

Usage and outcomes of the Synthetic Data Server

Lars Vilhuber¹ John Abowd¹

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

May 2016

Excerpt

Datasets

SIPP Synthetic Beta (SSB)

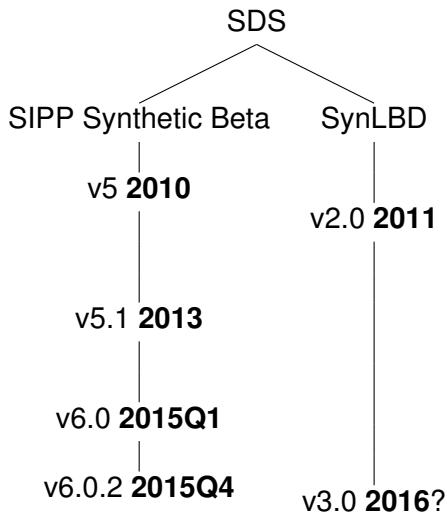
- ▶ provide access to linked data that are usually not publicly available
 - ▶ all variables (except two) are synthesized
 - ▶ gender and a link to the first reported marital partner are the exception.
 - ▶ estimate the joint distribution of all the variables in the data and taking random draws from 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.

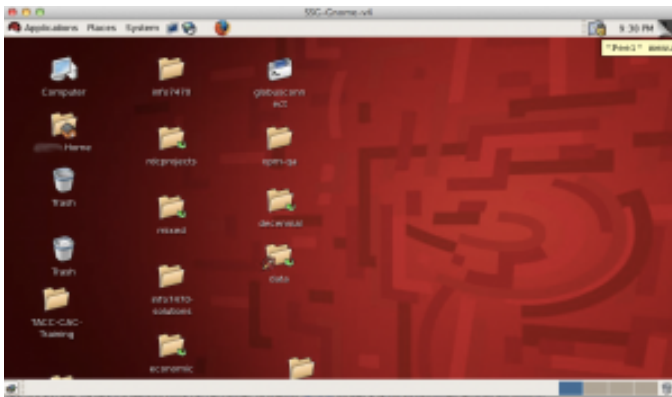
The Server

What is it?

Synthetic Data Server (SDS)

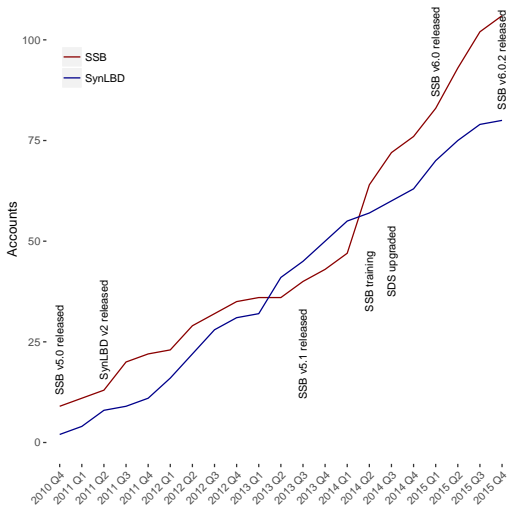
1. The Synthetic Data Server (SDS) at Cornell University was set up to provide early access to new synthetic data products (by the U.S. Census Bureau, others).
2. Remote graphical desktop, statistical software, emulates Census Bureau environment to a large extent

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 → 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

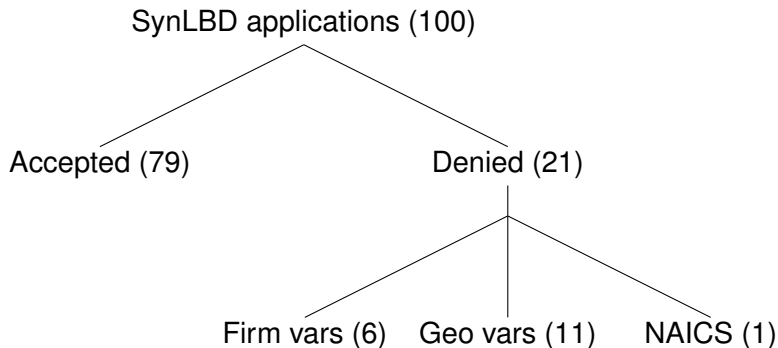
- ▶ In order to prevent users from removing datasets from the server, requests for removal are *moderated*, but **not** censored.
- ▶ To ensure replicability/validation, upload requests for auxiliary data are moderated

Server access

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

Outcomes

Not all applications get accepted



Key feature: Feedback loop

User feedback incorporated into each version

SSB

- ▶ Variables
- ▶ Structure

SynLBD

- ▶ NAICS
- ▶ firm-structure
- ▶ geography

→ V3.0

Validation

Validation

- ▶ No restrictions on type of model to be estimated
- ▶ However, validated results must pass disclosure-avoidance analysis → 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]

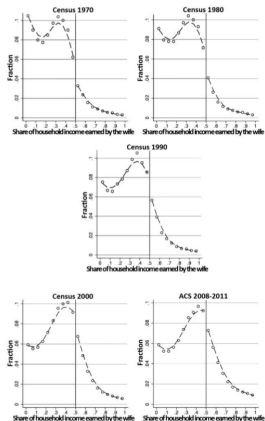


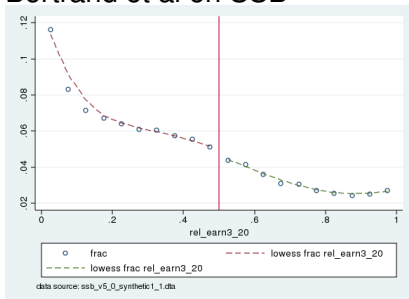
FIGURE III

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

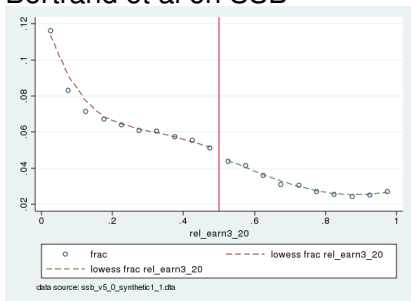
Bertrand et al on SSB



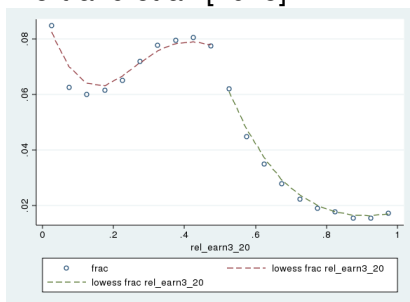
No such break in the synthetic data

How well does validation work

Bertrand et al on SSB



Bertrand et al. [2015]:



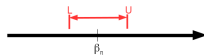
How well does validation work

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)



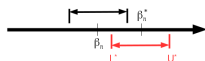
How well does validation work

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)



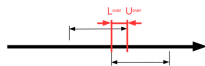
How well does validation work

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

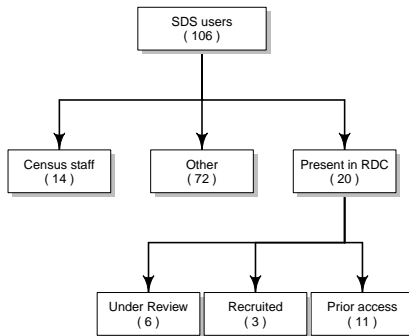
- ▶ Cannot adapt your model to the data
- ▶ Fundamental: will not work for non-congenial designs (f.i. regression discontinuity)

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...)