Usage and outcomes of the Synthetic Data Server

Lars Vilhuber[1] John Abowd[1]

[1]Labor Dynamics Institute, ILR, Cornell University, United States

May 2016

The Data

Synthetic Data

"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"

Synthetic Data

"...all variables are synthesized, or modeled, in a way that changes the record of each individual in a manner designed to preserve the underlying covariate relationships between the variables..."

SSB webpage

What type of synthetic data are they not?

These are not...

What type of synthetic data are they not?

These are not...

1. Univariate synthetic data ("test files") (used at IAB Germany, Statistics Canada)
2. Custom-generated synthetic data per project (SYLLS) [Nowok et al., 2016]
3. Differentially-private

Datasets

SIPP Synthetic Beta (SSB)

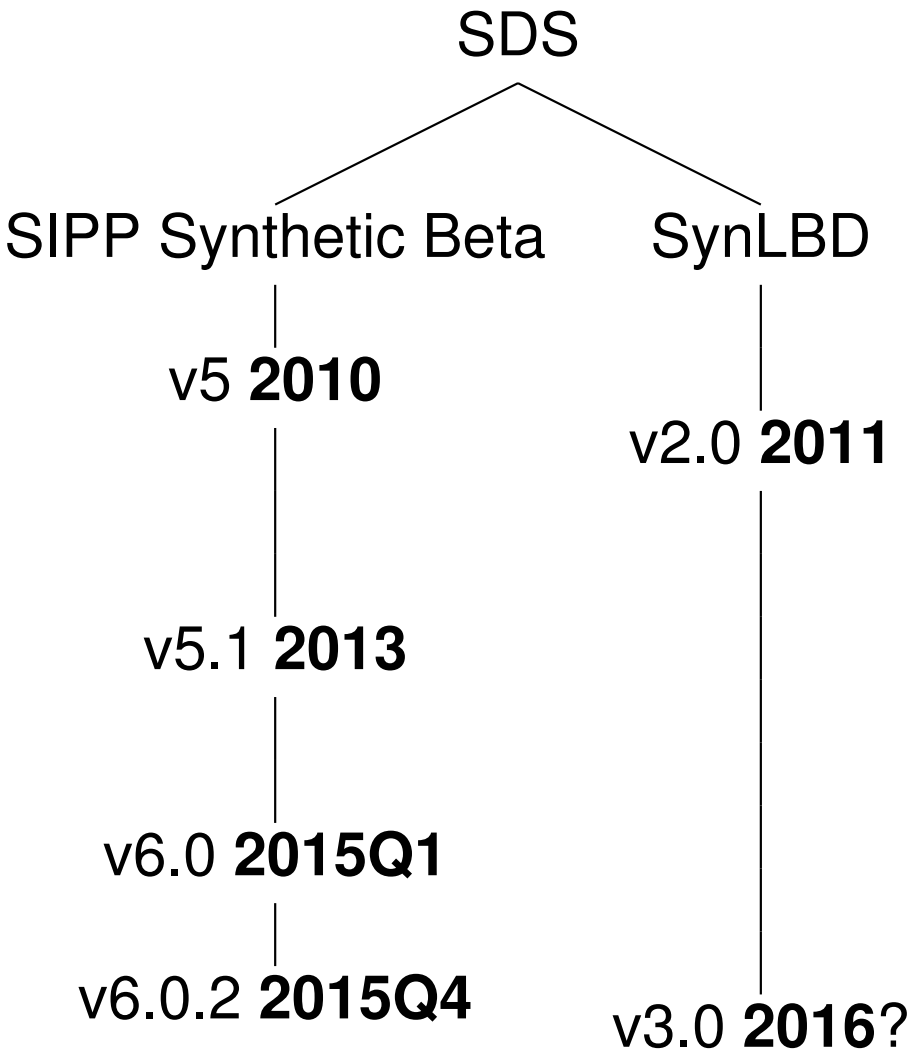
- provide access to linked data that are usually not publicly available
  - all variables (except two) are synthesized
  - gender and a link to the first reported marital partner are the exception.
  - estimate the joint distribution of all the variables in the data and taking random draws from this modeled distribution.
- The goal of the SSB is to produce results that are *qualitatively* the same as results from the Completed Gold Standard Files.
- Benedetto et al. [2013]

Datasets

Synthetic Longitudinal Business Database (LBD) (SynLBD)

- goal: provide users with access to a longitudinal business data product without disclosing confidential information.
- based on LBD: establishments' employment and payroll, establishments' birth and death years, and multi-unit status, conditional on industrial classification.
- Miranda and Jarmin [2002], Kinney et al. [2011]

History of datasets



The Audience

The Audience

Who's using this?

