Usage and outcomes of the Synthetic Data Server

Lars Vilhuber[1] John Abowd[1]

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

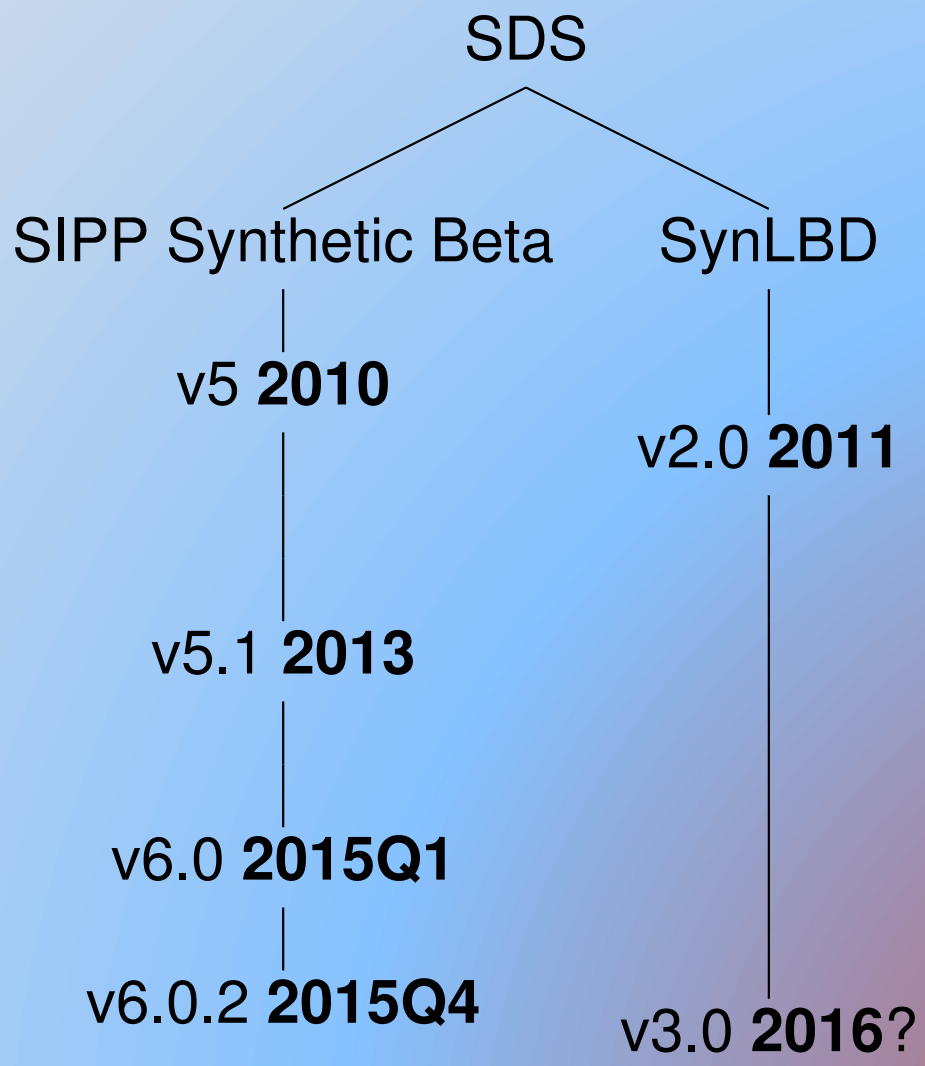
May 2016

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

History

