Synthetic Longitudinal Business Data

International User Seminar

Tuesday, May 9, 2017

Washington DC

Location:
The National Academies of Sciences, Engineering, and Medicine
Keck Center
500 Fifth Street, NW
Washington, DC 20001
Keck 209

This version last updated 2017-05-01.
Online version at this unprintable long URL.

Registration at https://goo.gl/forms/hUjcQaKlkWEU18Yd2

Funding for the workshop is provided by the National Science Foundation (CNS-1012593, SES-1131848) and the Alfred P. Sloan Foundation. The organizers thank the National Academies' Committee on National Statistics for hosting the seminar.

Goals and Methods of the Seminar

The main purpose of the US Census Bureau's Synthetic LBD is to facilitate researcher access to establishment microdata in a way that preserves the confidentiality of the underlying entities' data. Establishment and firm microdata pose many challenges in this dimension, as they are sparse and often unique. It is easy to think of firms or establishments that are dominant in a specific industry or geographic location to such degree that their identification would be trivial if their data were 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, or some other limited access. These restrictions on data access reduce the research output by increasing the cost to researchers of accessing the data.

In 2010, the US Census Bureau made available the first analytically valid synthetic establishment microdata, the Synthetic Longitudinal Business Database (SynLBD). 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 (Kinney et al. 2011). By making the disclosable *synthetic* microdata available through a remotely accessible data server, combined with a validation server, the SynLBD approach alleviates some of these restrictions. The approach is mutually beneficial to agency and researchers. Researchers can access public use servers at little or no cost within a few weeks of their initial application. Researchers validate their model-based inferences on the full confidential microdata. The statistical agency has an interest in improving future versions of the synthetic data along existing and new dimensions. They can do so by leveraging the diversity of the researchers' models and analyzing discrepancies. The Synthetic Data Server (SDS) at Cornell University provides the infrastructure to implement this approach for two different synthetic datasets, with funding from NSF and the Alfred P. Sloan Foundation.

In this seminar, we discuss with interested parties the conditions necessary to implement the SynLBD approach, with the goal of providing other statistical agencies 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 ideal end result is a series of country-specific datasets on establishments and/or firms available within the same computing environment. We discuss 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 seminar brings together academics working on cutting-edge methods for the protection of privacy in statistical databases, and researchers and implementers at statistical agencies that have started or are interested in starting a similar project.

Five sessions will touch on the full lifecycle of a SynLBD development and implementation, and will follow the same pattern. We will first discuss existing implementations and experiences, and will then as a group discuss issues as they pertain to the broader community. Emphasis should be on discussing open issues, specific solutions to specific problems. A designated scribe for each session will take notes. Post-workshop, the preliminary summaries will be finalized by the scribes and the organizers, published as a proceedings, and made available to all.

Some additional readings: (Vilhuber and Abowd 2016; Vilhuber, Abowd, and Reiter 2016; Kinney et al. 2011; Drechsler and Vilhuber 2014)

