A modern container-based approach for development of and access to confidential data

Presentation to the NASEM Consensus Panel "Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation"

Lars Vilhuber[1]

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

November 2022

Disclaimer

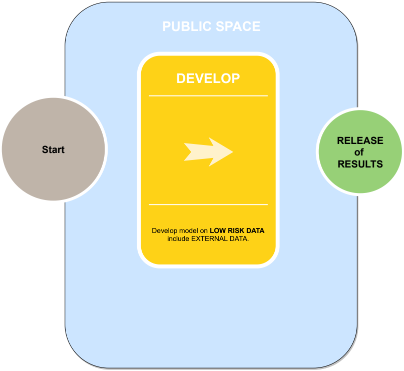
The opinions expressed in this talk are solely the author's, and do not represent the views of the U.S. Census Bureau, the American Economic Association, or any of the funding agencies.

Goal

Providing access to confidential data: a dichotomy?

- Apply SDL methods to render data less sensitive (coarsen, sample), publish **public-use data**

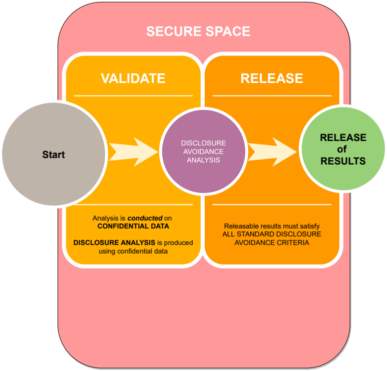
Public-use data process



Providing access to confidential data: a dichotomy?

- Apply SDL methods to render data less sensitive (coarsen, sample), publish **public-use data**
- Provide access to confidential data directly, apply SDL methods to **outputs of models** (**RDC model**)

Public-use data process


SECURE SPACE
VALIDATE
RELEASE
Start
DISCLOSURE AVOIDANCE ANALYSIS
RELEASE of RESULTS
Analysis is *conducted* on **CONFIDENTIAL DATA**
DISCLOSURE ANALYSIS is produced using confidential data
Releasable results must satisfy ALL STANDARD DISCLOSURE AVOIDANCE CRITERIA

Providing access to confidential data: broadening the toolkit

- Apply SDL methods to render data less sensitive (coarsen, sample), publish **public-use data**
- Generate synthetic data, publish **synthetic public-use data**
- Provide access to confidential data directly, apply SDL methods to **outputs of models** (**RDC model**)

What about inference validity?

- **Public-use data**: rarely disputed, sometimes tested, occassionally fails (unions, top earnings, ...)
- **Synthetic public-use data**: design goal, generally high target (replace access to confidential data)
- **RDC model**: generally undisputed, because access is to the unmodified data

What about confidentiality protection?

- **Public-use data**: old-style SDL, sometimes disputed. Known to fail for newer data (linked data, geographically precise data)
- **Synthetic public-use data**: design goal, generally high target (replace access to confidential data)
- **RDC model**: high, but at the cost of complex access mechanism

Synthetic Data

Synthetic Data

> *"Synthetic data are simulated data generated from statistical models."*

SSB webpage

Synthetic Data are protective

“They are designed to protect the confidentiality of the people and firms in the underlying confidential data”

SSB webpage

(Partially) 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

Also called “synthetic data”:

These are not...

1. “Test files” (= synthetic data drawn from univariate distributions) (used at various statistical agencies)
2. Custom-generated (bespoke) synthetic data per project SYLLS

   [Nowok et al., 2016b]

Variations in the data generating process

Synthetic data can be generated in a variety of ways

1. Drawn from univariate distributions of each variable
2. Sampled from a posterior-predictive distribution of a particular analytical model
3. Sampled from a posterior-predictive distribution of a model of the data
4. Sampled from a differentially private high-dimensional histogram

This affects both analytic (or inferential) validity as well as protective properties

Synthetic Microdata in the Wild

- **SIPP Synthetic Beta** (person-level survey + admin data) [Abowd et al., 2006, Benedetto et al., 2018]
- **Synthetic LBD** (establishment-level data) [Kinney et al., 2011]
- German establishment data [Drechsler, 2012]
- Canadian SyntheticLEAP (firm-level data) [Alam et al., 2020]

Synthetic Tabular Data in the Wild

- ACS group quarters data (tabular data)
- OnTheMap (tabular data) [Machanavajjhala et al., 2008]

Most of these are aimed at “inferentially valid” applications.

Quality assessment of synthetic data

Three dimensions of assessment

1. Inferential validity: are inferences based on synthetic data approximately the same as those drawn from unprotected data?
2. Confidentiality protection: are the synthetic data protective of the privacy of respondents in the unprotected data?
3. Analyst utility: are the synthetic data useful to analysts?

Datasets

SIPP Synthetic Beta (SSB)

- provide access to linked data that are usually not publicly available
  - mostly synthetic: all variables except two are synthesized
  - *gender* and a *link to the first reported marital partner* are the exception.
  - method: 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.
- Method: Benedetto et al. [2018]
- Codebook: Reeder et al. [2018]

Developing Synthetic Data

Enter the Validation Server

- Creating (general purpose) synthetic data that achieves high inferential validity and high confidentiality protection is **hard**.
- Target measures for known analyses/models are straightforward to achieve [Kinney et al., 2011, Drechsler, 2012, Abowd et al., 2006, Benedetto et al., 2018] and can even be achieved using **canned** procedures [Nowok et al., 2016a]
- **Broadening** the set of valid analyses is **hard** (problem of congeniality of model [Meng, 1994])

Approach taken: iterative approach combined with unstructured user feedback loop through validation [Reiter et al., 2009].

