Usage and outcomes of the Synthetic Data Server

Lars Vilhuber¹  John Abowd¹

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

May 2016
The 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”
“...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...

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
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

SDS

SIPP Synthetic Beta
- v5 2010
- v5.1 2013
- v6.0 2015Q1
- v6.0.2 2015Q4

SynLBD
- v2.0 2011
- v3.0 2016?
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
The Data  The Audience  The Server  Access  Outcomes  Validation  Next steps

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

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]:

Vilhuber, Abowd
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 $\beta_n$ (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 $\beta_n$ (from the confidential data )
- $(L^*, U^*)$ for $\beta_n^*$ (from synthetic data)
General approach: *interval overlap measure* $J_k$

[Karr et al., 2006]

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

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

Initial results from SynLBD

<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></td>
<td>19.8</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

- SDS users (106)
  - Census staff (14)
  - Other (72)
  - Present in RDC (20)
    - Under Review (6)
    - Recruited (3)
    - Prior access (11)
Other outcomes/interactions

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

→ 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!
Funding

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


