Utlity of two synthetic data sets mediated through a validation server: Experience with the Cornell Synthetic Data Server

Lars Vilhuber[1]

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

July 2019

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 various statistical agencies)
2. Custom-generated synthetic data per project SYLLS

   [Nowok et al., 2016]
3. Differentially-private

Datasets

SIPP Synthetic Beta (SSB)

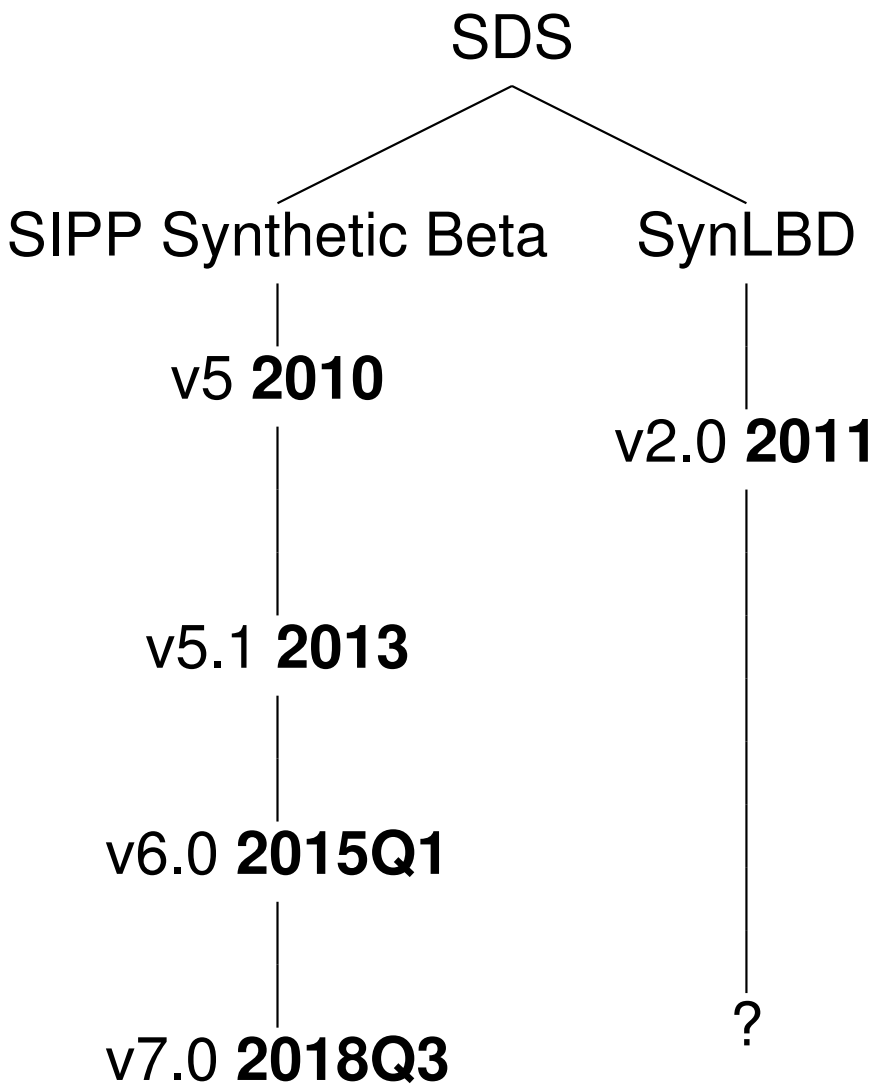
- provide access to linked data that are usually not publicly available
  - all variables (except two) are synthesized
  - gender and a link to the first reported marital partner are the exception.
  - estimate the joint distribution of all the variables in the data and taking random draws from this 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. [2018]
- Codebook: Reeder et al. [2018]

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]
- Codebook: Vilhuber [2013]

History of datasets



Similar efforts

- Drechsler and Reiter [2009] synthetic business microdata, released
- Alam, Dostie, Vilhuber [xxxx] Synthetic LEAP (Canada), synthetic business microdata, access through Canadian RDC network (no outcomes yet) [+ validation]
- Burman et al. [2018] administrative tax data with a synthetic public use file [+ validation], not released yet

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 (file system, software availability) to a large extent
3. Reflects 2010 technology

Why a dedicated server?