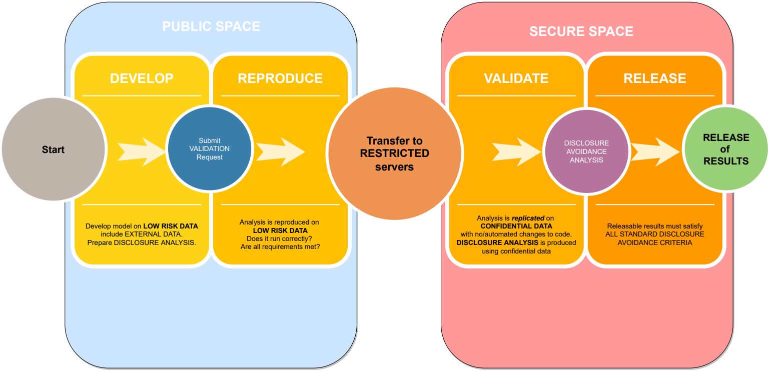
Validation Server

Validation

- Researcher develops analysis in unsecure environment (using synthetic data, public-use data): $M(\Xi(D)) = M(D^*) \rightarrow Y^*$
- Researcher submits analysis to data publisher
- Data publisher runs analyis on confidential data $M(D) \rightarrow Y$
- Data publisher reports back
  - Quality metric $Q(Y^*, Y)$ and/or
  - Confidentialized output from model $\Xi(Y) = Y^{'}$

In principle not restricted to synthetic data - any modified data (e.g., **public-use data**) can be “validated” through such a system.

Validation Server



PUBLIC SPACE

DEVELOP

REPRODUCE

Start

Submit VALIDATION Request

Develop model on **LOW RISK DATA** include EXTERNAL DATA. Prepare DISCLOSURE ANALYSIS.

Analysis is reproduced on **LOW RISK DATA** Does it run correctly? Are all requirements met?

Transfer to RESTRICTED servers

SECURE SPACE

VALIDATE

RELEASE

DISCLOSURE AVOIDANCE ANALYSIS

RELEASE of RESULTS

Analysis is ***replicated*** on **CONFIDENTIAL DATA** with no/automated changes to code. **DISCLOSURE ANALYSIS** is produced using confidential data

Releasable results must satisfy ALL STANDARD DISCLOSURE AVOIDANCE CRITERIA

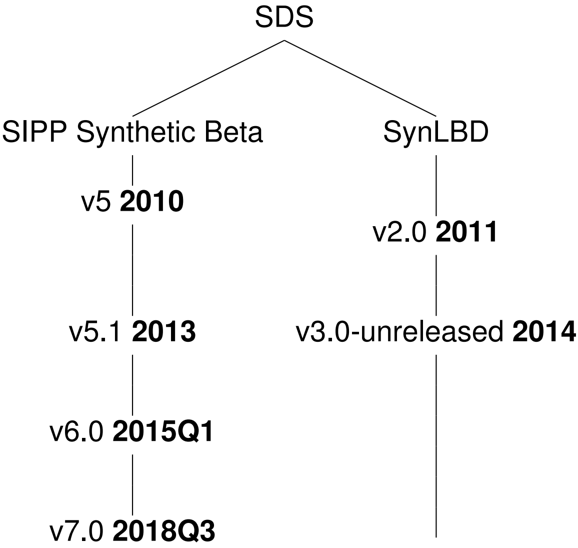
SDS Validation Server Cycle

The Synthetic Data Server (SDS) was set up with the specific goal to test a mechanism to gather broad user input and improve the synthetic data iteratively.

Iterate

- Make available v(X) of synthetic data
- Provide access to researchers
- Allow for validation
- Collect models, feedback
- Revise synthetic data generation, prepare v($X+1$)

History of datasets



SDS Validation Server Quid-pro-quo

The Synthetic Data Server (SDS) embodied a quid-pro-quo between researchers and the agency.

NSI

- Improve synthetic data generation, prepare v$(X+1)$
- To do so: Collect feedback from researcher

Researcher

- Convenient access to data
- Provide feedback

This is specific to the SDS - it is not a general feature of validation servers!

System

Goals

The SDS was designed around the following goals:

- Provide convenient researcher access to **suitable development environment**
- Provide “**guardrails**” to facilitate later validation in confidential environment
- Prevent **misperception** of quality of synthetic data (“beta”, not called “public-use” product)
- Allow for mostly **unrestricted** model development

Note: This version of SDS and the validation mechanism was not meant to scale without further development of the *mechanism*.

Validation environment: Census Bureau

Basic parameters

- Linux system
- Typically “batch submission” system
- Secure environment, no internet access
- Any outside files must be explicitly identified, copied onto system

Public Space: SDS

Synthetic Data Server (SDS)

- Remote graphical desktop on Linux
- emulate Census Bureau environment to a large extent
  - file system structure
  - software availability
  - batch submission system
- Accessible over the internet, but cannot access internet
- Any outside files must be explicitly identified, copied onto system

In essence, a *Virtual Desktop Environment (VDE)* as others provide it, but with Linux. All VDE are essentially 1990s desktop-centric technology.

Why this structure?

Reproducibility

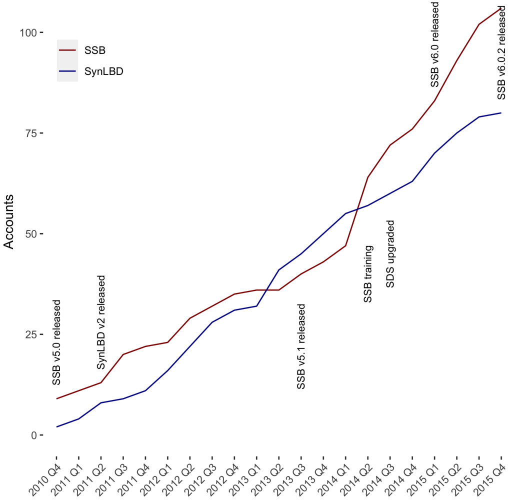
- The SDS emulates the target compute environment closely
- Allows researchers to create code that can be re-run on the confidential data

