Synthetic Establishment Data: Origins and Introduction to Current Research

John M Abowd

Cornell University, U.S. Census Bureau and IAB
john.abowd@cornell.edu

The papers in this section of the June *Statistical Journal of the IAOS* are an outgrowth of a project supported by the National Science Foundation (US-NSF), Census Bureau (US-Census), Internal Revenue Service (US-IRS), and Institut für Arbeitsmarkt- und Berufsforschung (DE-IAB) to promote collaboration between researchers at the statistical agencies (Census and IAB) and at Cornell, Duke, Michigan and Simon Fraser universities.[1] The goal of the research projects was to explore the potential for using synthetic data methods in the manner originally suggested by Little (1993) and Rubin (1993; see citation in Jarmin *et al.*, this section) for the protection and release of micro-data from establishment censuses, linked employer-employee administrative records, linked survey-administrative record data, large scale surveys, and household censuses. Summaries of the papers from this project were presented at the NSF-Census-IRS Workshop on Synthetic Data and Confidentiality Protection (2009).[2]

The appeal of synthetic data is founded on the observation that if the statistical methods used to produce these data are sufficiently accurate and also sufficiently protective, analysts can use them to do a wide variety of studies that would otherwise have required access to the original confidential data. But the programs described in this issue go further—they formally incorporate a feedback cycle between the users of the synthetic data and the custodians of the confidential data.

In this feedback system, the developers of the synthetic data, who are experts on both the structure and content of the confidential data, create an initial release of the synthetic product. This initial release mimics the logical structure of the underlying confidential data and is analytically valid for a particular set of statistical models that the developers preselected. The initial synthetic data release is then analyzed by external users who are not granted access to the underlying confidential data. These synthetic data analyses are conducted in a controlled environment that closely mimics the development environment that houses the confidential data. The Census Bureau then uses the synthetic data analyses—exactly the same ones the external users conducted—for validation studies conducted on the confidential data. The results of the confidential data analyses are released to the external users after conventional disclosure avoidance limitation methods have been applied. The confidential data custodians then use the results of these external user analyses to discover valuable enhancements of the synthetic data, which are then incorporated into the next release of the synthetic data. For the

---

[1] NSF ITR 0427889 http://www.nsf.gov/awardsearch/showAward?AWD_ID=0427889, cited on February 12, 2014.

[2] The program, papers, and presentations can be found at http://www2.vrdc.cornell.edu/news/workshops-and-sessions/nsf-census-irs-workshop2009/program/, cited on February 12, 2014.

Census Bureau's synthetic micro-data releases, the controlled environment is provided by the Synthetic Data Server run by the Cornell University VirtualRDC.[3]

In this *Issue* the authors provide further insight into the use of synthetic business data. There are a variety of definitions of synthetic micro-data. The papers in this issue define fully synthetic micro-data as records on which all variables are sampled from a formal probability model of the population of interest that has been fit using the confidential micro-data. Partially synthetic data are records on which some of the variables have been sampled from a formal probability model, again fit using the underlying confidential data, and other variables are copied directly from the confidential data.[4]

The first paper in this *Issue*: "Expanding the Role of Synthetic Data at the U.S. Census Bureau" by Ron Jarmin, Thomas Louis, and Javier Miranda recounts the history of the first two synthetic data products released by the Bureau. These were the SIPP-Synthetic Beta and the Synthetic Longitudinal Business Data Base. The former is based on longitudinally linked Surveys of Income and Program Participation, Internal Revenue Service tax data, and Social Security Administration benefits data. It formally incorporates the feedback mechanism described above and is currently in its fifth release.[5] The latter, which is the subject of the three other articles in this *Issue*, is based on longitudinally linked business establishment data from the Census Bureau's Longitudinal Business Database.[6] This synthetic data system also incorporates the feedback cycle, and is in its second version. After describing other R&D activities at the Census Bureau that also use synthetic data methods, Ron, Tom, and Javi call on all National Statistical Offices to "take the lead in [synthetic data] and other activities to increase the relevance and accessibility of high quality and reliable official statistics"—advice that the U.S. Census Bureau has taken to heart.