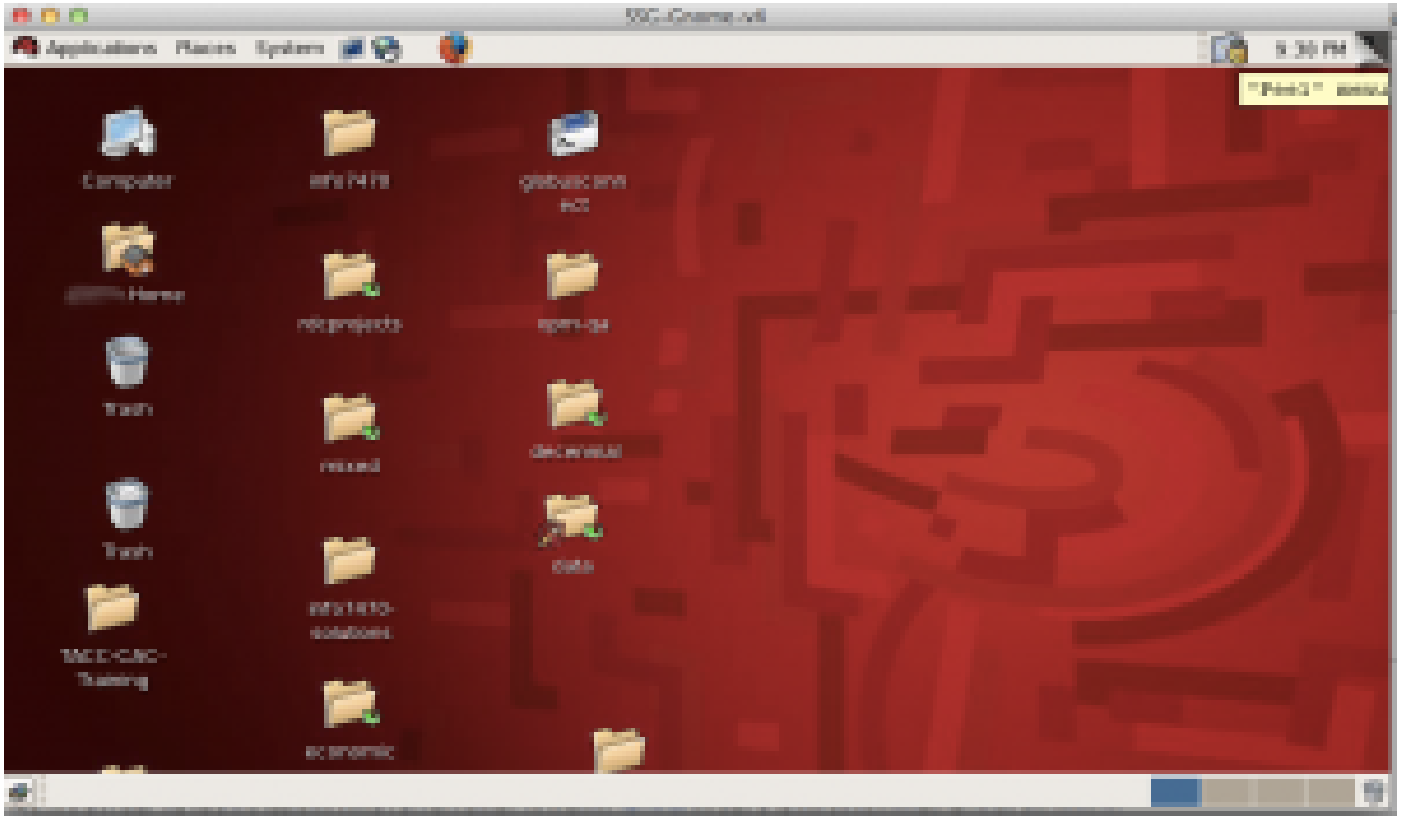
Reproducibility

- The SDS emulates the target compute environment closely
- Allows researchers to create code that can be re-run on the confidential data

Enforce non-redistribution

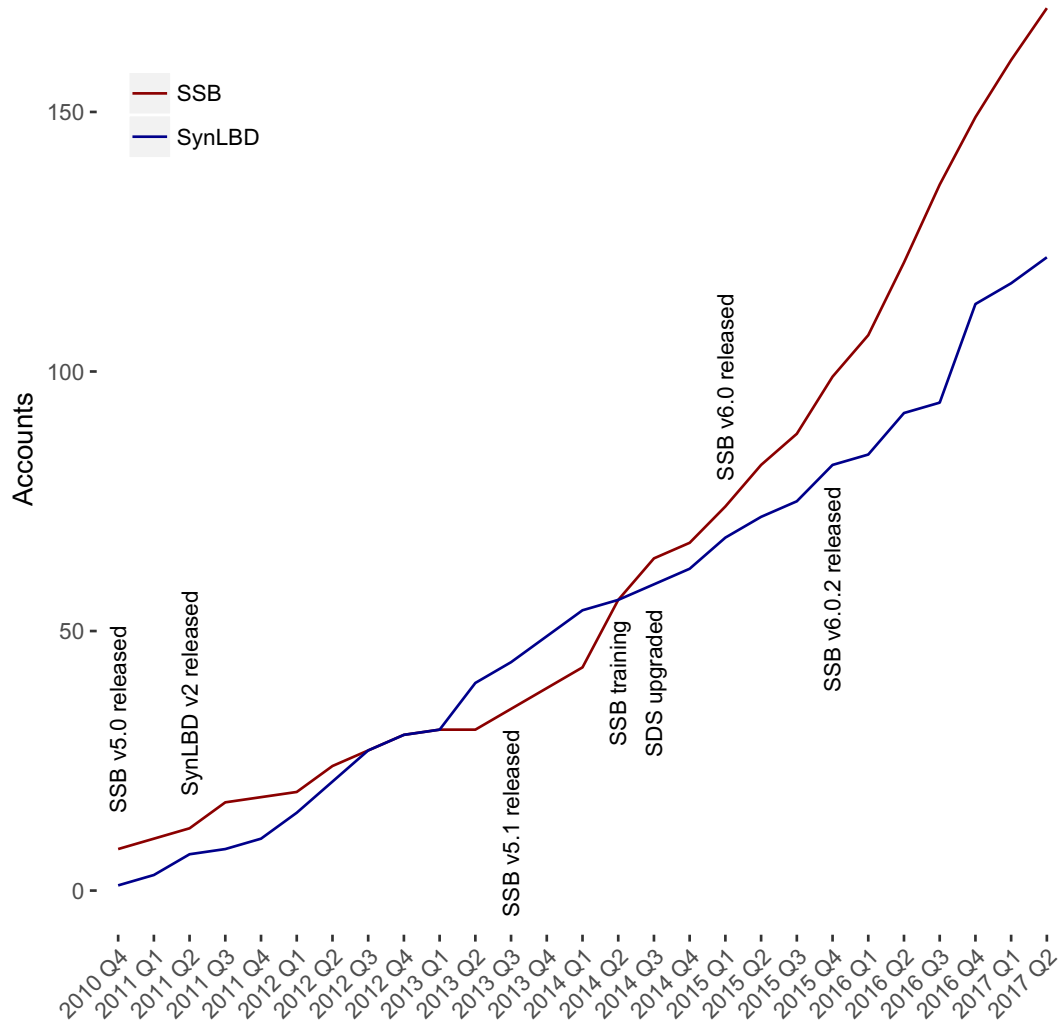
- No specific user license
- No guaranteed data quality - concerns about mis-representation of results obtained from synthetic data