Enforce non-redistribution of dataset

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

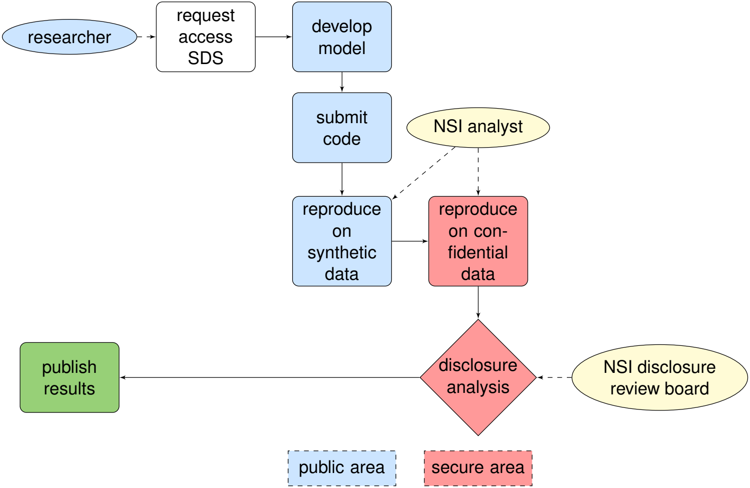
Usage

6 (versions of) synthetic datasets, over 200 users in first 5 years

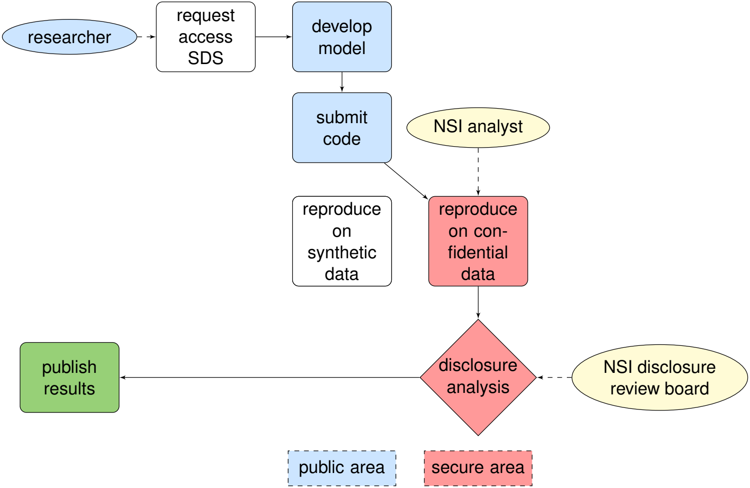


Growth substantially decreased afterwa

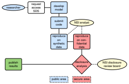
Workflow with Validation Server



Workflow with Validation Server: In practice



Access process

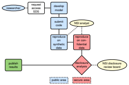


request access SDS

Simple access requests

- Access requests were sent to data custodians (separately for SynLBD and SSB)
- Access requests were only reviewed for feasibility (of the analysis on confidential data), but were not otherwise restricted.
- Once access was verified, the server provider (Cornell University) set up accounts on the system

Model Development



researcher

request access SDS

develop model

submit code

NSI analyst

reproduce on synthetic data

reproduce on confidential data

publish results

disclosure analysis

NSI disclosure review board

public area

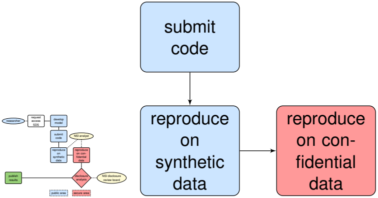
secure area

develop model

Researchers work from their own offices +

- No need to travel
- Low to zero cost
- Access to system and software at no charge
- No restrictions on type of model to be estimated
- Upload of files is moderated (enforce data provenance documentation, reproducibility, necessary for target environment)

Validation


researcher
request access SDS
develop model
submit code
NSI analyst
reproduce on synthetic data
reproduce on con-fidential data
publish results
disclosure analysis
NSI disclosure review board
public area
secure area
submit code
reproduce on synthetic data
reproduce on con-fidential data

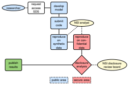
Requirements

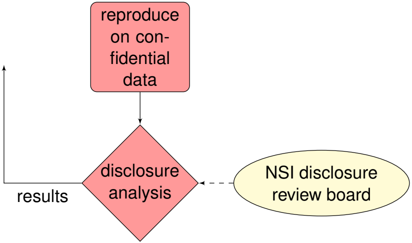
- all programs and auxiliary input files,
- documentation of the results similar to a *disclosure review request at Federal Statistical Research Data Center (FSRDC)* ,
- all programs run error-free (replicability requirement)

Process

- Conducted by program staff at Census Bureau

Obtaining results



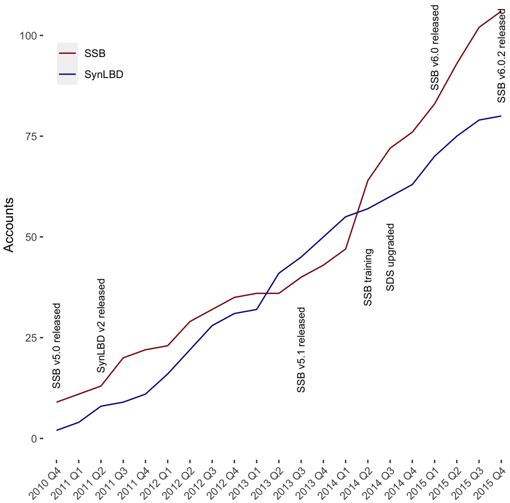


Requirements

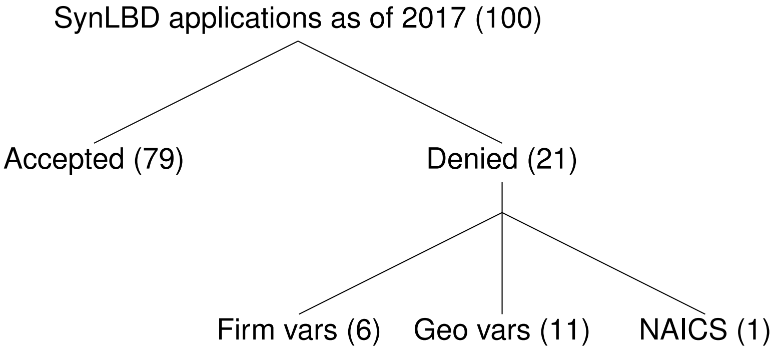
- documentation of the results similar to a *disclosure review request at FSRDC* (assisted by program staff, no researcher interaction)

