Using containers for analysis validation at scale

Lars Vilhuber

August 2024

Keywords: synthetic data; verification server; confidential data; reproducibility; validation

1 Overview

Researchers have, for the past 50 years, argued for greater access to the detailed but confidential data that statistical agencies collect and curate. From the 1965 proposal by the Social Science Research Council for "a federal data center, with public access for researchers" (Anderson, 2015, pg. 219) to later calls for similar expanded access (e.g. Card et al., 2010), these requests have been met by increasing options for providing such access (United Nations, 2007; Schouten and Cigrang, 2003; Weinberg et al., 2007; Cole et al., 2021). Public use data is just one of those dissemination mechanisms, but even when data are publicly available, researchers have regularly obtained access to confidential data to assess and verify the accuracy and reliability of public use data relative to alternate sources of data (to cite just a few, Larrimore et al., 2008; Armour et al., 2016; Alexander et al., 2010; Abraham et al., 2013, 2020). A steadily increasing number of physical or virtual access portals to confidential US data (through the Federal Statistical Research Data Center (FSRDC) or virtual enclaves like those provided by the Bureau of Labor Statistics (BLS) or the Economic Research Service of the USDA) provide researchers with access. Yet that access pales with the quantity of publications that use the public use data.[1]

One approach to improving access to confidential data is by creating reasonable "facsimiles" of confidential data, or "synthetic data." First proposed by Rubin (1993), various methods have been proposed since then (see Vilhuber et al., 2016; Raghunathan, 2021, for overviews). However, researchers, when faced with novel technology and datasets, are rightly suspicious of the data quality and appropriateness for their analyses. Abowd et al. (2006) describe a synthetic data file (SIPP Synthetic Beta (SSB)) which was made available through a publicly accessible server (described later), where researchers could prepare their analyses, and then submit these to the statistical agency for "validation."[2] Kinney et al. (2011) subsequently relied on the same mechanism for the Synthetic Longitudinal Business Database (LBD) (SynLBD).[3] Reiter et al. (2009) proposed an analogous idea of a "verification server" in a more general context with any public-use data, and Barrientos et al. (2018) expanded the concept to differentially private verification.

The provision of SSB and SynLBD as pilot projects was not meant to be scaled, and involved substantial learning on behalf of researchers, the statistical agency, and the scientific community more generally. I was involved from the start in setting up various iterations of the server, from the first (limited) version to make the SSB available around 2007, and maintaining them through 2022, when the last publicly accessible version was shut down (Vilhuber and Abowd, 2022).

Understanding researcher constraints, technologically feasible options, and using those to balance privacy choices and costs of access remain key, as various other presentations at this conference (Reiter and Park, 2024a; Raghunathan and

[1]Public use data are available through many sources, but Integrated Public Use Microdata Series (IPUMS) alone counts nearly 5,000 publications that use census, American Community Survey (ACS), and Current Population Survey (CPS) data in the past 5 years (as of 2024-08-14), which is almost surely an undercount of the overall usage. The Census Bureau's Center for Economic Studies' working paper series is the closest proxy for the number of publications that use data in the FSRDC, as it includes papers by both staff and researchers (though it does not include papers that use other agencies' data exclusively). It lists 241 working papers over the same time period, from usage of all data sources available within the FSRDC.

[2]See also Benedetto et al. (2013); U.S. Census Bureau (2015); Reeder et al. (2018) for additional details on the SSB.

[3]See (U.S. Census Bureau, 2011; Vilhuber, 2013; Kinney et al., 2014) for additional details.

Hotz, 2024) and the underlying National Academies of Science, Engineering, and Medecine (NASEM) panel reports (Raghunathan and Chaney, 2023; Reiter and Park, 2024b) attest to.

2 The Cornell Synthetic Data Server: History and Lessons Learned

When the SSB first became available, the need for users to access and use the synthetic data was a key part of the improvement plan. However, self-validation would have involved proposing projects to be conducted in the FSRDC (then still called the Census RDC), which at the time involved very long approval delays. Around 2007, with approval from the federal agencies involved, a first server was set up at Cornell University that was structured in such a way as to facilitate the preparation of statistical analyses using a (for the time, novel) graphical remote desktop. Researchers then notified Census Bureau staff, who transferred the code to a separate, secure agency computing system, and re-executed the code using the confidential data. If the code ran and produced output, staff verified compliance with disclosure avoidance rules in effect at the time, and subsequently released the results obtained with the confidential data to the researchers. This was called "validation."

Subsequently, National Science Foundation (NSF) funding was obtained,[4] and a new, more powerful server implemented to support both the SSB as well as the SynLBD, which would be released shortly thereafter (Kinney et al., 2011). Additional funding supported the server through an additional hardware upgrade and maintenance phase until 2022.[5] Usage statistics are available for the 2010-2015 time period, and depicted in Figure 1. They show a relatively steady (linear) increase in the number of registered users. User growth declined somewhat in the following years; by the end of the project, there were about 300 registered users.