What's it look like?



Usage

9 years, 6 (versions of) synthetic datasets, over 200 users

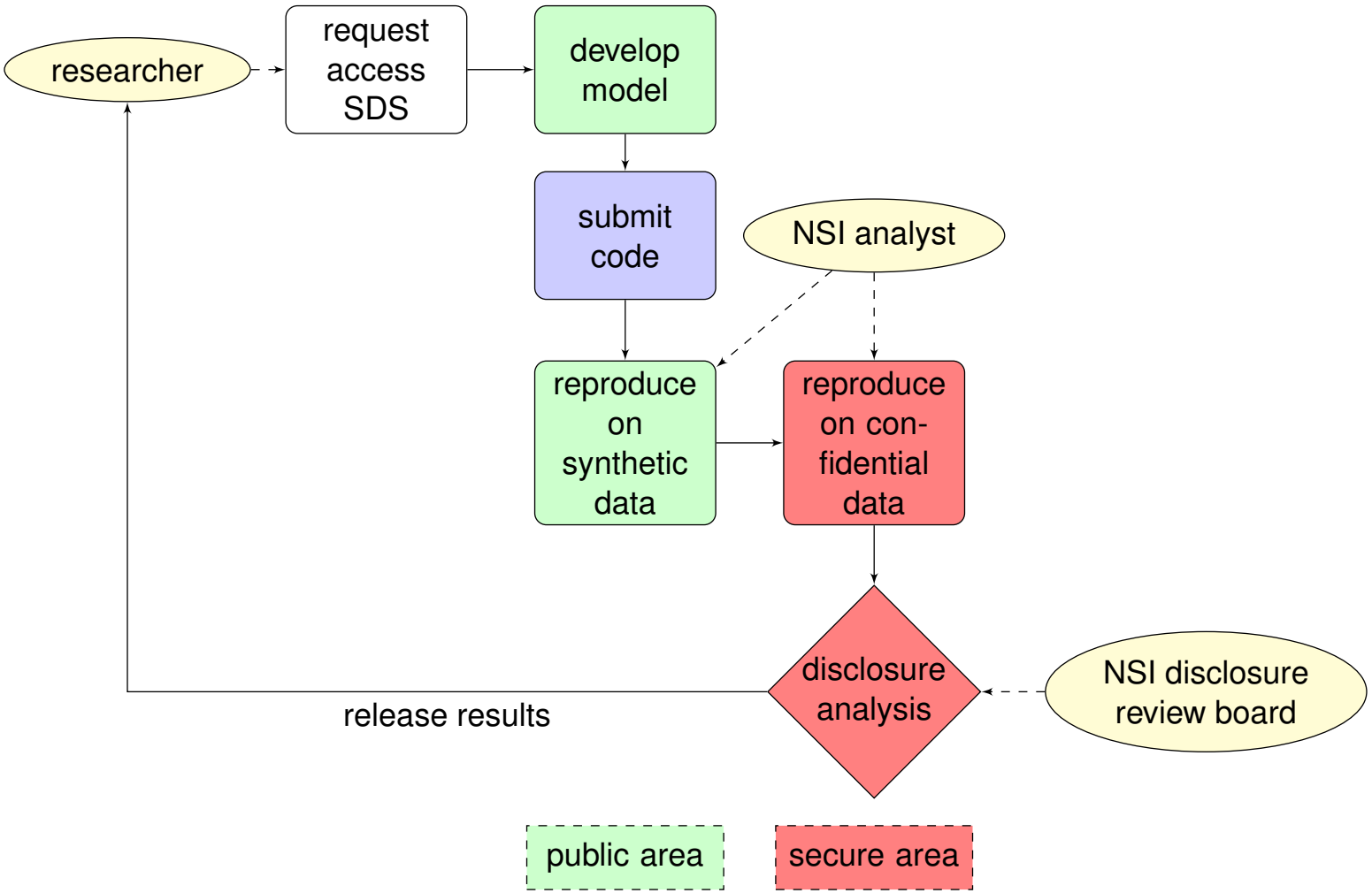


More information

www.vrdc.cornell.edu/sds

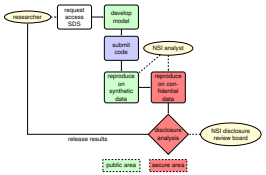
Workflow with Validation Server

Workflow with Validation Server



Access

Access is fast

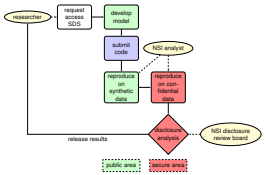


request
access
SDS

Simple access requests

- Access requests are sent to data custodians
- Access requests are only reviewed for feasibility (of the analysis on confidential data), 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