Outcomes

Accounts created (as of 2015)



Analysis of denied applications: Feasibility


SynLBD applications as of 2017 (100)

Accepted (79)

Denied (21)

Firm vars (6)

Geo vars (11)

NAICS (1)

Based on analysis of universe of applications as of 2017.

Key feature: Feedback loop

User feedback (also survey) incorporated into each version

SSB

- Variables
- Structure

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

SynLBD

- NAICS
- firm-structure
- geography

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

Validation

For both datasets

about 8 out every 100 projects request validation

An ideal example: Bertrand et al (2015)

Bertrand et al. [2015]

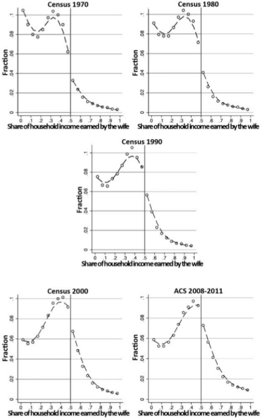


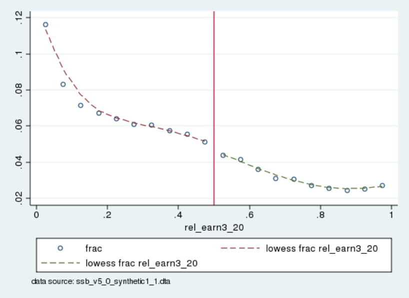
Figure III

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)

An ideal example: Bertrand et al (2015)

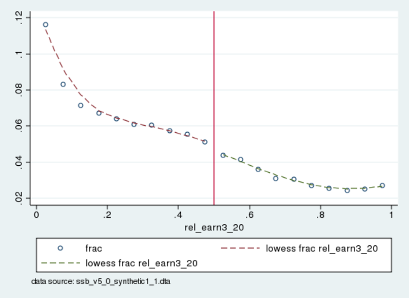
Bertrand et al on SSB



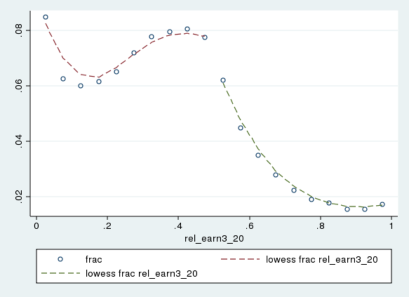
No such break in the synthetic data

An ideal example: Bertrand et al (2015)

Bertrand et al on SSB



Bertrand et al. [2015]:



Analytic validity

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 **higher** the fraction of **insignificant** tests across all estimated models and parameters, the closer the synthetic and confidential models are under the estimated models.

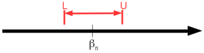
Analytic validity, more generally

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)



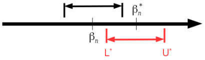
Analytic validity, more generally

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)



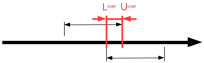
Analytic validity, more generally

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



Analytic validity

Then the overlap in confidence intervals is

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

Challenges in measurement

No systematic mechanism implemented from the start

- In collaboration with Census Bureau, we added code to users' SAS/Stata programs ex-post, attempting to capture parameter values estimated
- **Difficult**, because no homogeneous structure
- **Limiting**, since no easy way to identify "main parameters" of interest from "nuissance" parameters (controls)

Analytic validity

Proximity $t_{\Delta \beta_{k,m}}$

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

Fraction of test statistics that are not significant. Higher is better.

Analytic validity

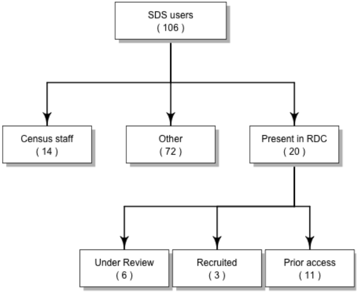
Coverage J_k^*

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

Distribution of coverage across all parameters. Higher is better.

Gateway data

Figure: Connection between Census RDC usage and Synthetic Data Server



Based on a manual analysis of (FS)RDC projects in 2017.

SDS and FSRDC

Other outcomes/interactions

- at least some of the RDC projects were direct continuation of SDS projects(source: private conversations)
- average delay (project start (SDS), project start (RDC)) : 400 days.
- (reminder: turnaround for validation = [7, 90] days...)