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

History of datasets



© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

The Server

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

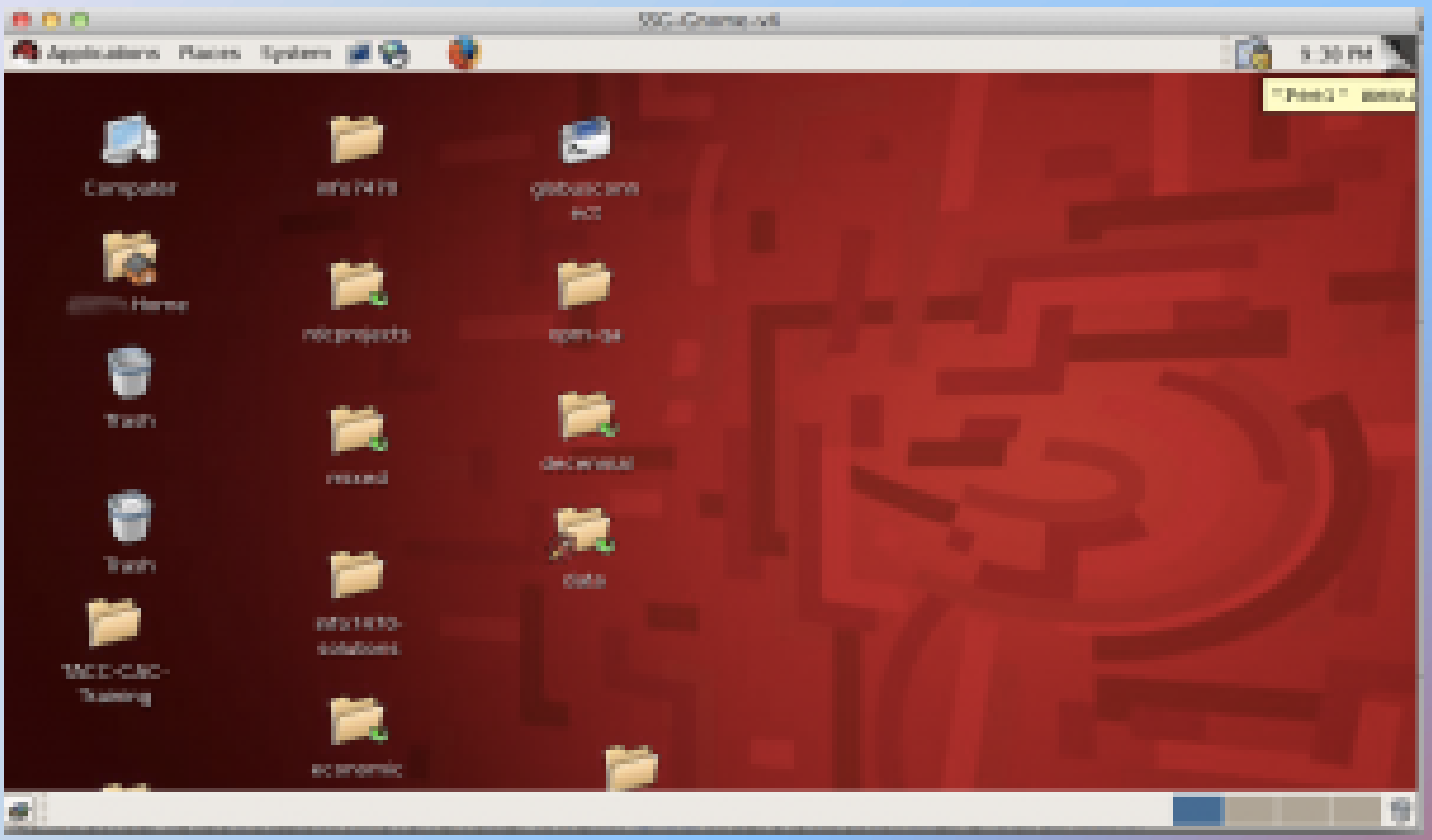
What is it?