1. The datasets are made available to interested researchers in a controlled environment, prior to a more generalized release.
2. Academic researchers, worldwide.

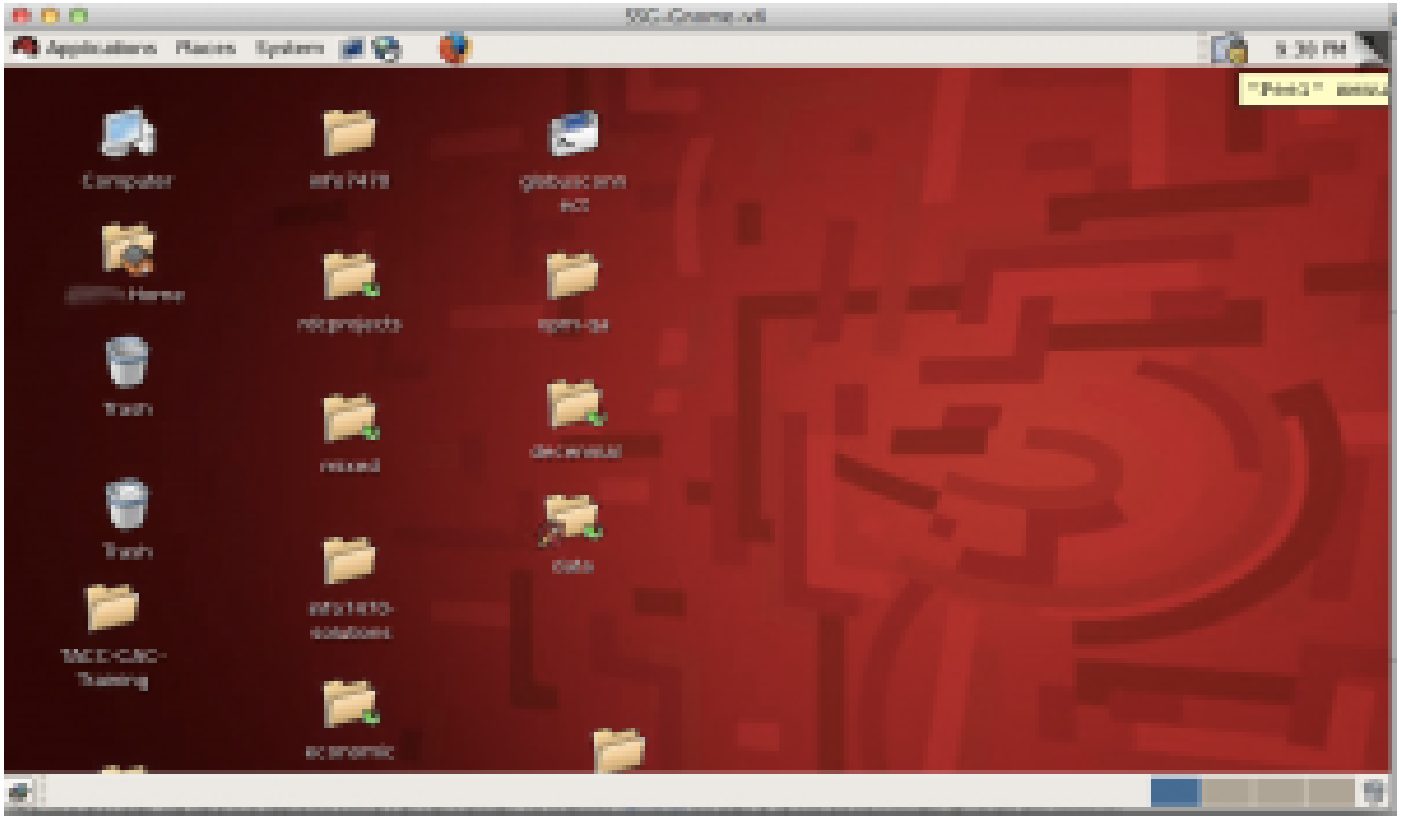
The Server

What is it?

