Statistical Journal of the IAOS 32 (2016) 65–68
DOI 10.3233/SJI-160964
IOS Press

Synthetic establishment microdata around the world

Lars Vilhuber[a,*], John M. Abowd[a] and Jerome P. Reiter[b]
[a]*Labor Dynamics Institute, Cornell University, Ithaca, NY, USA*
[b]*Department of Statistical Science, Duke University, Durham, NC, USA*

Abstract. In contrast to the many public-use microdata samples available for individual and household data from many statistical agencies around the world, there are virtually no establishment or firm microdata available. In large part, this difficulty in providing access to business microdata is due to the skewed and sparse distributions that characterize business data. Synthetic data are simulated data generated from statistical models. We organized sessions at the 2015 World Statistical Congress and the 2015 Joint Statistical Meetings, highlighting work on synthetic *establishment* microdata. This overview situates those papers, published in this issue, within the broader literature.

Keywords: Business data, confidentiality, international comparison, multiple imputation, synthetic, differential privacy

1. Introduction

Synthetic data are simulated data generated from statistical models. They are designed to protect the confidentiality of the people and firms in the underlying confidential data. The basic ideas can be traced back to Little [1] and Rubin [2]. Multiple imputation is often used for data that are missing due to non-response or some other feature of the data collection process that is outside of the data collector's control. In contrast, synthetic data for confidentiality protection scales this idea up to the entire dataset– explicitly replacing some or all observed data with model-generated data in order to protect the confidentiality of the underlying responding units. Whether used to address missing data, confidentiality protection or both, the methods share the goal of allowing users to obtain estimates with known statistical properties of at least some population parameters of interest.[1]

Synthetic microdata have been used to provide access to detailed confidential datasets in a secure fashion, and thus are also linked to a broader discussion of how best to provide access to such datasets to researchers [4–7]. Other methods include access to confidential microdata in secure data enclaves (e.g., research data centers of the U.S. Federal Statistical System, of the German Federal Employment Agency, others), and via remote submission system. Remote submission systems often provide researchers with test data, sometimes also called "synthetic data", so they can prepare analysis code for remote submission on their local computers. Such test data differ from the synthetic data in this overview in that they explicitly make no claim of statistical validity of any inferences made from the synthetic data.

In contrast to the many public-use microdata samples available for individual and household data from many statistical agencies around the world, there are virtually no establishment or firm microdata available. In large part, this difficulty in providing access to business microdata is due to the skewed and sparse distributions that characterize business data. In 2013, we organized a session at the World Statistical Congress[2] to

*Corresponding author: Lars Vilhuber, Labor Dynamics Institute, Cornell University, Ithaca, NY 14853, USA. E-mail: lars.vilhuber@cornell.edu.

[1]See [3] for a review of the theory and applications of the synthetic data methodology.

[2]2013 World Statistical Congress: http://2013.isiproceedings.

1874-7655/16/$35.00 © 2016 – IOS Press and the authors. All rights reserved
This article is published online with Open Access and distributed under the terms of the Creative Commons Attribution Non-Commercial License.

highlight work on synthetic *establishment* microdata, subsequently published in this journal [8–12].

As a follow-up, we organized similar sessions at the 2015 World Statistical Congress,[3] and at the 2015 Joint Statistical Meetings.[4] This overview, and the additional articles in this issue, stem from those sessions.

2. Synthetic longitudinal business database

In the United States, a key research file in the secure data enclaves of the U.S. federal statistical system is the Longitudinal Business Database (LBD) [13,14], a longitudinally-linked version of the U.S. employer business register. Using the LBD as the primary input, a synthetic dataset called the Synthetic Longitudinal Business Database (LBD) (SynLBD) [15] was generated, and released to an easily web-accessible computing environment [16] (a synthetic data set of a German business dataset was released at about the same time [17]). In addition, the Business Dynamics Statistics (BDS) are tabulated from the LBD, and protected using primary/complementary suppression techniques. The BDS were designed as public-use data that explicitly tabulated some of the estimates needed to study phenomena that cannot be studied with traditional tabulations (gross job creations and destructions, which are an establishment-level concept). For instance, [18] show that much job creation is driven by small and medium firms; however, in the published BDS, many of the suppressed cells are for precisely those types of firms and events. The article by Miranda and Vilhuber on "Using partially synthetic microdata to protect sensitive cells in business statistics" describes a potential use of the SynLBD for publishing tabulations for precisely those small cells, and thus potentially improving the analytical quality of the published statistics, without increasing confidentiality risk.[5] While their final conclusion is tentative until newer work on improving the SynLBD is made available [11], the method proposed is very much in the spirit of the original Rubin ideas [2]. Their work is also part of a broader effort to make consistently generated synthetic establishment microdata available.