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

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

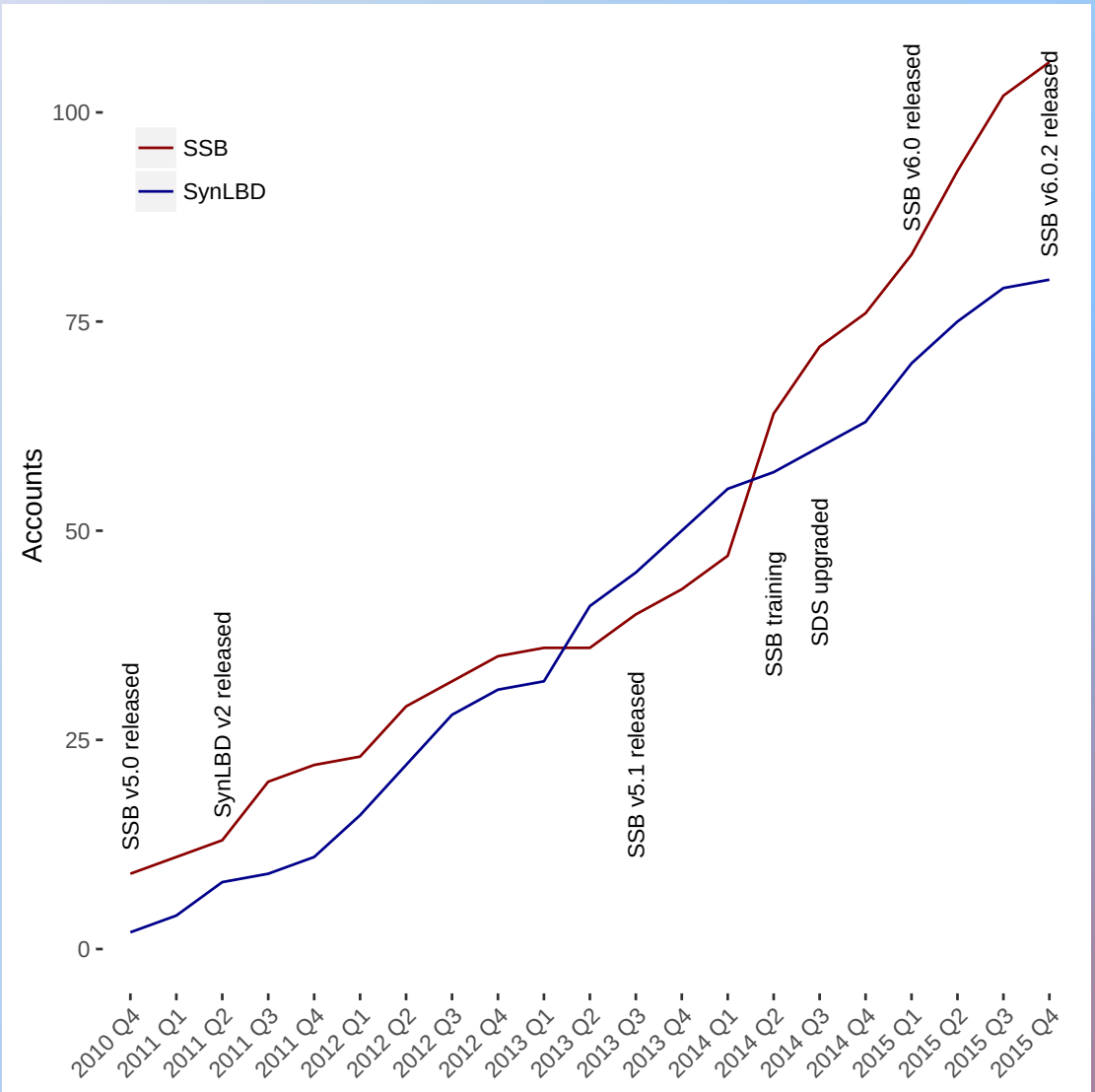
What's it look like?



© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

Usage

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



© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

More information

www.vrdc.cornell.edu/sds

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

Access

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

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

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

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).