Synthetic Data Server (SDS)

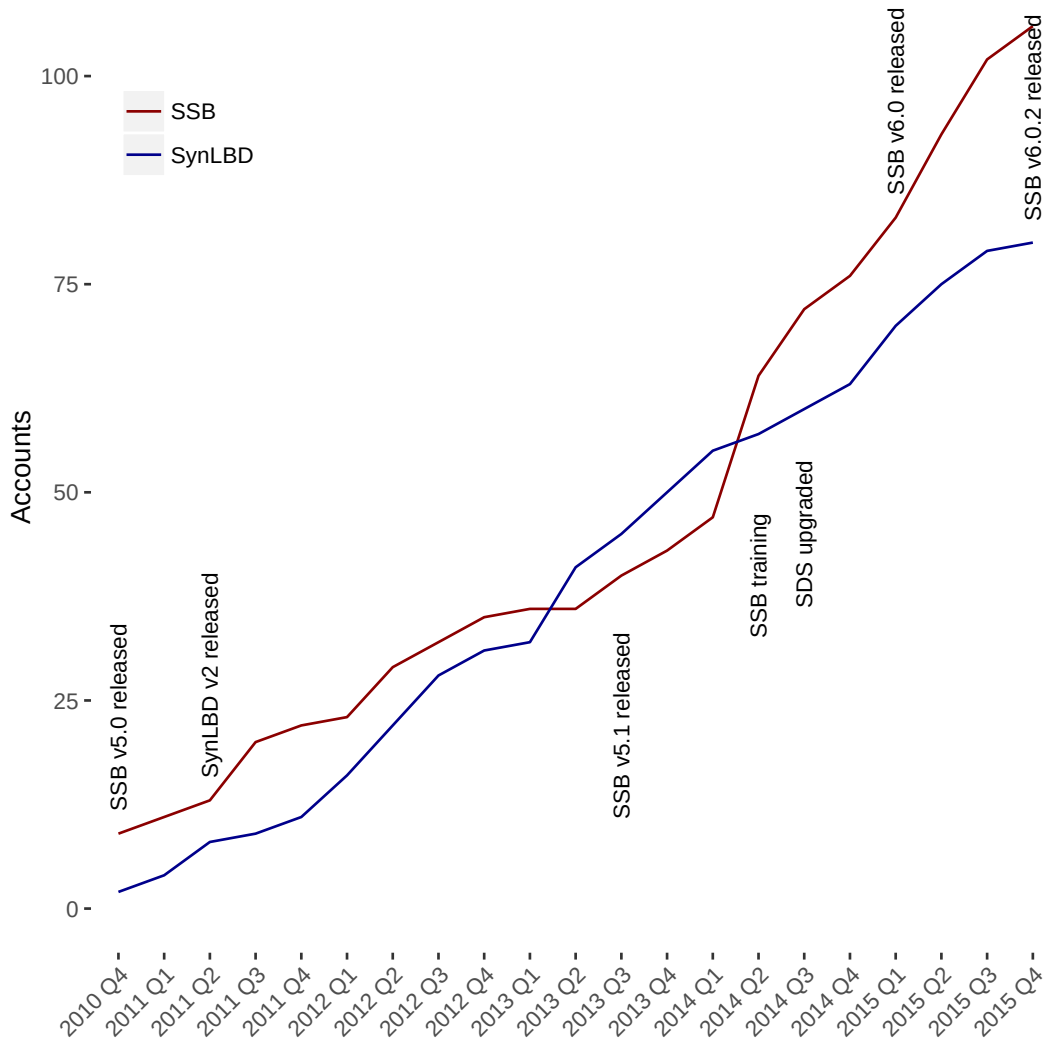
1. The Synthetic Data Server (SDS) at Cornell University was set up to provide early access to new synthetic data products (by the U.S. Census Bureau, others).
2. Remote graphical desktop, statistical software, emulates Census Bureau environment to a large extent

What's it look like?



Usage

6 years, 5 (versions of) synthetic datasets, over 180 users



More information

www.vrdc.cornell.edu/sds

Access

Access is fast

Simple access requests

- Access requests are sent to data custodians (a centralized application form is under development)
- Access requests are only reviewed for feasibility, but are not otherwise restricted.
- Once access is verified, the server provider (Cornell University) sets up accounts on the system
- Typical turnaround time is 1-10 days

Validation

- No restrictions on type of model to be estimated
- However, validated results must pass disclosure-avoidance analysis $\rightarrow$ some limitation (quantity, count restrictions)
- requires that users provide
  - all programs and auxiliary input files,
  - documentation of the results similar to a disclosure review request at Federal Statistical Research Data Center (FSRDC),
  - all programs run error-free (replicability requirement).