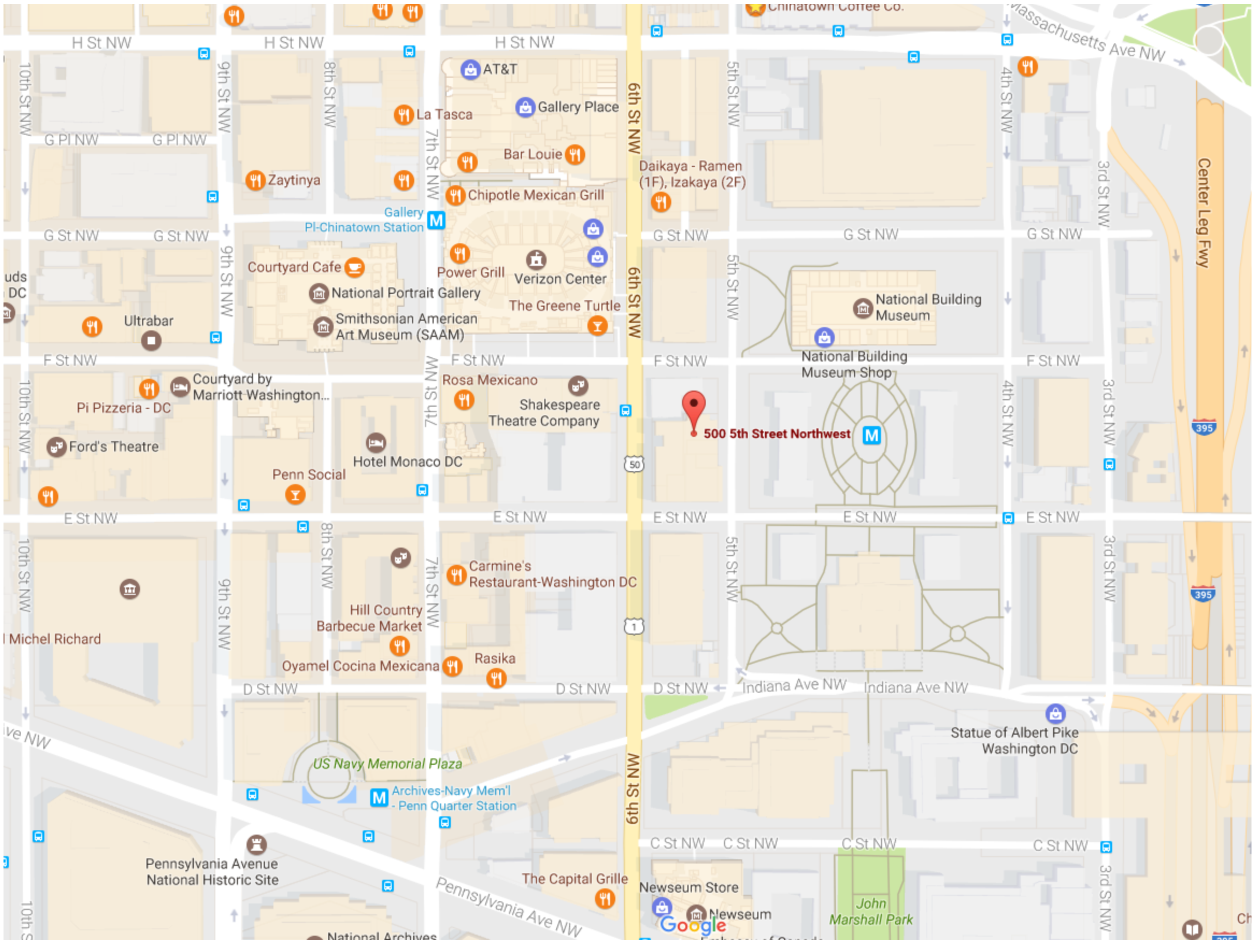
Agenda

Start	*Duration*	*Topic*
8:00	0:30	Coffee
8:30	0:15	**Welcome** messages
8:45	0:45	**Overview**: Current SynLBD methodology - scope, application, pre conditions
9:30	0:45	**Inputs**: What issues to consider when collecting input data (data cleaning, imputation, versioning, time-consistent coding)
10:15	0:20	Coffee break
10:35	0:45	**Confidentiality**: How confidential is synthetic LBD? What are your data provider's concerns regarding confidentiality?
11:20	0:45	**Validation:** Considerations for validation proposals, running your own validation service
12:05	1:10	Lunch (for purchase in the NAS cafeteria)
13:15	0:45	**Future**: what are next steps for SynLBD? What is needed? (additional variables, cross-national comparison)
14:00		**Conference ends**

Note: Full breakfast is available for purchase at the NAS cafeteria starting at 7AM.

Funding: NSF CNS-1012593, SES-1131848, Sloan Foundation.