© John M. Abowd and Lars Vilhuber 2016, all rights reserved

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

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

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

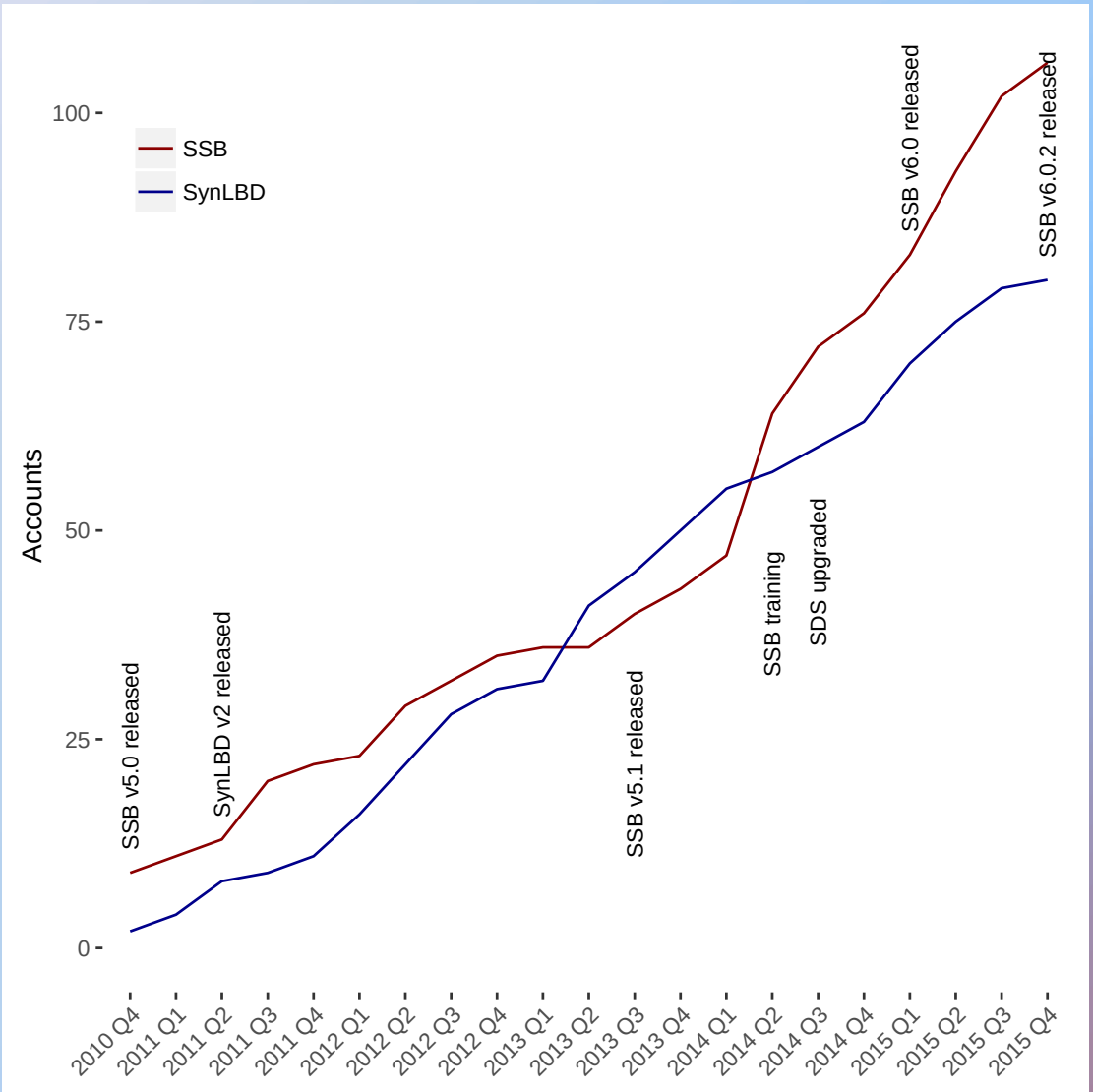
- software is limited to **SAS, Stata**.
- R, Matlab, Python may be available upon special request and upon coordination with data custodians.