A few restrictions

Server access

- In order to prevent users from removing datasets from the server, requests for removal are *moderated*, but **not** censored.
- To ensure replicability/validation, upload requests for auxiliary data are moderated

A few restrictions

Server access

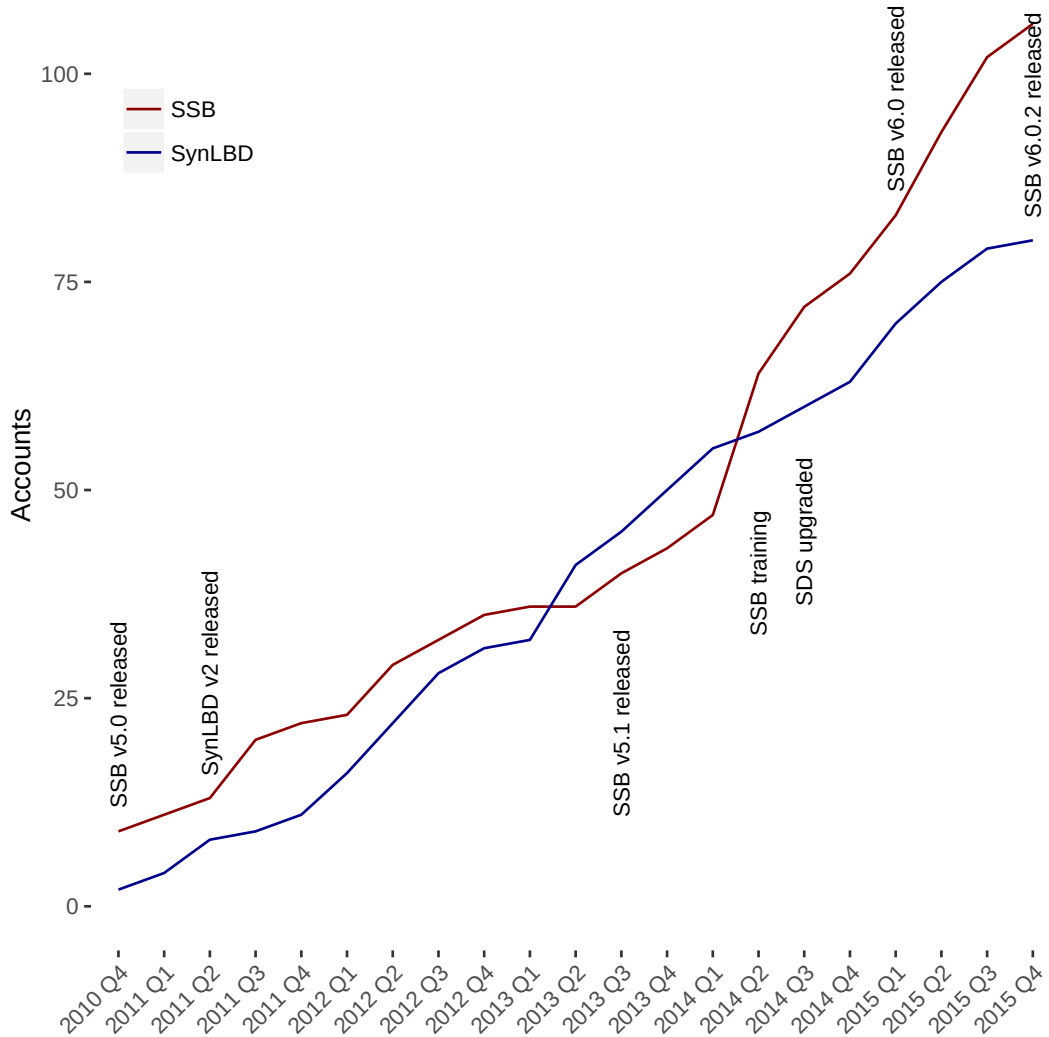
- In order to prevent users from removing datasets from the server, requests for removal are *moderated*, but **not** censored.
- To ensure replicability/validation, upload requests for auxiliary data are moderated