The modeling strategy used for the SynLBD does not constrain the resulting synthetic data to match marginals in the confidential data. Wei and Reiter address this issue in "Releasing Synthetic Magnitude Microdata Constrained to Fixed Marginal Totals." By using mixtures of Poisson distributions, they can guarantee that the synthetic data, drawn from the posterior predictive distribution of the model, sum to the marginal totals produced from the confidential data, and illustrate this on dataset on manufacturing establishments.

3. Strengthening the protection mechanisms

McClure and Reiter's article "Assessing Disclosure Risks for Synthetic Data with Arbitrary Intruder Knowledge" investigates disclosure risks for synthetic data under different levels of intruder knowledge. Using simulation studies, they use Bayesian posterior probabilities to compute disclosure risks. They show that, in their studies, risks appear low for ordinary records but are higher for unusual records. They also show how intruders' abilities to infer about confidential values lessen with decreasing intruder information.

One of the scenarios considered by McClure and Reiter, when the intruder knows every data point but one, is closely related to the assumptions in differential privacy mechanisms. Schmutte, in "Differentially Private Publication of Data on Wages and Job Mobility," explicitly considers a differentially private publication mechanism for business-level statistics, and investigates the tradeoff between the privacy guaranteed to individuals present in the population, and the accuracy of the released statistics. He characterizes the realized tradeoff in generated data, but also finds, as in other cases, that model inference in the differentially-private synthetic data is poor when the analysis model is *uncongenial* [20] to the model generating the synthetic data. This point has been made in other contexts as well [7].

Finally, the article by Abowd and McKinney on "Noise Infusion as a Confidentiality Protection Measure for Graph-Based Statistics," while not formally on synthetic data models, addresses an issue quite prevalent in the analysis of linked data that includes establishments: how to publish information for graph-based statistics, for instance for flows of workers between establishments. Their solution leverages an existing noise-infusion protection mechanism [21], and extends it to the protection of the statistics generated

org/, accessed Dec 20, 2015

[3] http://www.isi2015.org/, accessed Dec 20, 2015.

[4] https://www.amstat.org/meetings/jsm/2015/, accessed December 20, 2015.

[5] Preliminary results from the same research effort were presented in [19].

from the projection of the employer-employee graph onto a single set of nodes. However, similar to the synthetic data models, no data on actual respondents are ever published. The method proposed here is used to protect the U.S. Census Bureau's newly released Job-to-Job flows [22].

4. Conclusions

Synthetic data methods and related protection mechanisms are an important part in the toolkit of agencies seeking to disseminate microdata. While applied in this context to establishment and firm data, synthetic data methods are valuable in the context of person and household data as well [23,24]. The aforementioned SIPP Synthetic Beta file (SSB) [25] is one particular example of synthetic data in a feedback loop, and the provision of synthetic data for custom extracts from the England and Wales Longitudinal Study (ONS LS), Scottish Longitudinal Study (SLS) and Northern Ireland Longitudinal Study (NILS) as part of the Synthetic Data Estimation for UK Longitudinal Studies (SYLLS) project is another,[6] The next step of creating even stronger privacy guarantees for synthetic data, through the use of differentially private mechanisms, as evidenced by two papers in this issue, as well as ongoing work on validation (such as the Synthetic Data Server (SDS), [16]) and verification servers [26], is expected to make additional microdata available to researchers from an increasing number of sources.

Acknowledgment

Abowd and Vilhuber acknowledge support through NSF Grant SES-1042181. Reiter acknowledges support through NSF grant SES-11-31897.

References

[1] R.J.A. Little, Statistical Analysis of Masked Data, *Journal of Official Statistics* **9**(2) (1993), 407–426.

[2] D.B. Rubin, Discussion of Statistical Disclosure Limitation, *Journal of Official Statistics* **9**(2) (1993), 461–468.

[3] J. Drechsler, Synthetic Datasets for Statistical Disclosure Control-Theory and Implementation, *New York: Springer*, 2011.

[4] S. Bender, The RDC of the Federal Employment Agency as a part of the German RDC Movement, In: Comparative Analysis of Enterprise Data, 2009 Conference, 2009. Available from: http://gcoe.ier.hit-u.ac.jp/CAED/index.html.

[5] L. Vilhuber, Methods for Protecting the Confidentiality of Firm-Level Data: Issues and Solutions, Labor Dynamics Institute, 2013, **19**. Available from: http://digitalcommons.ilr.cornell.edu/ldi/19/.

[6] J.M. Abowd and J.I. Lane, New Approaches to Confidentiality Protection Synthetic Data, Remote Access and Research Data Centers, in: *Privacy in Statistical Databases*, 2004, pp. 282–289. Available from: http://www.springer.com/la/book/9783540221180.

[7] J.M. Abowd and I. Schmutte, Economic analysis and statistical disclosure limitation. *Brookings Papers on Economic Activity*. Fall 2015. Available from: http://www.brookings.edu/about/projects/bpea/papers/2015/economic-analysis-statistical-disclosure-limitation.