Development is more convenient



develop model

Researchers work from their own offices

- No need to travel
- Low to zero cost

Challenges/ Lessons

- Researchers must incorporate future disclosure avoidance requirements into their analysis
- Researchers must avoid hard-coded (data-informed) programming, and use data-dependent (automated) code
- Researchers develop code interactively, but must ultimately submit to an (semi-) automated system

A few restrictions

Server access

- In order to prevent users from removing datasets from the server, requests for removal of *results* are *moderated*, but **not** censored.
- Requests to download the data are denied
- To ensure replicability/validation, upload requests for auxiliary data are moderated (enfore data provenance documentation, reproducibility)

A few restrictions

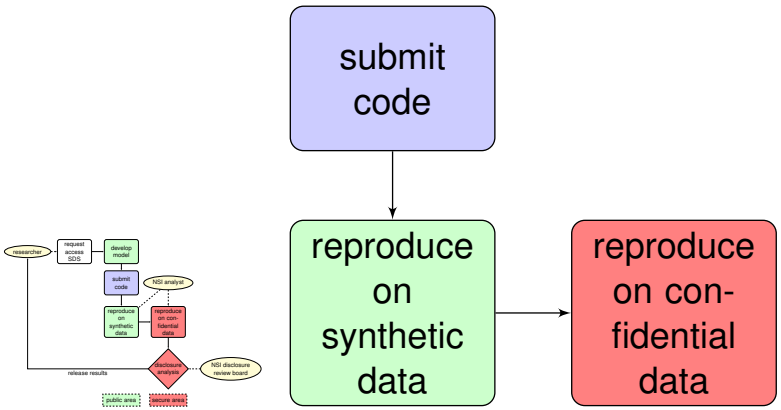
Server access

- In order to prevent users from removing datasets from the server, requests for removal of *results* are *moderated*, but **not** censored.
- Requests to download the data are denied
- To ensure replicability/validation, upload requests for auxiliary data are moderated (enfore data provenance documentation, reproducibility)

Server access

- software is limited to **SAS, Stata**.
- R, Matlab, Python may be available upon special request and upon coordination with data custodians (limitation imposed by target environment).

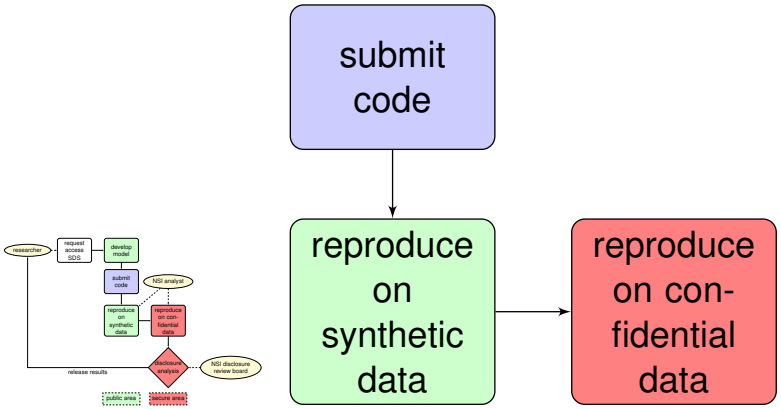
Validation



Notes

- No restrictions on type of model to be estimated
- 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).

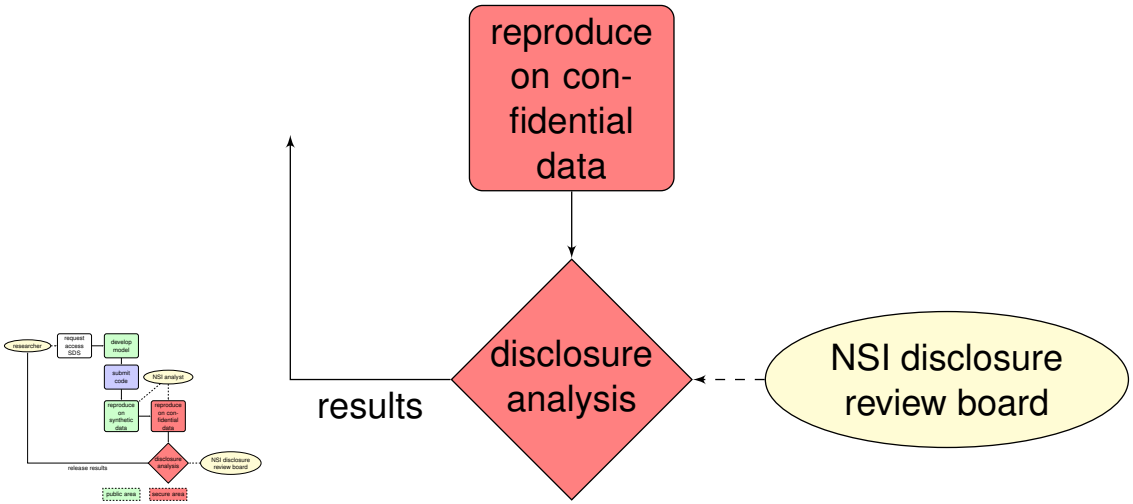
Validation



Caveats

- Code often fails upon first attempt (see above)
- Currently not automated (submission, feedback, reproduction), involves manual labor