Server access

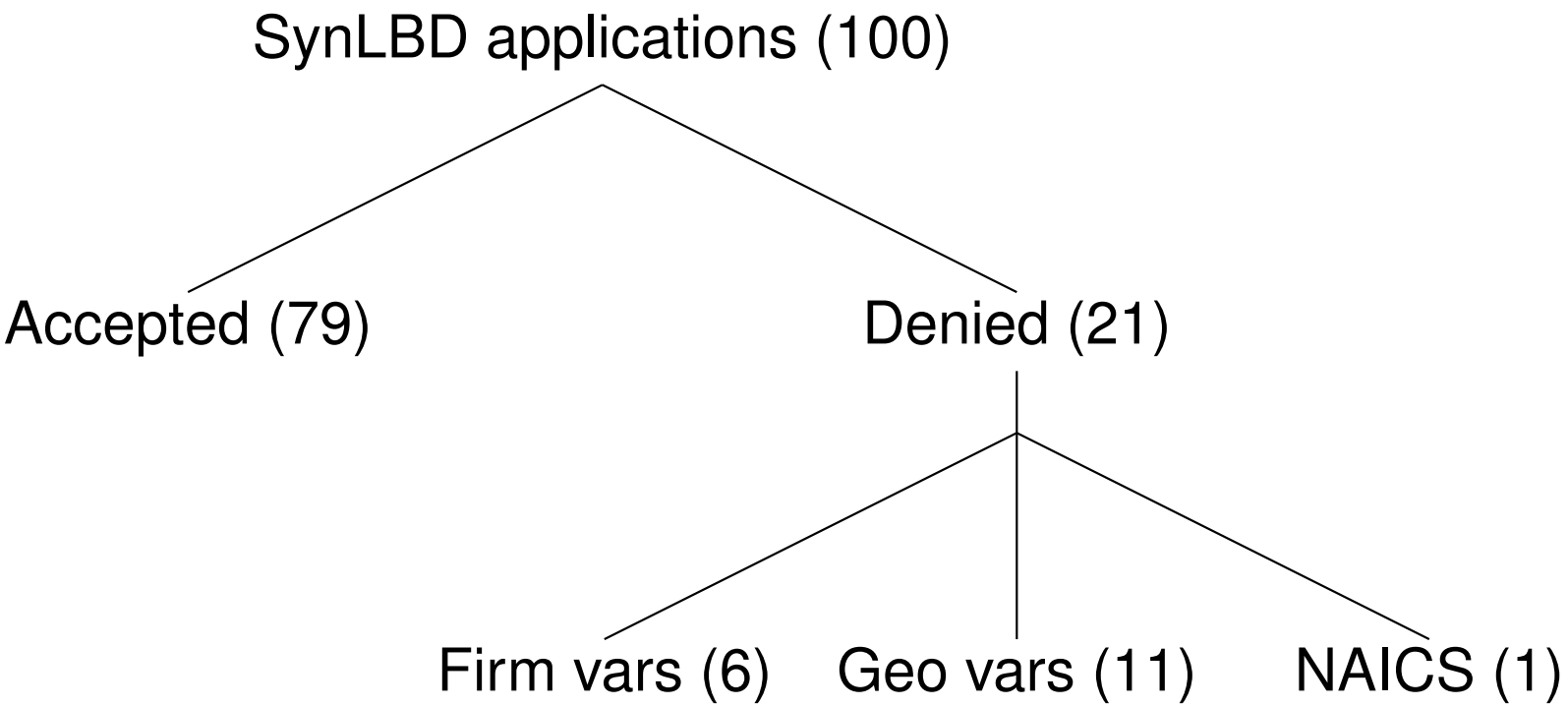
- software is limited to **SAS, Stata**.
- R, Matlab, Python may be available upon special request and upon coordination with data custodians.

Outcomes

Accounts created



Not all applications get accepted



Key feature: Feedback loop

User feedback incorporated into each version

SSB

- Variables
- Structure

SynLBD

- NAICS
- firm-structure
- geography

$\rightarrow$ V3.0

Validation

Validation

- No restrictions on type of model to be estimated
- However, validated results must pass disclosure-avoidance analysis $\rightarrow$ some limitation (quantity, count restrictions)
- requires that users provide
  - all programs and auxiliary input files,
  - documentation of the results similar to a disclosure review request at FSRDC,
  - all programs run error-free (replicability requirement).

Validation

SynLBD

As of 2015-08-10: 5 out of 79 projects have requested validation

SSB

As of yesterday: about 7 or 8 out of about 100 have requested validation

How well does validation work

Bertrand et al. [2015]