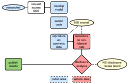
Lessons

Lessons

Some Lessons Learned from the Pilot

[TLDR]

Access process



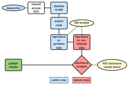
request access SDS

Access requests

- **Manual** process
- Simple, when compared to access to FSRDC
- Typical turnaround time is **1-10 days**

Not homogeneous across datasets

Model Development

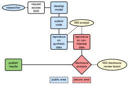


develop model

Challenges/ Lessons for Researchers

- do not work in their **usual** work environment (*perceived or real lack of flexibility*)
- must incorporate future disclosure avoidance requirements into their analysis (note: *little guidance, no tools*)
- must avoid hard-coded (data-informed) programming, and use data-dependent (automated) code (this seems *challenging for many researchers*)
- develop code interactively, but must ultimately submit to an (semi-) automated system (*= reproducibility, not always clearly communicated*)

Model Development

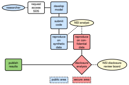


develop model

Server transfers

- Requests for removal of *results* are ***manually** moderated* - **take a few days**
- *Upload requests for auxiliary data are **manually** moderated*

Model Development



develop model

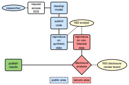
Server transfers

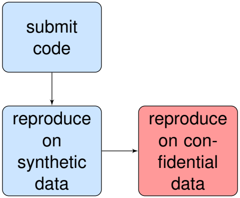
- Requests for removal of *results* are ***manually** moderated* - **take a few days**
- *Upload requests for auxiliary data are **manually** moderated*

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*, mostly accomodated).

Validation

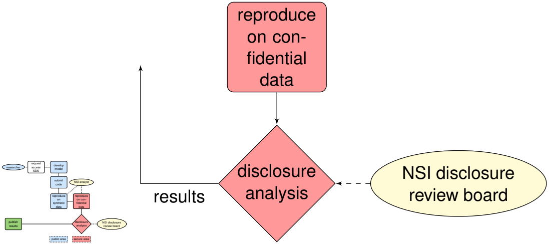




Challenges

- Anecdotally, code often fails upon first validation attempt
- Not automated (submission, feedback, reproduction), involves **(lots of) manual labor**

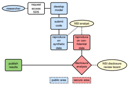
Obtaining results



Challenges

- Validated results must **pass disclosure-avoidance analysis** → some limitation (quantity, count restrictions)
- requires that users provide documentation of the results similar to a *disclosure review request at FSRDC* (**Challenging, and these have evolved over time!**),
- delays have increased (**substantially!**) over time

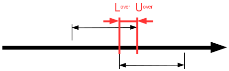
Some tentative conclusions



Process

- Multiple friction points: Disclosure of results, Initial access, Development, Removal of results based on synthetic data (in decreasing order of importance)
- Due to nature of pilot, labor intensity is high (no automation)
- User support should be improved: Tools to prepare and assess disclosure avoidance measures, success of porting model

Some tentative conclusions



Data and models

- In general, data quality is sufficient for model development
- In general, data quality is insufficient to replace confidential data!
- Almost no user published results based on synthetic data (Exception: Perla et al. [2021])
- Even simple papers have over 100 parameters (Ex. Carr and Wiemers [2018])

Next steps

Basic goals

Maintain general approach

- Allow for arbitrary model development (subject to minimal constraints)
- Allow for broad software usage

Basic Goals

Reduce friction points

- Streamlined and fast (minutes, not months) access to data, validation
- User-centered development
- Integrated disclosure avoidance for validation
- Greater conformance of database schema
- Lower barriers of entry

Basic Goals

Reduce cost

- Automation and self-checks
- User-hosted development
- Lower analytic validity of the synthetic data, easier confidentiality protection

Lower analytic validity

Trade-off

- Reduce target analytic validity (limit target model coverage)
- Easier confidentiality protection enables easier access model
- Allows for full schema coverage (all variables present in confidential and synthetic data, with varying fidelity)

Data Access

Licensed access to synthetic data

- Provide licensed access to the synthetic data.
- Data can be downloaded to arbitrary system, subject to recorded license agreement *[analog: IPUMS]*
- License specifies the limitations of the data, acknowledgement by user *[analog: MIT software license]*

Note: Emphasis that the synthetic data is **not analytically valid**/**robust** is key!

Validation Access

Controlled access to validation

- Provide API for validation *[analog: geocoding systems]*
- API requires simple sign-up, with acknowledgement of limitations *[analog: IPUMS Beta, industry models]*
- API can be wrapped into software-specific packages *[goal: ease of use]*
- Validation is **fully automated**! *(Including disclosure avoidance)*
- **API can be used to verify reproducibility** of submitted packages prior to validation

Example of API

Simplify and Scale

New school

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

Validation Output

Blended outputs

Combine outputs to reduce disclosure risk:

- If Verification [Reiter et al., 2009, Barrientos et al., 2018] is positive ($Q(Y, Y^*) > \overline{q}$), provide output from synthetic data $[M(\Xi(D)) = Y^*]$
- If Verification fails, provide (partial?) output from validation against confidential data $[\Xi(M(D)) = Y']$
- If quality of confidentialized output is too low ($Q(Y') < \overline{q}$), suppress and prepare FSRDC proposal...

Conditions for success

Reproducibility of code

It is crucial that submitted code be reproducible.

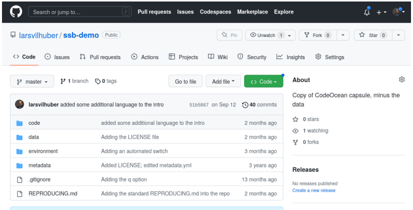
Automation of disclosure avoidance

Disclosure avoidance tools must be built in (easy to verify without manual intervention, prior to submission)

- Set expectations
- Streamline / speed-up result release

Example

Containerized processing



larsvilhuber added some additional language to the intro 51b5867 on Sep 12 40 commits

code	added some additional language to the intro	2 months ago
data	Adding the LICENSE file	2 months ago
environment	Adding an automated switch	3 months ago
metadata	Added LICENSE; edited metadata.yml	3 years ago
.gitignore	Adding the q option	13 months ago
REPRODUCING.md	Adding the standard REPRODUCING.md into the repo	2 months ago

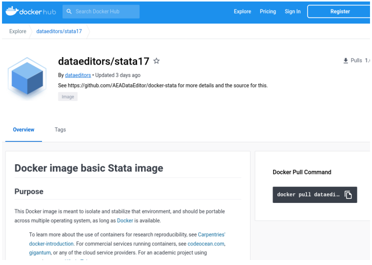
About

Copy of CodeOcean capsule, minus the data

0 stars
1 watching
0 forks

Releases

No releases published
Create a new release



dataeditors/stata17

By dataeditors • Updated 3 days ago

See https://github.com/AEADataEditor/docker-stata for more details and the source for this.