Obtaining results

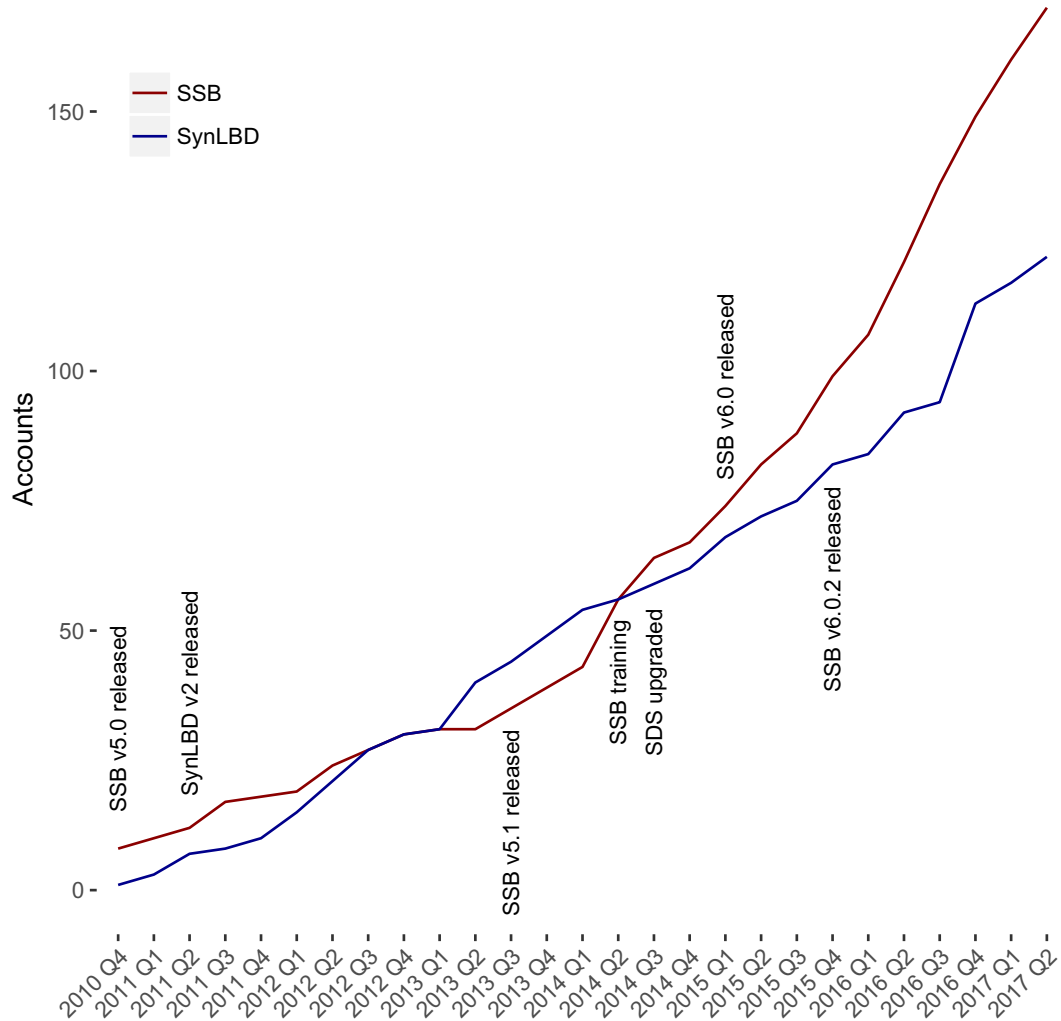


Notes

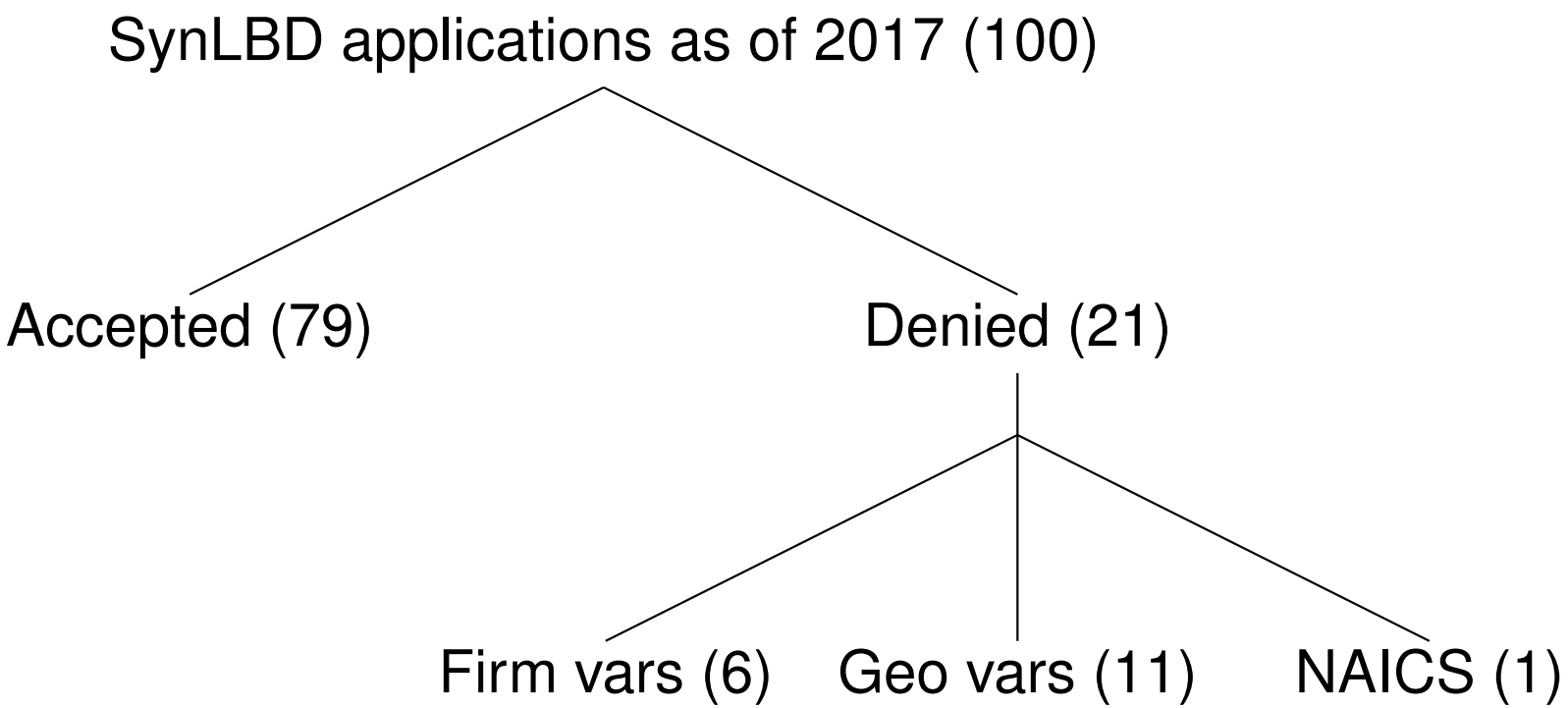
- Validated results must pass disclosure-avoidance analysis $\rightarrow$ some limitation (quantity, count restrictions)
- requires that users provide documentation of the results similar to a disclosure review request at FSRDC
- delays have increased over time

Outcomes

Accounts created (as of 2017)



Not all applications get accepted



Key feature: Feedback loop