[8] J. Miranda and L. Vilhuber, Looking Back On Three Years Of Using The Synthetic LBD Beta. *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics*. 2014, **30**. Available from: http://iospress.metapress.com/content/X415V18331Q33150.

[9] J. Drechsler and L. Vilhuber, 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) (2014), Available from: http://iospress.metapress.com/content/X415V18331Q33150.

[10] R.S. Jarmin, T.A. Louis and J. Miranda, 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) (2014), Available from: http://iospress.metapress.com/content/fl8434n4v38m4347/?p=00c99b98bf2f4701ae806ee638594915&pi=0.

[11] S.K. Kinney, J.P. Reiter and J. Miranda, Improving The Synthetic Longitudinal Business Database, *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics* **30**(2) (2014).

[12] J.M. Abowd, Synthetic establishment data: Origins and introduction to current research, *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics* **30**(2) (2014). Available from: http://iospress.metapress.com/content/76707M55W510VT16.

[13] R. Jarmin and J. Miranda, The Longitudinal Business Database. U.S. Census Bureau, Center for Economic Studies; 2002. CES-WP-02-17.

[14] U S Census Bureau, Longitudinal Business Database (LBD). Washington, DC USA: U.S. Census Bureau [distributor]; 2012. Available from: https://www.census.gov/ces/dataproducts/ datasets/lbd.html.

[15] S.K. Kinney, J.P. Reiter, A.P. Reznek, J. Miranda, R.S. Jarmin and J.M. Abowd, Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database, *International Statistical Review* **79**(3) (2011), 362–384. Available from: http://ideas.repec.org/a/bla/istatr/v79y2011i3p362-384.html.

[16] J.M. Abowd and L. Vilhuber, Synthetic Data Server; 2010. Available from: http://www.vrdc.cornell.edu/sds/.

[17] J. Drechsler, Synthetische Scientific-Use-Files der Welle 2007 des IAB-Betriebspanels, Institute for Employment Research, Nuremberg, Germany; 2011. 201101_de. Available from: http://ideas.repec.org/p/iab/iabfme/201101_de.html.

[6] http://www.lscs.ac.uk/projects/synthetic-data-estimation-for-uk-longitudinal-studies/. accessed on December 20, 2015.

[18] J.C. Haltiwanger, R.S. Jarmin and J. Miranda, Who Creates Jobs? Small vs. Large vs. Young. National Bureau of Economic Research, Inc; 2010. 16300. Available from: https://ideas.repec.org/p/nbr/nberwo/16300.html.

[19] J. Miranda and L. Vilhuber, Using Partially Synthetic Data to Replace Suppression in the Business Dynamics Statistics: Early Results, in: *Privacy in Statistical Databases, J. Domingo-Ferrer, ed., vol. **8744** of Lecture Notes in Computer Science. Springer International Publishing*; 2014, pp. 232–242. Available from: http://dx.doi.org/10.1007/978-3-319-11257-2_18.

[20] X.L. Meng, Multiple-imputation inferences with uncongenial sources of input, **Statistical Sciences 9**(4) (1994), 538–573.

[21] J.M. Abowd, B.E. Stephens, L. Vilhuber, F. Andersson, K.L. McKinney, M. Roemer et al., The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators, in: *T. Dunne, J.B. Jensen, M.J. Roberts, eds, Producer Dynamics: New Evidence from Micro Data. University of Chicago Press*, 2009.

[22] H. Hyatt, E. McEntarfer, K. McKinney, S. Tibbets, L. Vilhuber and D. Walton, Job-to-Job Flows: New Statistics on Worker Reallocation and Job Turnover, U.S. Census Bureau; 2015. Available from: http://lehd.ces.census.gov/doc/jobtojob_documentation_long.pdf.

[23] J. Drechsler and J.P. Reiter, Sampling With Synthesis: A New Approach for Releasing Public Use Census Microdata, *Journal of the American Statistical Association* **105**(492) (2010), 1347–1357. Available from: http://ideas.repec.org/a/bes/jnlasa/v105i492y2010p1347-1357.html.

[24] J. Hu, J.P. Reiter and Q. Wang, Dirichlet Process Mixture Models for Nested Categorical Data, ArXiv e-prints. 2014 Dec.

[25] J.M. Abowd, M. Stinson and G. Benedetto, Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project, U.S. Census Bureau; 2006. Available from: http://www2.vrdc.cornell.edu/news/?p=308.

[26] J.P. Reiter, A. Oganian and A.F. Karr, Verification servers: Enabling analysts to assess the quality of inferences from public use data, *Computational Statistics & Data Analysis* **53**(4) (2009), 1475–1482. Available from: http://dx.doi.org/10.1016/j.csda.2008.10.006.