[4]NSF grant SES-1042181.

[5]Additional funding came through NSF Grant BCS-0941226 and from the Alfred P. Sloan Foundation. Funding in the last years was provided through John Abowd's Edmund Ezra Day chair at Cornell University.

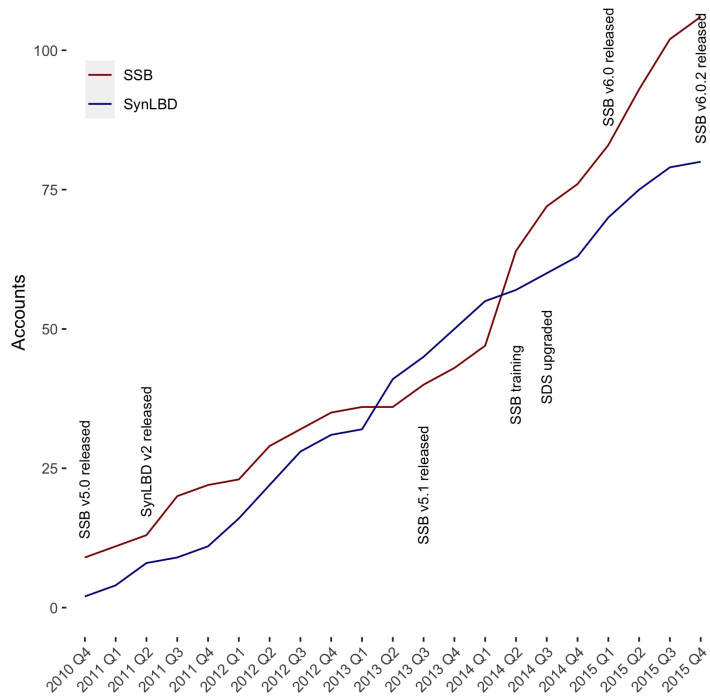

Figure 1: Computer accounts on the SDS over time

The development cycle was primarily active for the SSB. Launched with version 4.1 on the Synthetic Data Server (SDS) in 2010, updates were released in 2013 (v5.1), 2015 (v6.0), and 2018 (v7.0). The version of the SynLBD available throughout the time period was v2.0, though additional work to improve the SynLBD was undertaken (Kinney et al., 2014).

Several lessons emerged from the SDS mechanism. While many researchers used the data to write papers, and even organized conference sessions specifically around the use of the data,[6] even more researchers only "tried out" the data. Over 100 researchers were granted access to the server to access the SSB in the first five years of its availability (Figure 1), but far fewer published using the SSB data.[7]

[6]LERA session "Data Gold! Exploiting the Rich Research Potential of Lifetime Administrative Earnings Data Linked to the Census Bureau's Household SIPP Survey", at the Allied Social Sciences 2016 Annual Meeting (American Economic Association, 2016).

[7]All publications directly funded by the supporting NSF grant, or using the NSF-funded

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

Table 1: Distribution of Parameter-specific Confidence Interval Overlap, for selected projects

Almost none of the published articles actually used the results produced using the synthetic data. Comparison of parameters obtained from synthetic data and from confidential data using confidence interval overlap, a measure of congruence between the synthetic data and the confidential data introduced by Karr et al. (2006), was very heterogeneous even for a given dataset across and within projects (Table 1). A more recent assessment, presented as part of this same conference, finds generally similar findings (Carr et al., 2023; Stanley and Totty, 2023). Authors were rightly hesitant to use the parameters estimated on the synthetic data.

Thus, a core goal of the synthetic data — to replace the confidential data in researchers' analyses — was not being met, even when the synthetic data actually is a very good test dataset. Nevertheless, the synthetic data were complex enough to allow for development of models without access to the confidential data, what I would call "good enough data."

Anecdotal evidence from both my own and Census staff's attempts to use author-provided computer code to run the analysis on the confidential data demonstrated challenges in reproducibility. Authors might hard-code intermediate findings, rather than letting the data drive the analysis, and would otherwise not fully leverage the similarity between the two computing environments. These lead to

server, are listed at `https://www.zotero.org/groups/5595570/sds-nsf-1042181/library`. Some publications were prepared by NSF-funded project personnel and should not be directly included in a publication count of "users." Most publications were included in this list after a bibliographic full-text search for the grant identifiers. Some researchers may not have reported the published article to the project team, or mentioned the support of the grant to the server they used in their acknowledgements.

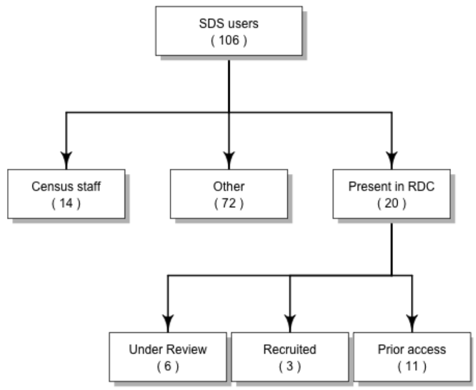

Figure 2: SDS users and access to FSRDC

time-intensive human debugging, or multiple rounds with authors, neither of which are an efficient and satisfying process.