User feedback incorporated into each version

SSB

- Variables
- Structure

→ V5, V6, V7 (see Benedetto et al. [2018] for details)

SynLBD

- NAICS
- firm-structure
- geography

→ V3.0 (see Kinney et al. [2014] for plans)

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

For both datasets

about 8 out every 100 projects request validation

How well does validation work

Bertrand et al. [2015]

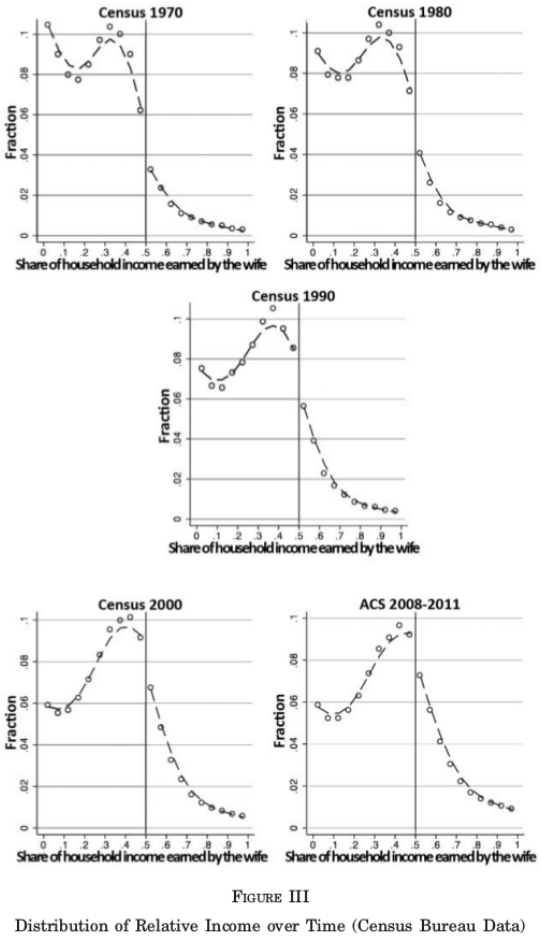
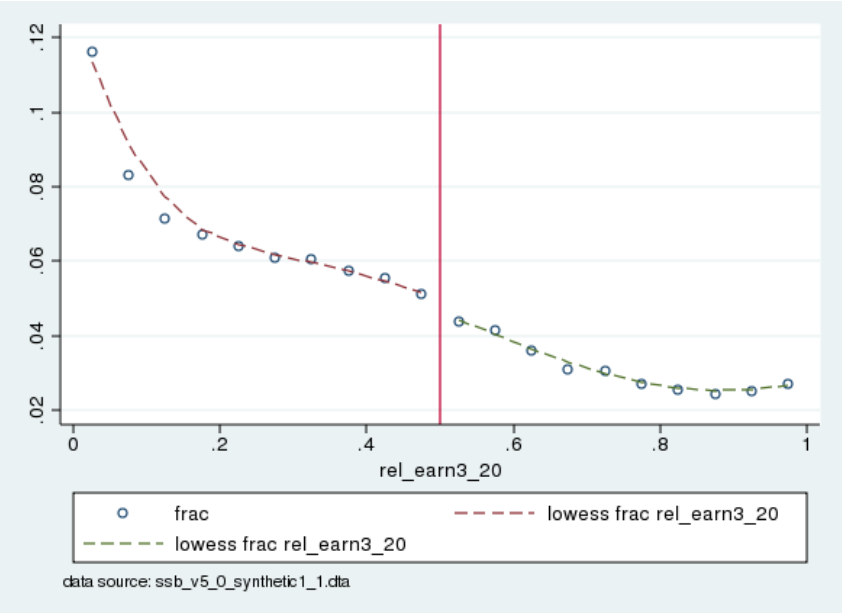