Docker image basic Stata image

Purpose

This Docker image is meant to isolate and stabilize that environment, and should be portable across multiple operating system, as long as Docker is available.

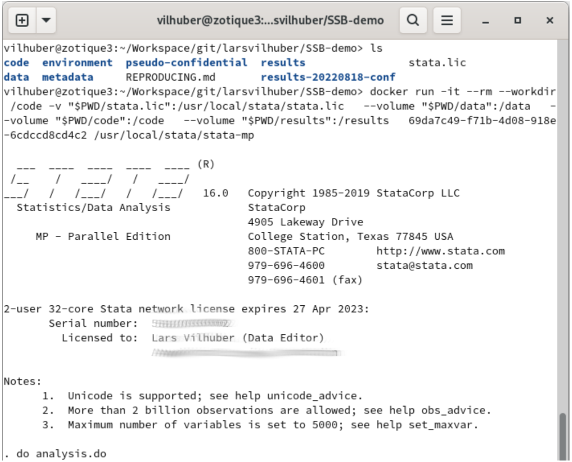
> To learn more about the use of containers for research reproducibility, see Carpentries' docker-introduction. For commercial services running containers, see codeocean.com, gigantum, or any of the cloud service providers. For an academic project using

Docker Pull Command

```
docker pull dataedi…
```

Example

Container can be run on author’s computer



```
vilhuber@zotique3:~/Workspace/git/larsvilhuber/SSB-demo> ls
code  environment  pseudo-confidential  results              stata.lic
data  metadata     REPRODUCING.md       results-20220818-conf
vilhuber@zotique3:~/Workspace/git/larsvilhuber/SSB-demo> docker run -it --rm --workdir
 /code -v "$PWD/stata.lic":/usr/local/stata/stata.lic   --volume "$PWD/data":/data   -
-volume "$PWD/code":/code   --volume "$PWD/results":/results   69da7c49-f71b-4d08-918e
-6cdccd8cd4c2 /usr/local/stata/stata-mp

  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   16.0   Copyright 1985-2019 StataCorp LLC
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     MP - Parallel Edition            College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

2-user 32-core Stata network license expires 27 Apr 2023:
       Serial number:  [redacted]
         Licensed to:  Lars Vilhuber (Data Editor)
                       [redacted]

Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  More than 2 billion observations are allowed; see help obs_advice.
      3.  Maximum number of variables is set to 5000; see help set_maxvar.

. do analysis.do
```

Example

Validating reproducibility

Run code non-interactively

```
docker run -it --rm --workdir /code \
       -v "$PWD/stata.lic":/usr/local/stata/stata.lic  \
       --volume "$PWD/data":/data        \
       --volume "$PWD/code":/code        \
       --volume "$PWD/results":/results  \
       69da7c49-f71b-4d08-918e-6cdccd8cd4c2 \
       \
       \
       /usr/local/stata/stata-mp analysis.do
```

Example

Submitting for validation

```
docker run -it --rm \
       --volume "$PWD/data":/data        \
       --volume "$PWD/code":/code        \
       --volume "$PWD/results":/results  \
       69da7c49-f71b-4d08-918e-6cdccd8cd4c2 \
       \
       \
       /special/validate $APIKEY $myemail
```

Example

Submitting for validation

```
trace submit --validate .
```

Example

Validation

```
docker run -it --rm \
       --volume "/path/to/confidential/data":/data       <====
       --volume "$PWD/code":/code        \
       --volume "$PWD/results":/results  \
       69da7c49-f71b-4d08-918e-6cdccd8cd4c2 \
       \
       \
       /usr/local/stata/stata-mp analysis.do
```

Science fiction?

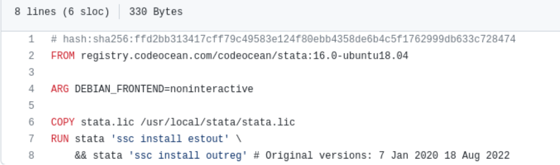
All but one exist

`/special/validate` does not exist yet.

Security?

Isolation can be brought to bear

Containerization not per-se security software, but only clear-text data transmitted



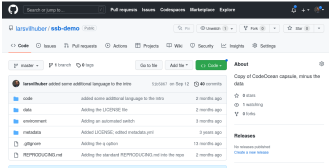
```
8 lines (6 sloc)  |  330 Bytes

# hash:sha256:ffd2bb313417cff79c49583e124f80ebb4358de6b4c5f1762999db633c728474
FROM registry.codeocean.com/codeocean/stata:16.0-ubuntu18.04

ARG DEBIAN_FRONTEND=noninteractive

COPY stata.lic /usr/local/stata/stata.lic
RUN stata 'ssc install estout' \
    && stata 'ssc install outreg' # Original versions: 7 Jan 2020 18 Aug 2022
```

Cannot install Docker?

NOTE: Not an endorsement.

