Usage and outcomes of the Synthetic Data Server

Lars Vilhuber\textsuperscript{1}    John Abowd\textsuperscript{1}

\textsuperscript{1}Labor Dynamics Institute, ILR, Cornell University, United States

May 2016
History
History of datasets

SDS

SIPP Synthetic Beta

v5 2010

v5.1 2013

v6.0 2015Q1

v6.0.2 2015Q4

SynLBD

v2.0 2011

v3.0 2016?
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
Accounts created

- SSB v5.0 released
- SynLBD v2 released
- SSB v5.1 released
- SSB training
- SSB v6.0 released
- SSB v6.0.2 released
- SDS upgraded

© John M. Abowd and Lars Vilhuber 2016, all rights reserved
Not all applications get accepted

SynLBD applications (100)

- Accepted (79)
- Denied (21)
  - Firm vars (6)
  - Geo vars (11)
  - NAICS (1)
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 10 out of about 100 have requested validation
How well does validation work?

![Graphs showing the distribution of income](image)

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

![](chart1.png)

?!

![](chart2.png)
How well does validation work

General approach: *interval overlap measure $J_k$*

[?] Consider the overlap of **confidence intervals** for variable $n$

- $(L, U)$ for $\beta_n$ (from the confidential data)
How well does validation work

General approach: *interval overlap measure* $J_k$

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

- $(L, U)$ for $\beta_n$ (from the confidential data)
- $(L^*, U^*)$ for $\beta_n^*$ (from synthetic data)
How well does validation work

General approach: *interval overlap measure* $J_k$

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

- $(L, U)$ for $\beta_n$ (from the confidential data)
- $(L^*, U^*)$ for $\beta^*_n$ (from synthetic data)
- Let $L^{\text{over}} = \max(L, L^*)$ and $U^{\text{over}} = \min(U, U^*)$. 
How well does validation work

Then the overlap in confidence intervals is

\[ J_k^* = \frac{1}{2} \left[ \frac{U^{\text{over}} - L^{\text{over}}}{U - L} + \frac{U^{\text{over}} - L^{\text{over}}}{U^* - L^*} \right] \]
How well does validation work

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>75th</th>
<th>95th</th>
<th>Max</th>
<th>PctGrtThan0</th>
</tr>
</thead>
<tbody>
<tr>
<td>All models</td>
<td>0.206</td>
<td>0</td>
<td>0.504</td>
<td>0.791</td>
<td>0.995</td>
<td></td>
</tr>
<tr>
<td>User 1</td>
<td>0.101</td>
<td>0</td>
<td>0</td>
<td>0.726</td>
<td>19.8</td>
<td></td>
</tr>
<tr>
<td>User 2</td>
<td>0.212</td>
<td>0</td>
<td>0.507</td>
<td>0.791</td>
<td></td>
<td>38.0</td>
</tr>
</tbody>
</table>
How well does validation work

Initial results from SSB

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>75%</th>
<th>95%</th>
<th>Max</th>
<th>PctGrtThan0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.49</td>
<td>0.54</td>
<td>0.79</td>
<td>0.91</td>
<td>0.98</td>
<td>82.38</td>
</tr>
<tr>
<td>2</td>
<td>0.39</td>
<td>0.52</td>
<td>0.56</td>
<td>0.71</td>
<td>0.94</td>
<td>73.20</td>
</tr>
</tbody>
</table>
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
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...)

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

SynLBD

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

→ cross-national analysis on establishment-level data
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.
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!
Thank you!
Funding
NSF Grants #1042181 and #0941226, Alfred P. Sloan Foundation.
Bibliography