FIGURE III
Distribution of Relative Income over Time (Census Bureau Data)

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

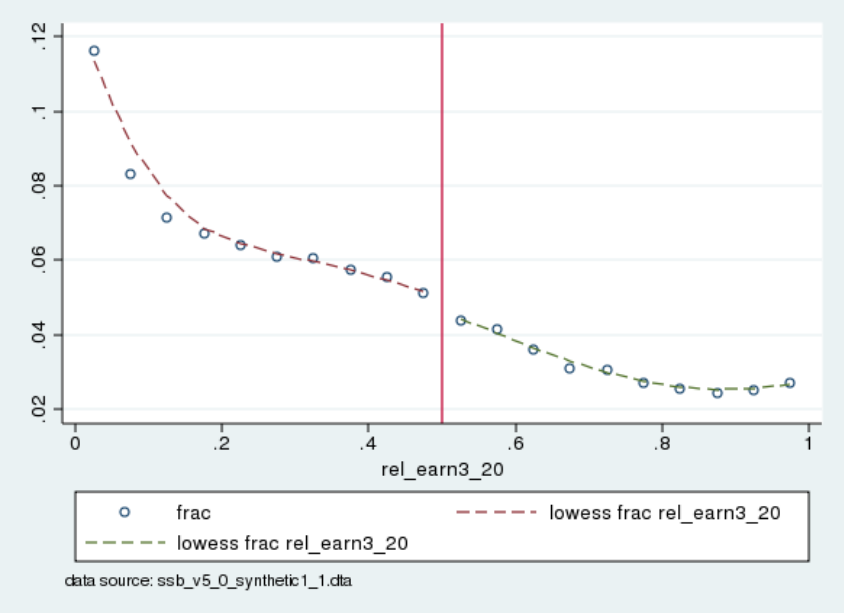
Bertrand et al on SSB



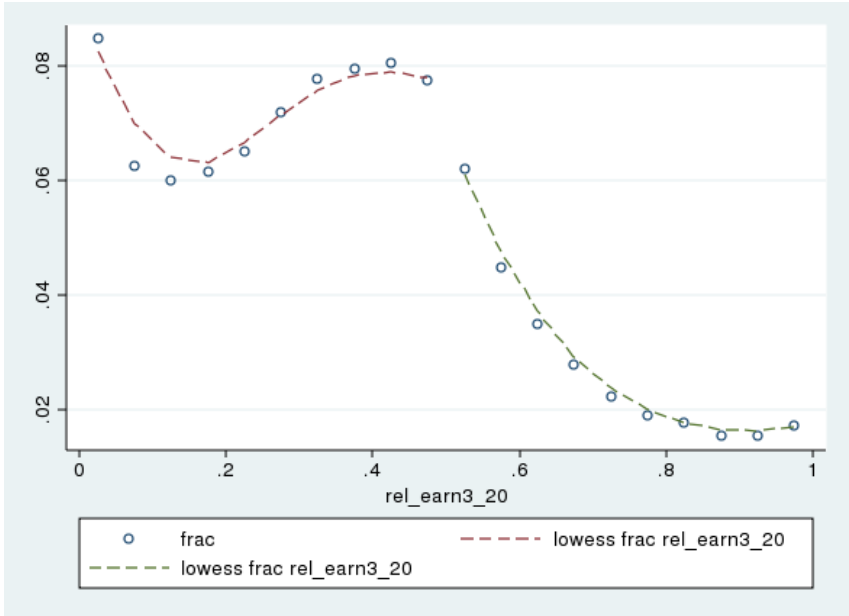
No such break in the synthetic data

How well does validation work

Bertrand et al on SSB



Bertrand et al. [2015]:



How well does validation work

General approach: proximity of coefficients $t_{\Delta\beta_{k,m}}$

We compute

$$t_{\Delta\beta_{k,m}} = \frac{\beta_{k,m} - \beta^*_{k,m}}{\sqrt{s^2_{k,m} + s^{*2}_{k,m}}}$$

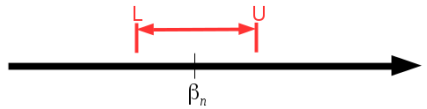
and assess its statistical significance (90% bilateral). The fraction of insignificant tests across all estimated models and parameters is an indicator of how close the synthetic and confidential models are under the estimated models.