Use cloud service

- AWS, Google Cloud, Azure, Github Codespaces
- (free or low-cost: pennies per compute hour)



larsvilhuber added some additional language to the intro 51b5867 on Sep 12 40 commits

code	added some additional language to the intro	2 months ago
data	Adding the LICENSE file	2 months ago
environment	Adding an automated switch	3 months ago
metadata	Added LICENSE; edited metadata.yml	3 years ago
.gitignore	Adding the q option	13 months ago
REPRODUCING.md	Adding the standard REPRODUCING.md into the repo	2 months ago

About

Copy of CodeOcean capsule, minus the data

0 stars

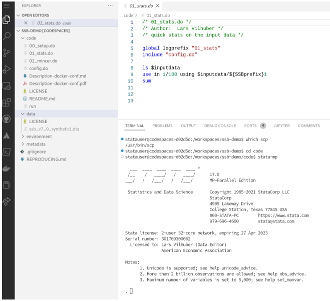
1 watching

0 forks

Releases

No releases published

Create a new release



```
/* 01_stats.do */
/* Author:  Lars Vilhuber */
/* quick stats on the input data */

global logprefix "01_stats"
include "config.do"

ls $inputdata
use in 1/100 using $inputdata/${SSBprefix}1
sum
```

```
statauser@codespaces-d02d5d:/workspaces/ssb-demo$ which scp
/usr/bin/scp
statauser@codespaces-d02d5d:/workspaces/ssb-demo$ cd code
statauser@codespaces-d02d5d:/workspaces/ssb-demo/code$ stata-mp

  ___  ____  ____  ____  ____ ®
 /__    /   ____/   /   ____/      17.0
___/   /   /___/   /   /___/       MP—Parallel Edition

 Statistics and Data Science       Copyright 1985-2021 StataCorp LLC
                                   StataCorp
                                   4905 Lakeway Drive
                                   College Station, Texas 77845 USA
                                   800-STATA-PC        https://www.stata.com
                                   979-696-4600        stata@stata.com

Stata license: 2-user 32-core network, expiring 27 Apr 2023
Serial number: 501709300062
  Licensed to: Lars Vilhuber (Data Editor)
               American Economic Association

Notes:
      1. Unicode is supported; see help unicode_advice.
      2. More than 2 billion observations are allowed; see help obs_advice.
      3. Maximum number of variables is set to 5,000; see help set_maxvar.

. 
```

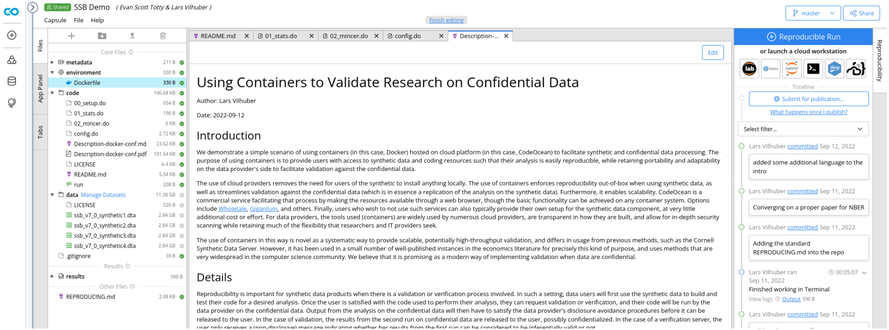
Not user friendly?

Correct

- Could be integrated into software packages (`ssc install censusvalidate` or `install.packages("censusvalidate")`)
- Use existing online systems (free or low-cost)

Using Commercial Services

NOTE: Not an endorsement.



Core Files

metadata	211 B
environment	330 B
Dockerfile	330 B
code	146.68 KB
00_setup.do	654 B
01_stats.do	198 B
02_mincer.do	6 KB
config.do	2.72 KB
Description-docker-conf.md	23.62 KB
Description-docker-conf.pdf	101.54 KB
LICENSE	6.4 KB
README.md	5.34 KB
run	208 B
data Manage Datasets	11.38 GB
LICENSE	526 B
ssb_v7_0_synthetic1.dta	2.84 GB
ssb_v7_0_synthetic2.dta	2.84 GB
ssb_v7_0_synthetic3.dta	2.84 GB
ssb_v7_0_synthetic4.dta	2.84 GB
.gitignore	39 B

Results

results	590 B

Other Files

REPRODUCING.md	2.08 KB

Using Containers to Validate Research on Confidential Data

Author: Lars Vilhuber

Date: 2022-09-12

Introduction

We demonstrate a simple scenario of using containers (in this case, Docker) hosted on cloud platform (in this case, CodeOcean) to facilitate synthetic and confidential data processing. The purpose of using containers is to provide users with access to synthetic data and coding resources such that their analysis is easily reproducible, while retaining portability and adaptability on the data provider's side to facilitate validation against the confidential data.

The use of cloud providers removes the need for users of the synthetic to install anything locally. The use of containers enforces reproducibility out-of-box when using synthetic data, as well as streamlines validation against the confidential data (which is in essence a replication of the analysis on the synthetic data). Furthermore, it enables scalability. CodeOcean is a commercial service facilitating that process by making the resources available through a web browser, though the basic functionality can be achieved on any container system. Options include Wholetale, Gigantum, and others. Finally, users who wish to not use such services can also typically provide their own setup for the synthetic data component, at very little additional cost or effort. For data providers, the tools used (containers) are widely used by numerous cloud providers, are transparent in how they are built, and allow for in-depth security scanning while retaining much of the flexibility that researchers and IT providers seek.

The use of containers in this way is novel as a systematic way to provide scalable, potentially high-throughput validation, and differs in usage from previous methods, such as the Cornell Synthetic Data Server. However, it has been used in a small number of well-published instances in the economics literature for precisely this kind of purpose, and uses methods that are very widespread in the computer science community. We believe that it is promising as a modern way of implementing validation when data are confidential.

Details

Reproducibility is important for synthetic data products when there is a validation or verification process involved. In such a setting, data users will first use the synthetic data to build and test their code for a desired analysis. Once the user is satisfied with the code used to perform their analysis, they can request validation or verification, and their code will be run by the data provider on the confidential data. Output from the analysis on the confidential data will then have to satisfy the data provider's disclosure avoidance procedures before it can be released to the user. In the case of validation, the results from the second run on confidential data are released to the user, possibly confidentialized. In the case of a verification server, the user only receives a (non-disclosive) message indicating whether her results from the first run can be considered to be inferentially valid or not.