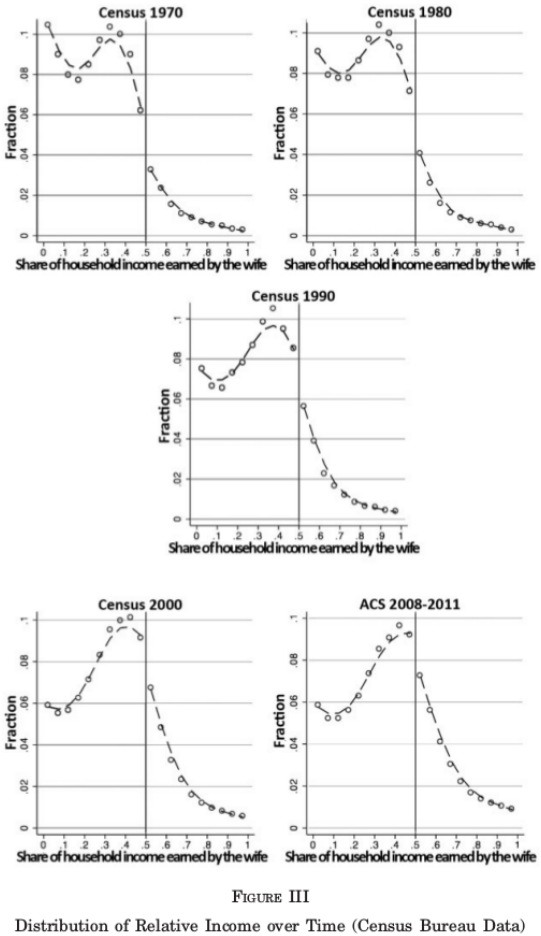
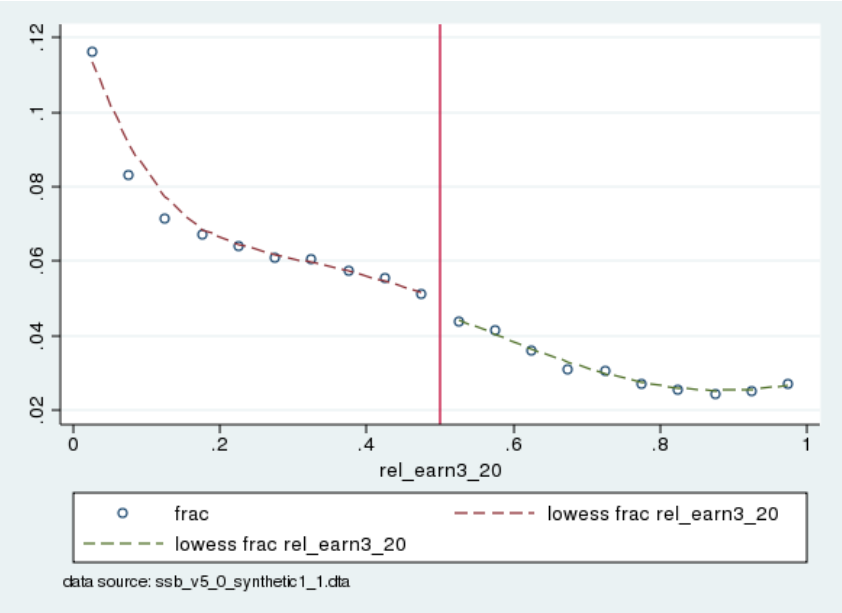


FIGURE III
Distribution of Relative Income over Time (Census Bureau Data)

There is a distinct break in the distribution of couples when the wife's income surpassed 50% (their Figure 3)