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

Outcomes

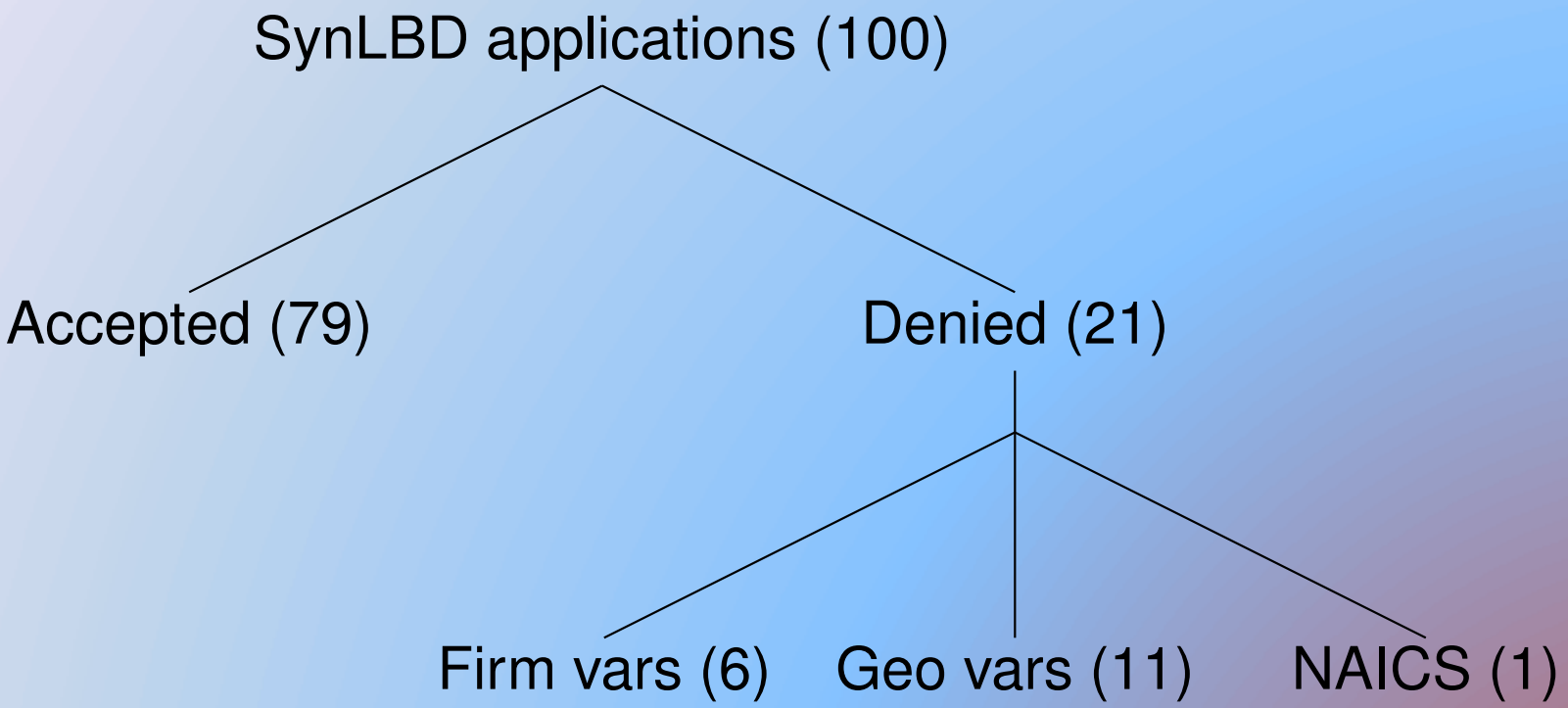
© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

Accounts created



© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

Not all applications get accepted



© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

Key feature: Feedback loop

User feedback incorporated into each version

SSB

- Variables
- Structure

SynLBD

- NAICS
- firm-structure
- geography

$\rightarrow$ V3.0

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

Validation

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

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).