Lars Vilhuber committed Sep 12, 2022

added some additional language to the intro

Lars Vilhuber committed Sep 11, 2022

Converging on a proper paper for NBER

Lars Vilhuber committed Sep 11, 2022

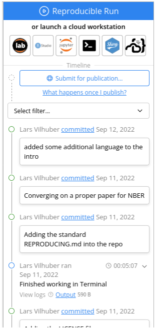
Adding the standard REPRODUCING.md into the repo

Lars Vilhuber ran 00:05:07
Sep 11, 2022
Finished working in Terminal
View logs Output 590 B

Lars Vilhuber committed Sep 11, 2022

Using Commercial Services

NOTE: Not an endorsement.


Reproducible Run
or launch a cloud workstation
lab
Studio
jupyter
Shiny
Timeline
Submit for publication...
What happens once I publish?
Select filter...

Lars Vilhuber committed Sep 12, 2022

added some additional language to the intro

Lars Vilhuber committed Sep 11, 2022

Converging on a proper paper for NBER

Lars Vilhuber committed Sep 11, 2022

Adding the standard REPRODUCING.md into the repo

Lars Vilhuber ran Sep 11, 2022 00:05:07

Finished working in Terminal

View logs Output 590 B

Lars Vilhuber committed Sep 11, 2022

Recommendations

Recommendations

1. Provide scalable mechanism - packages, API, submission websites, licensed data, etc.
2. Provide clear disclosure avoidance criteria and **tools**
3. Provide set of results: Verification [Reiter et al., 2009, Barrientos et al., 2018], validation, only then FSRDC
4. Provide fast turnaround time (minutes, not months)

Current Barriers

1. Provide scalable mechanism - packages, API, submission websites, licensed data, etc.
2. **Provide clear disclosure avoidance criteria and tools**
3. Provide set of solutions: Verification [Reiter et al., 2009, Barrientos et al., 2018], validation, only then FSRDC (Statistics of this)
4. Provide fast turnaround time (minutes, not months)

Thank you!

Funding

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

Bibliography

J. M. Abowd, M. Stinson, and G. Benedetto. Final report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. Technical report, U.S. Census Bureau, 2006. URL http://hdl.handle.net/1813/43929.

M. J. Alam, B. Dostie, J. Drechsler, and L. Vilhuber. Applying data synthesis for longitudinal business data across three countries. Statistics in Transition New Series, 21(4):212–236, 2020. ISSN 1234-7655, 2450-0291. doi: 10.21307/stattrans-2020-039. URL https://www.exeley.com/statistics_in_transition/doi/10.21307/stattrans-2020-039.

A. F. Barrientos, A. Bolton, T. Balmat, J. P. Reiter, J. M. de Figueiredo, A. Machanavajjhala, Y. Chen, C. Kneifel, and M. DeLong. Providing Access to Confidential Research Data Through Synthesis and Verification: An Application to Data on Employees of the U.S. Federal Government. The Annals of Applied Statistics, June 2018. URL http://arxiv.org/abs/1705.07872.

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.

M. D. Carr and E. E. Wiemers. New evidence on earnings volatility in survey and administrative data. AEA Papers and Proceedings, 108:287–91, 2018. doi: 10.1257/pandp.20181050. URL http://www.aeaweb.org/articles?id=10.1257/pandp.20181050.

J. Drechsler. New data dissemination approaches in old Europe – synthetic datasets for a German establishment survey. Journal of Applied Statistics, 39(2):243–265, Feb. 2012. ISSN 0266-4763, 1360-0532. doi: 10.1080/02664763.2011.584523. URL http://www.tandfonline.com/doi/abs/10.1080/02664763.2011.584523.

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.

Bibliography II

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.

A. Machanavajjhala, D. Kifer, J. M. Abowd, J. Gehrke, and L. Vilhuber. Privacy: Theory meets practice on the map. International Conference on Data Engineering (ICDE), pages 277–286, 2008. doi: 10.1109/ICDE.2008.4497436. URL http://dx.doi.org/10.1109/ICDE.2008.4497436.

X.-L. Meng. Multiple-Imputation Inferences with Uncongenial Sources of Input. Statistical Science, 9(4):538 – 558, 1994. doi: 10.1214/ss/1177010269. URL https://doi.org/10.1214/ss/1177010269.

B. Nowok, G. M. Raab, and C. Dibben. synthpop : Bespoke creation of synthetic data in r. Journal of Statistical Software, 74:1–26, 2016a. URL https://www.jstatsoft.org/article/view/v074i11.

B. Nowok, G. M. Raab, J. Snoke, and C. Dibben. synthpop: Generating Synthetic Versions of Sensitive Microdata for Statistical Disclosure Control, 2016b. URL https://CRAN.R-project.org/package=synthpop. R package version 1.2-1.

J. Perla, C. Tonetti, and M. E. Waugh. Equilibrium Technology Diffusion, Trade, and Growth. American Economic Review, 111(1):73–128, Jan. 2021. ISSN 0002-8282. doi: 10.1257/aer.20151645. URL https://pubs.aeaweb.org/doi/10.1257/aer.20151645.

L. B. Reeder, J. C. Stanley, and L. Vilhuber. Codebook for the SIPP Synthetic Beta v7.0 [codebook file]. DDI-C document, Cornell Institute for Social and Economic Research and Labor Dynamics Institute [distributor]. Cornell University, Ithaca, NY, USA, 2018. URL http://www2.ncrn.cornell.edu/ced2ar-web/codebooks/ssb/v/v7.

J. P. Reiter, A. Oganian, and A. F. Karr. Verification servers: Enabling analysts to assess the quality of inferences from public use data. Computational statistics & data analysis, 53(4):1475–1482, Feb. 2009. ISSN 0167-9473. doi: 10.1016/j.csda.2008.10.006.