How well does validation work

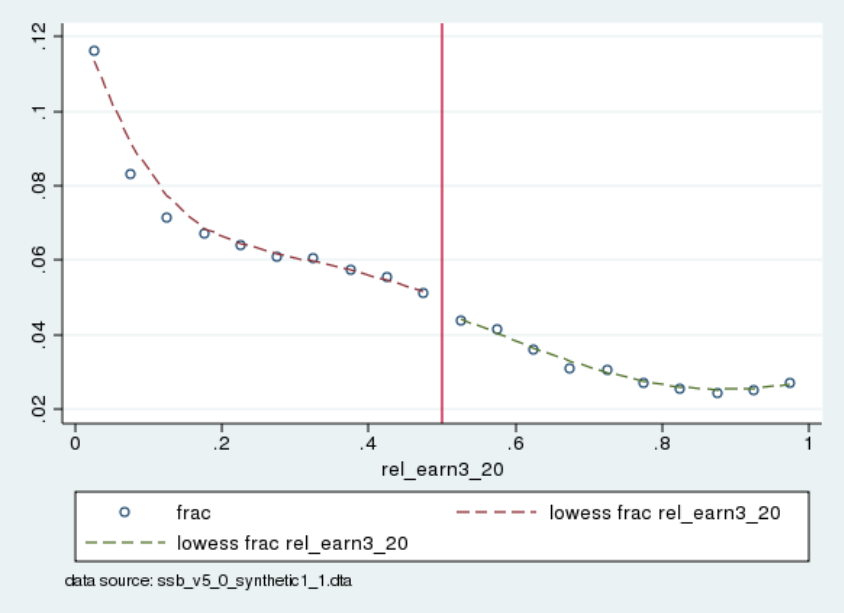
Bertrand et al on SSB



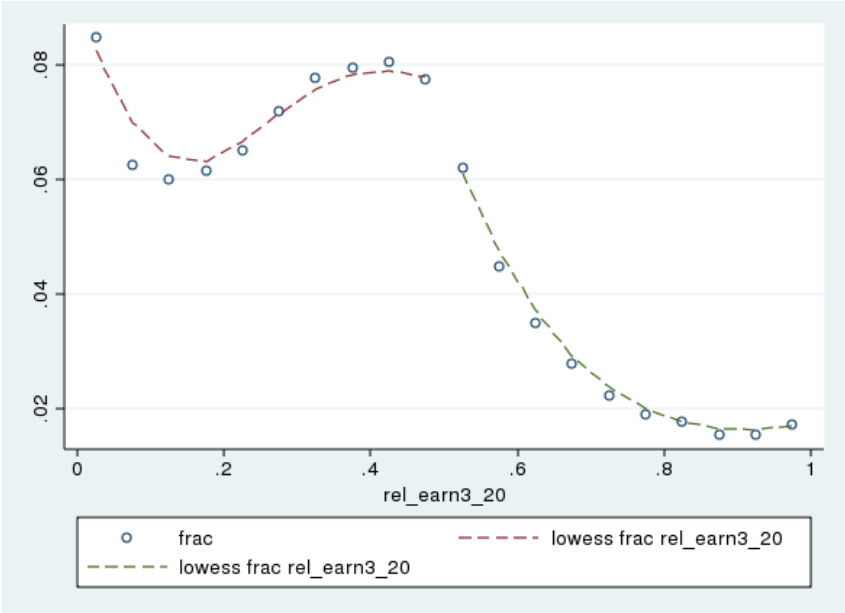
No such break in the synthetic data

How well does validation work

Bertrand et al on SSB



Bertrand et al. [2015]:



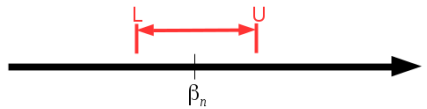
How well does validation work

General approach: *interval overlap measure* J_k

[Karr et al., 2006]

Consider the overlap of **confidence intervals** for variable *n*

- (L, U) for β_n (from the confidential data)

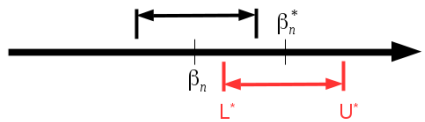


How well does validation work

General approach: *interval overlap measure* J_k

[Karr et al., 2006]
Consider the overlap of **confidence intervals** for variable *n*

- (L, U) for β_n (from the confidential data)
- (L^*, U^*) for β_n^* (from synthetic data)

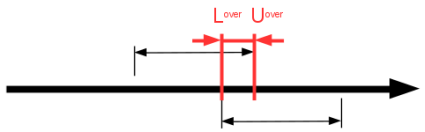


How well does validation work

General approach: *interval overlap measure* J_k

[Karr et al., 2006]
Consider the overlap of **confidence intervals** for variable *n*

- (L, U) for β_n (from the confidential data)
- (L^*, U^*) for β_n^* (from synthetic data)
- Let $L^{over} = \max(L, L^*)$ and $U^{over} = \min(U, U^*)$.



How well does validation work

Then the overlap in confidence intervals is

$$J_k^* = \frac{1}{2}\left[\frac{U^{over} - L^{over}}{U - L} + \frac{U^{over} - L^{over}}{U^* - L^*}\right]$$

How well does validation work

Initial results from SynLBD

	Mean	Median	75th	95th	Max	PctGrtThan0
All models	0.206	0	0.504	0.791	0.995	
User 1	0.101	0	0	0.726		19.8
User 2	0.212	0	0.507	0.791		38.0

How well does validation work

Initial results from SSB

	Mean	Median	75%	95%	Max	PctGrtThan0
1	0.49	0.54	0.79	0.91	0.98	82.38
2	0.39	0.52	0.56	0.71	0.94	73.20

How well does validation work

Downside

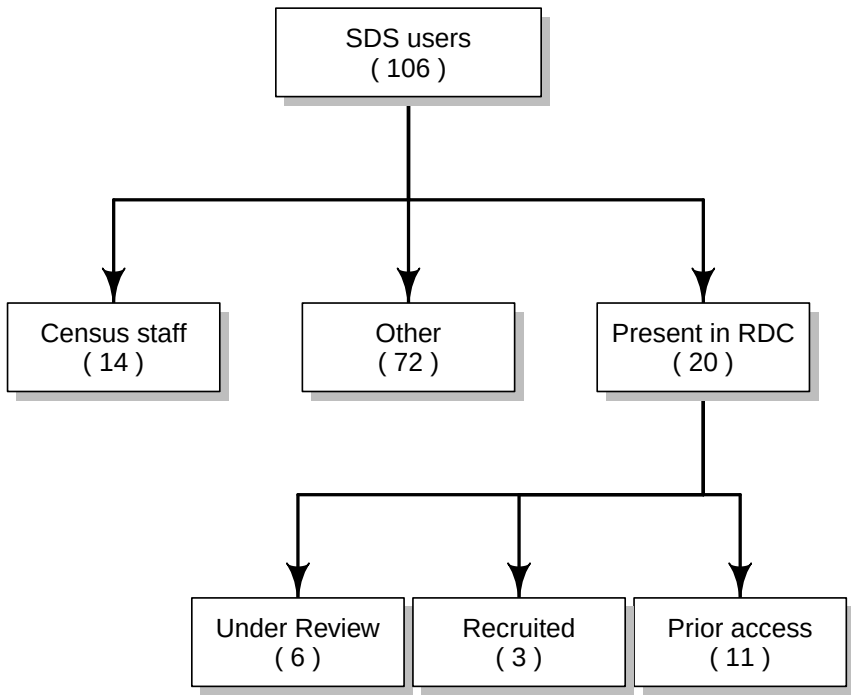
- Cannot adapt your model to the data
- Fundamental: will not work for non-congenial designs (f.i. regression discontinuity)