As I have already mentioned, an important feature of the synthetic data development process is the feedback loop that allows continuous improvement of the synthetic data quality and continuous enhancement of the underlying confidential data. The second paper in this *Issue*: "Looking back on three years of using the Synthetic LBD Beta" by Javier Miranda and Lars Vilhuber reports on the experience of the feedback loop for the synthetic U.S. business establishment data (SynLBD). SynLBD is in its second version. Version 1 was an experimental release. Version 2 (the current version) was the first with demonstrated analytical validity. This paper documents how 25 researchers who had used the SynLBD as of July 2013 were able to access the synthetic data, develop their scientific models, then have those models estimated on the underlying confidential LBD. They report on what has been learned about SynLBD and on how Version 3 will benefit from this feedback.

---

[3] http://www2.vrdc.cornell.edu/news/synthetic-data-server/, cited on February 12, 2014.

[4] Computer scientists in the data privacy domain often use a different definition in which the probability distribution used to create the sampled data is unrelated to any statistical model of the confidential data. See Abowd and Vilhuber (2008; see citation in Kinney *et al.*, this section).

[5] http://www.census.gov/programs-surveys/sipp/methodology/sipp-synthetic-beta-data-product.html, cited on February 12, 2014.

[6] https://www.census.gov/ces/dataproducts/synlbd/, cited on February 12, 2014.

Which brings us to the third paper in this *Issue*: "Improving the Synthetic Longitudinal Business Database" by Satkartar Kinney, Jerome Reiter, and Javier Miranda, who were the key developers of Version 2 of SynLBD. In this paper they describe the "second generation of the SynLBD," which will be Version 3 in the official numbering. They discuss and document how the next version of the SynLBD uses methods that address many of the key shortcomings identified by the SynLBD users that Javi and Lars studied. Reading this paper is like peering into the feedback machine while it is still turning. A knowledgeable user of the LBD (synthetic or confidential) will see in this work a serious attempt to confront and model statistical difficulties that plague any longitudinally linked establishment data base: how to handle births and deaths to control bias in job creation/destruction statistics; how to model the relationship of establishments to the business entities that own them; and what to do with the skewed distribution of employment and payroll at the establishment level, among many others.

The final paper in this *Issue*: "A first step to a German SynLBD: Constructing a German Longitudinal Business Database" by Jörg Drechsler and Lars Vilhuber documents their attempt to port the technology that was associated with the U.S. SynLBD for use with similar German business establishment data. Jörg was a collaborator in the original NSF project discussed above, and his work on that project led to single-year applications of synthetic data methods to Germany establishment surveys. Lars is the principal architect of the feedback system used for the SIPP Synthetic Data and SynLBD, as well as a very experienced user of the confidential LBD and related data. Jörg and Lars walk us through a case study in generalizing the concepts, methods, and computer code used for one application (SynLBD), then re-adapting them to the German case (Synthetic GLBD). They are particularly detailed in their discussion of how they first had to work with a different administrative record system (the German Social Security Data) in order to develop establishment-level variables comparable to the employment and payroll measures commonly found in such systems, but not in Germany. Then, they tackle the problem of standardizing the way activity and geography identifiers are conceptualized and implemented so that cross-national comparisons of the synthetic data are both feasible and sensible. Theirs is a project in the early stages of the feedback process, but one that has already benefited from the SynLBD's first cycle.

There is another aspect to research on the uses of synthetic data that is also reflected in this *Issue*. Two of the authors, Saki Kinney and Jörg Drechsler, as well as two of the original designers of the SIPP Synthetic data, Martha Stinson and Gary Benedetto, incorporated significant work on synthetic data into their doctoral theses. That, of course, expanded the knowledge base. But equally important to the task at hand, they stayed at the statistical agencies (or, in Saki's case, at the National Institute of Statistical Sciences, a long-time agency collaborator) where their research and subject-matter knowledge have nurtured a better understanding of the strengths, limitations and potential of synthetic data methods. We hope that others are inspired to do likewise by what the read in this *Issue*.

References not cited elsewhere in the Section

Little, Roderick J.A. (1993) "Statistical Analysis of Masked Data," Journal of Official Statistics, Vol. 9, No. 2, pp. 407-26.