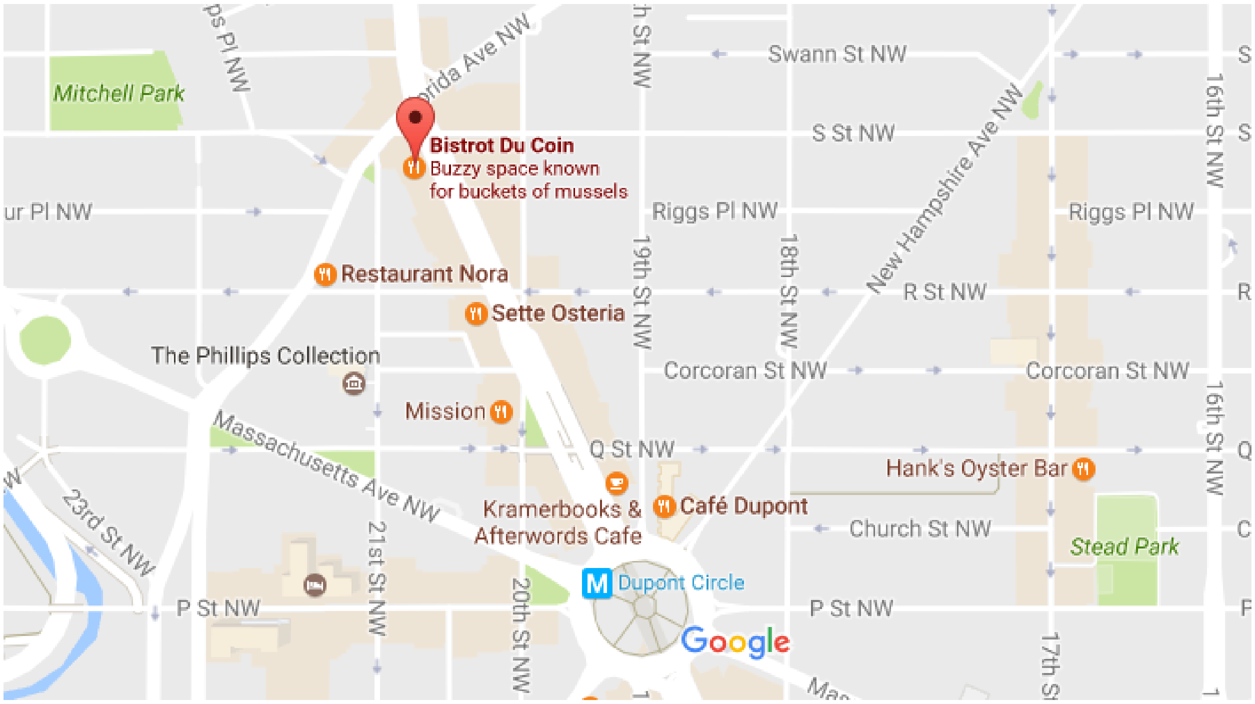
Getting there



Pre-workshop Dinner

Bistrot du Coin, 7:00 p.m. Monday, May 8

1738 Connecticut Avenue, NW - Washington, DC 20009



References

Drechsler, Jörg, and Lars Vilhuber. 2014. “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 (14-13). doi:10.3233/SJI-140812.

Kinney, Satkartar K., Jerome P. Reiter, Arnold P. Reznek, Javier Miranda, Ron S. Jarmin, and John M. Abowd. 2011. “Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database.” *International Statistical Review = Revue Internationale de Statistique* 79 (3). Blackwell Publishing Ltd: 362–84. doi:10.1111/j.1751-5823.2011.00153.x .

Little, Roderick J. A. 1993. “Statistical Analysis of Masked Data.” *Journal of Official Statistics* 9 (2): 407–26. http://www.jos.nu/articles/abstract.asp?article=92407 .

Rubin, Donald B. 1993. “Discussion: Statistical Disclosure Limitation.” *Journal of Official Statistics* 9 (2): 461–68. http://www.jos.nu/Articles/abstract.asp?article=92461 .

Vilhuber, Lars, and John M. Abowd. 2016. “Presentation: SOLE 2016: Usage and Outcomes of the Synthetic Data Server.” presented at the SOLE 2016. http://hdl.handle.net/1813/43883.

Vilhuber, Lars, John M. Abowd, and Jerome P. Reiter. 2016. “Synthetic Establishment Microdata around the World.” *Statistical Journal of the International Association for Official Statistics* 32 (1): 65–68. doi:10.3233/SJI-160964.