Upside

- Cannot adapt your model to the data
- Rapid turnaround (about 1 week) to get result from confidential answer

Outcomes other than validation

Figure: Connection between Census RDC usage and Synthetic Data Server



SDS and FSRDC

Other outcomes/interactions

- at least some of the RDC projects were direct continuation of SDS projects(private conversations)

SDS and FSRDC

Other outcomes/interactions

- at least some of the RDC projects were direct continuation of SDS projects(private conversations)
- average delay (project start (SDS), project start (RDC)) : 400 days.

SDS and FSRDC

Other outcomes/interactions

- at least some of the RDC projects were direct continuation of SDS projects(private conversations)
- average delay (project start (SDS), project start (RDC)) : 400 days.
- (reminder: average turnaround for validation = 7 days...)

Next steps

Expansion

SynLBD

- German SynLBD [Drechsler and Vilhuber, 2014]
- Canadian SynLBD (about to start!)
- Brazilian SynLBD (awaiting data)
- interest from a few other quarters

$\rightarrow$ cross-national analysis on establishment-level data

Iterative synthetic data

Differentially private data generation

Abowd and Schmutte [2015]: interactively build optimal synthetic data. Conditions: users that issue “queries” to the system, plus data that is of interest to users.

Improvements and training: SSB

Planning underway for SSB v7

- Survey sent to current users of SSB, requesting feedback on where to make improvements to SSB, first results coming in
- Interested users should contact me! Feedback from any interested user!

Training for interested users

- As part of NCRN-Michigan mission, provided by Census staff, with support from NCRN-Cornell (me)
- See ncrn.info for announcement and request for applications!

Stay tuned!

www.vrdc.cornell.edu/sds

Thank you!

```
$Id: Presentation-subdoc.tex 6130 2016-05-06 14:10:22Z lv39 $
```

Funding

NSF Grants #1042181 and #0941226, Alfred P. Sloan Foundation.

Bibliography

J. M. Abowd and I. Schmutte. Economic analysis and statistical disclosure limitation. *Brookings Papers on Economic Activity*, Fall 2015, 2015. ISSN 00072303. URL http://www.brookings.edu/about/projects/bpea/papers/2015/economic-analysis-statistical-disclosure-limitation.

G. Benedetto, M. Stinson, and J. M. Abowd. The creation and use of the sipp synthetic beta. Technical report, US Census Bureau, 2013. URL http://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/SSBdescribe_nontechnical.pdf.

M. Bertrand, E. Kamenica, and J. Pan. Gender identity and relative income within households. *The Quarterly Journal of Economics*, 130(2), 2015. doi: 10.1093/qje/qjv001. URL http://qje.oxfordjournals.org/content/early/2015/04/11/qje.qjv001.abstract.

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. doi: 10.3233/SJI-140812. URL http://content.iospress.com/articles/statistical-journal-of-the-iaos/sji00812.

A. F. Karr, C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil. A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 60(3):1–9, 2006. doi: 10.1198/000313006X124640.

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): 362–384, 2011. ISSN 1751-5823. doi: 10.1111/j.1751-5823.2011.00153.x. URL http://dx.doi.org/10.1111/j.1751-5823.2011.00153.x.

J. Miranda and R. Jarmin. The Longitudinal Business Database. Discussion Paper CES-WP-02-17, U.S. Census Bureau, Center for Economic Studies, 2002. URL http://ideas.repec.org/p/cen/wpaper/02-17.html.

B. Nowok, G. M. Raab, J. Snoke, and C. Dibben. *synthpop: Generating Synthetic Versions of Sensitive Microdata for Statistical Disclosure Control*, 2016. URL https://CRAN.R-project.org/package=synthpop. R package version 1.2-1.