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)

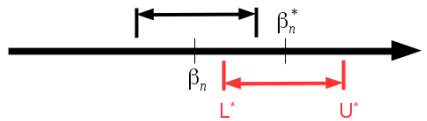


How well does validation work

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)

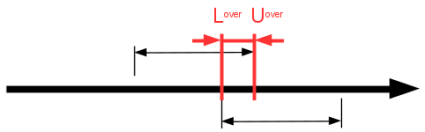


How well does validation work

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

Proximity

User	Request	Fraction	Dataset
A	1	0.10	SynLBD
A	2	0.06	SynLBD
B	1	0.87	SynLBD
C	1	0.17	SynLBD
D	1	0.63	SSB
E	1	0.62	SSB

How well does validation work

Coverage

User	Request	Mean	75th	90th	Max	Dataset
A	1	0.16	0.25	0.72	0.89	SynLBD
A	2	0.10	0.00	0.52	0.92	SynLBD
B	1	0.87	1.00	1.00	1.00	SynLBD
C	1	0.22	0.51	0.72	0.99	SynLBD
D	1	0.49	0.79	0.87	0.98	SSB
E	1	0.39	0.56	0.63	0.94	SSB

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 data

How well does validation work

Downside

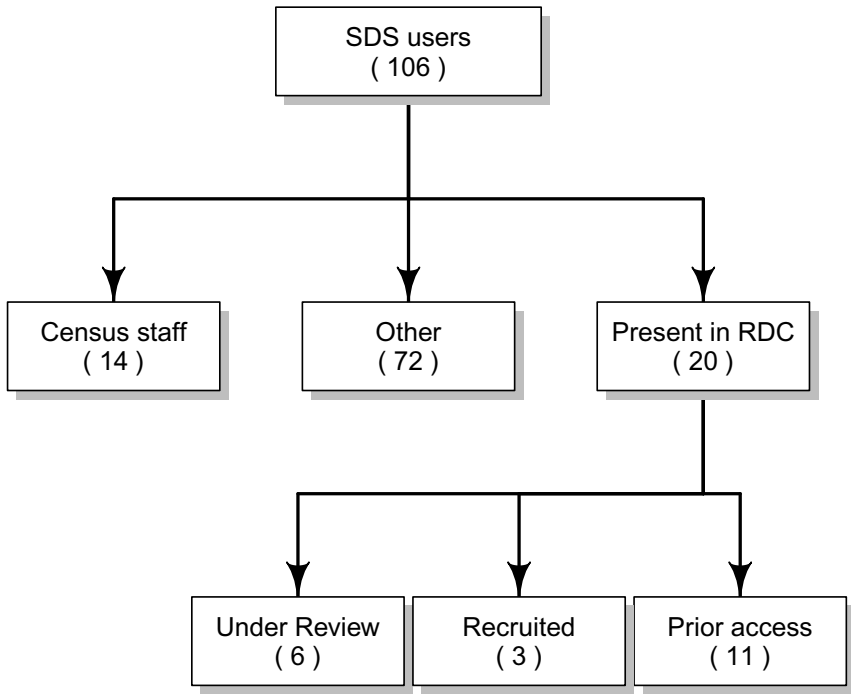
- Cannot adapt your model to the data
- Fundamental: will not work for non-congenial designs (f.i. regression discontinuity)

Upside

- Cannot adapt your model to the data
- ~~Rapid turnaround (about 1 week)~~ Faster turnaround than FSRDC to get result from confidential data

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: turnaround for validation = [7, 90] days...)

Next steps

Ongoing efforts

- Alam, Dostie, Vilhuber [xxxx] Synthetic LEAP (Canada)
  - Limited pilot
  - Accessible only within Canadian RDCs (instead of traveling to Ottawa, Canada [150, 3500] kms)
  - Outcomes unknown yet
- Burman et al. [2018] administrative tax data
  - not released yet

A new model for validation

Validation = Replication

- Use modern technologies (Docker, Rmarkdown, Jupyter)
- Leverage pervasive infrastructure (CodeOcean, Binder, Whole Tale, Gigantum, etc.)

Build analysis on a container, submit container.

Push back to user software

Simplify and Scale

Old school

```
sed -i 's/synthetic/confidential/g' submitted_code.
submit submitted_code.R {SERVER} | \
   drb_system_filter > mail
```

New school

```
import validate from census_validation
# test the analysis
myanalysis.syntheticdata.output
# validate the analysis
validate.authenticate()
myanalysis.validate.output
```

Stay tuned!

www.vrdc.cornell.edu/sds

Thank you!

```
$Id: Presentation-subdoc.tex 6886 2017-07-19 21:29:29Z 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, J. C. Stanley, and E. Totty. The creation and use of the SIPP Synthetic Beta v7.0. Technical report, U.S. Census Bureau, Nov. 2018. URL https://www.census.gov/content/dam/Census/library/working-papers/2018/adrm/Creation_SSBv7.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.

L. E. Burman, A. Engler, S. Khitatrakun, J. R. Nunns, S. Armstrong, J. Iselin, G. MacDonald, and P. Stallworth. Administrative tax data: Creating a synthetic public use file and a validation server. document, Tax Policy Center, Urban Institute and Brookings Institution, 2018. URL https://www.urban.org/research/publication/safely-expanding-research-access-administrative-tax-data-creating-synthetic-public-use-fi

J. Drechsler and J. P. Reiter. Disclosure risk and data utility for partially synthetic data: An empirical study using the German IAB establishment survey. Journal of Official Statistics, 25:589–603, 2009.

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.

S. K. Kinney, J. P. Reiter, and J. Miranda. Improving The Synthetic Longitudinal Business Database. Working Papers 14-12, Center for Economic Studies, U.S. Census Bureau, Feb. 2014. URL http://ideas.repec.org/p/cen/wpaper/14-12.html.

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://