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

Validation

SynLBD

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

SSB

As of yesterday: about 10 out of about 100 have requested validation

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

How well does validation work

?

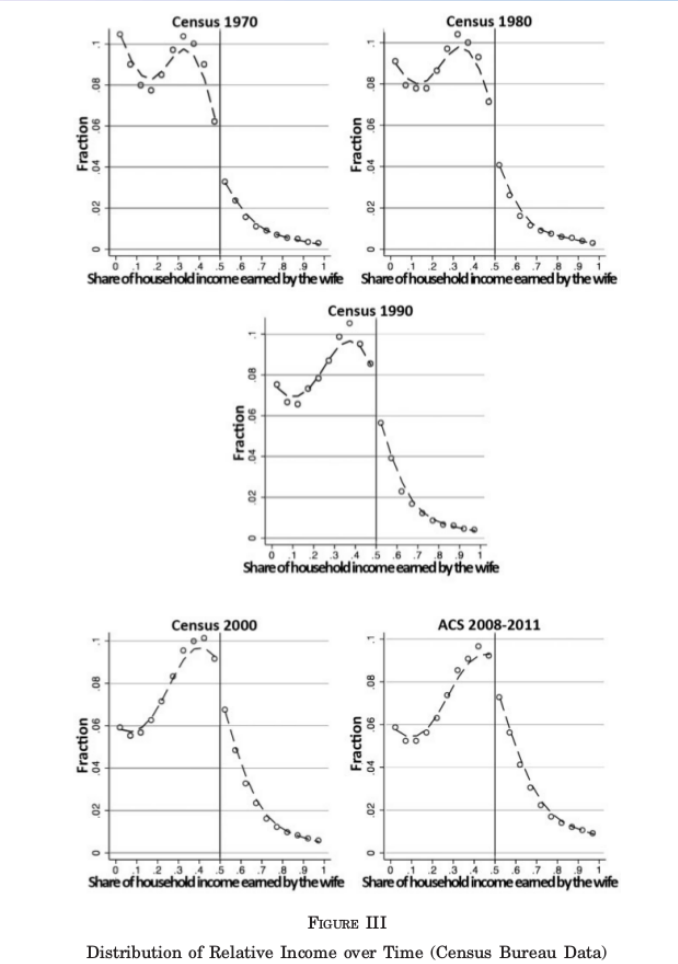


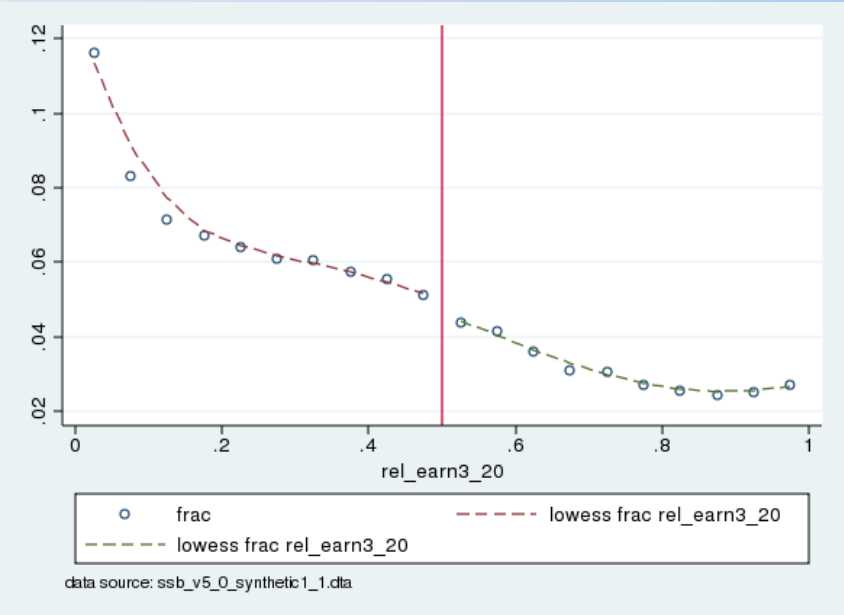
FIGURE III
Distribution of Relative Income over Time (Census Bureau Data)

