pay by race in the u. s. government. Technical Report 1705.07872v1, arXiv.

Carr, M. D. and Workers, E. E. (2016). Education, gender, and earnings volatility: Evidence from sipp linked administrative data. Presentation at annual meetings, Allied Social Sciences Associations.

Drechsler, J. (2012). New data dissemination approaches in old europe – synthetic datasets for a german establishment survey. *Journal of Applied Statistics*, 39(2):243–265.

Drechsler, J. and Vilhuber, L. (2014). 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.

Kinney, S. K., Reiter, J. P., and Miranda, J. (2014). Improving The Synthetic Longitudinal Business Database. Working Papers 14-12, Center for Economic Studies, U.S. Census Bureau.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011a). Appendix to 'Towards unrestricted public use business microdata: The Synthetic Longitudinal Business Database'. online document, Center for Economic Studies, U.S. Census Bureau.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011b). Towards unrestricted public use business microdata: The Synthetic Longitudinal Business Database. *International Statistical Review*, 79(3):362–384.

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 2(2):407–426.

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

Nowok, B., Raab, G. M., and Dibben, C. (2016). synthpop : Bespoke creation of synthetic data in r. *Journal of Statistical Software*, 74:1–26.

Rubin, D. B. (1993). Satisfying Confidentiality Constraints Through Use of Synthetic Multiply-imputed Microdata (Discussion: Statistical Disclosure Limitation). *Journal of Official Statistics*, 9:461–468.

Rutledge, M. S., Wu, A. Y., and Vitagliano, F. (2016). Do tax incentives increase 401(k) retirement saving? evidence from the adoption of catch-up contributions. Presentation at annual meetings, Allied Social Sciences Associations.

Shore-Sheppard, L. (2016). Education, earnings, and the timing of fertility. Presentation at annual meetings, Allied Social Sciences Associations.

Vilhuber, L. and Abowd, J. M. (2016). Usage and outcomes of the synthetic data server. Presentation at society of labor economics meetings, Cornell University, Labor Dynamics Institute.

Wicks-Lim, J. (2016). Low-wage careers in a changing labor market. Presentation at annual meetings, Allied Social Sciences Associations.

Woodcock, S. D. and Benedetto, G. (2009). Distribution-preserving statistical disclosure limitation. *Computational Statistics & Data Analysis*, 53(12):4228–4242.