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

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

How well does validation work

Bertrand et al on SSB

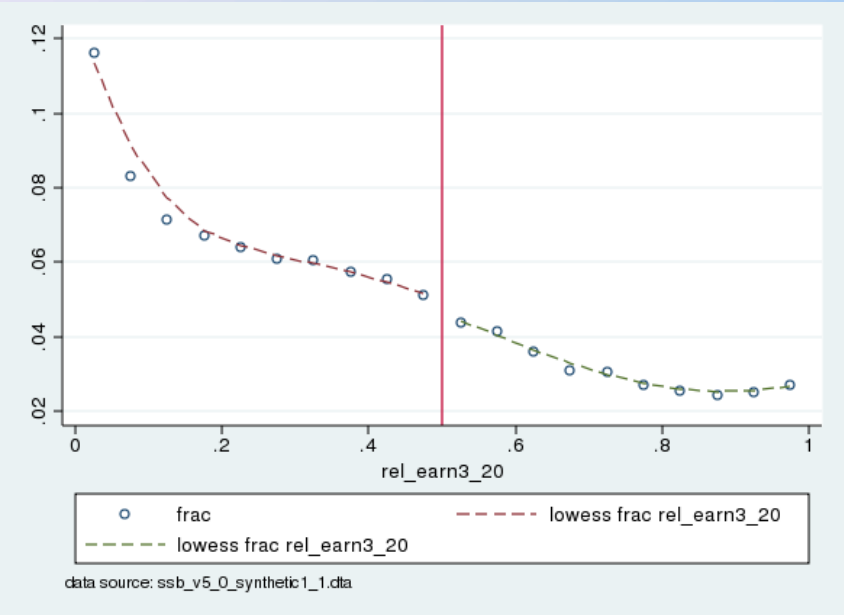


No such break in the synthetic data

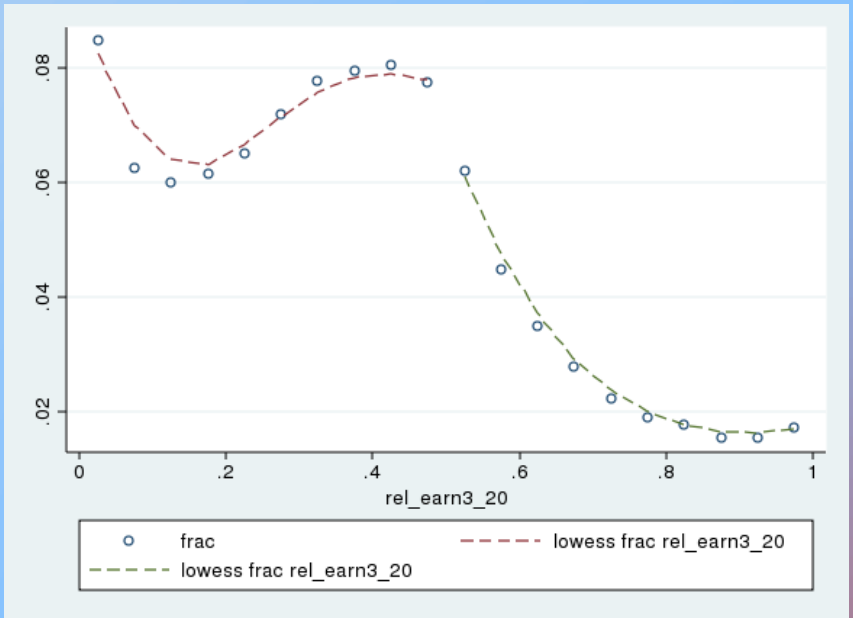
© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

How well does validation work

Bertrand et al on SSB



?:



© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

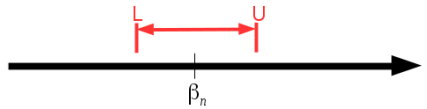
How well does validation work

General approach: *interval overlap measure* J_k

[**?**]

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

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



© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

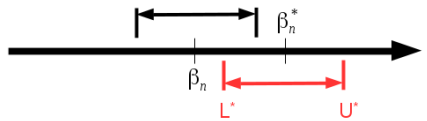
How well does validation work

General approach: *interval overlap measure* J_k

[**?**]

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)



© John M. Abowd and Lars Vilhuber 2016, all rights reserved

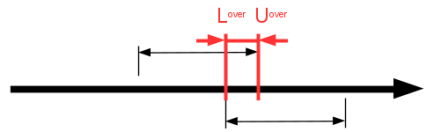
How well does validation work

General approach: *interval overlap measure* J_k

[**?**]

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 $U^{over} = \min(U, U^*)$.



© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

How well does validation work

Then the overlap in confidence intervals is

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

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

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

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

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

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

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)

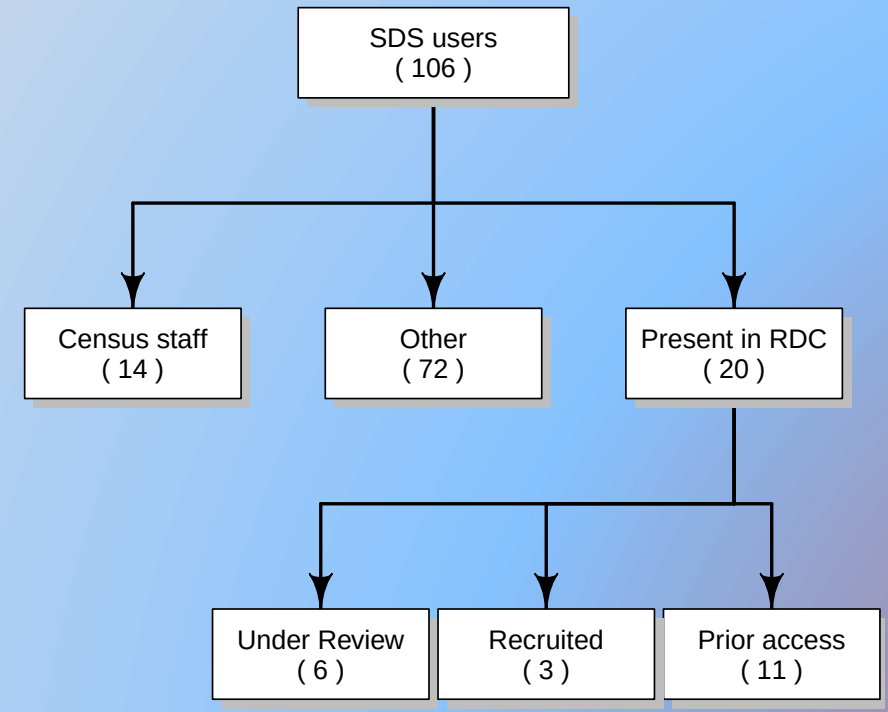
Upside

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

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

Outcomes other than validation

Figure: Connection between Census RDC usage and Synthetic Data Server



© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

SDS and FSRDC

Other outcomes/interactions

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

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

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.

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

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...)

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

Next steps

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

Expansion

SynLBD

- German SynLBD [**?**]
- Canadian SynLBD (about to start!)
- Brazilian SynLBD (awaiting data)
- interest from a few other quarters

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

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

Iterative synthetic data

Differentially private data generation

?: interactively build optimal synthetic data. Conditions: users that issue “queries” to the system, plus data that is of interest to users.

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

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
- See ncrn.info for announcement and request for applications!

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

Thank you!

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

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

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

Funding

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

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved

Bibliography

© John M. Abowd and Lars Vilhuber 2016,
all rights reserved