More interestingly, multiple authors treated the synthetic data access as a gateway process for access to the confidential data. Knowing that the synthetic data did not contain all the features they needed for their analysis, but having to wait for permission to access the more detailed confidential data in the FSRDC, authors used the synthetic data to prepare analyses and explore the data. Figure 2 shows an analysis of the first 106 users of the SDS, and subsequent usage of the FSRDC.

Importantly, in the initial phase of the projects, turnaround (submission of validation request and receipt of validated and privacy-protected results) was quite fast - single-digit weeks, rather than the multi-month process of obtaining access to the FSRDC. However, the introduction of new disclosure avoidance procedures at the Census Bureau, and the lack of integration of those procedures into the validation process, greatly increased the time lag in the second half of the projects.

3 Scaling up access to confidential data via synthetic data

If data cannot be made available due to intractable disclosure avoidance issues, yet access should be broadened, what can agencies do? The first-order solution is, of course, to greatly accelerate the process of granting access to the confidential data, but a secondary problem — reviewing the output — may still bind, even if all the security vetting issues are solvable.

The pilot projects described earlier were not set up to scale. Statistical agencies and research institutes have explored various ways to scale up access to confidential data. To cite a few examples, Statistics Canada provides the Real-time Remote Access (RTRA) process, Norway has the Microdata.no system, the Bank of Portugal uses a two-stage system combining a remote desktop and validation (Guimarães, 2023), and Barrientos et al. (2018) proposed a differentially private verification server.

Many such processes have restrictions that limit their utility for researchers. The aforementioned Statistics Canada and microdata.no systems strongly limit the type of analysis that is feasible by restricting the software keywords that can be used (RTRA), by creating a structured new statistical language (microdata.no), or by limiting the types of analysis that can be run and validated (Barrientos et al., 2018). Users of the Bank of Portugal's system still need to use the remote desktop system, similar to the SDS outlined before, because the data hosted there is not authorized as a full public-use product.

The issue is compounded by well-documented problems with the reproducibility of code in the social sciences. Heuristically, many of the problems with the SDS arose because the code failed to reproduce during validation, even though it was run in a very similar environment to the development environment. Researchers in the social sciences appear to rely heavily on interactive computing, with code produced subsequently failing simple reproducibility tests. In a sample of over 8,000 replication packages associated with high-profile economics articles, only 30% had some sort of master script, allowing for "push-button" reproducibility.[8] While

[8] Code run in November 2023, searching for any filename that contained the strings

"push-button" reproducibility may be optional for a general replication package, it is a requirement for a scalable remote-submission system. In part to cater to researchers' demand for interactive systems, remote-access or local secure access in the form of physical or virtual secure data enclaves are still the dominant — but expensive — way to access confidential data. The dominant method of access thus forces researchers to choose between lower quality data in an environment that corresponds to their preferred computing method, and higher quality confidential data in environments that are expensive for researchers, data providers, or both.

3.1 Desiderata

Drawing on the experience from the SDS pilot projects and other remote access methods used in the past, as well as looking at newer technologies that have emerged in the last decade, I suggest that a new, scalable mechanism to provide access to confidential data should have the following desirable characteristics:

1. the mechanism must support arbitrary modeling approaches and ideally a large number of programming languages
2. the mechanism must allow for development of models by researchers that are close to their "normal" method of developing models
3. the mechanism must be low-cost for the data provider, scaling at best sub-linearly with the number of users of those datasets
4. the mechanism must be low-cost for the data user, imposing at best marginal costs on their existing research infrastructure (software, computers)
5. the privacy-protected data provided as part of such mechanisms must be good enough to allow for complex modeling
6. validation, if necessary, must be fast - on the order of hours

Note that public-use data files, as historically provided by statistical agencies, satisfy all of these criteria, except for the last one, which can take years. Should statistical agencies actually offer validation even for such public use data, as Reiter et al. (2009) have argued? Traditionally, they do not, and leave it up to individual

'main' or 'master', the most common name used for control code in economics.

researchers to "self-validate" by requesting access to confidential data in a time-consuming fashion.[9]

4 A Proposal using Containers

In VILHUBER-HDSR, I demonstrate a simple scenario that satisfies most of the desiderata, using containers. 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. Containers, often referred to using the name of a particular implementation by a commercial provider (Docker), are technology most often, but not exclusively associated with Linux, which enables computer processes and code libraries to be bundled and constrained.[10] In essence, a container bundles into a single file all the dependencies and code required to run an application or to conduct a researcher's statistical analysis. This file can then be run on any computer without (much) further ado. Containers can be hosted on a cloud platform, but can also run on researcher compute platforms (laptops).

Containers are well-understood in the computer science and statistics community (Boettiger, 2015; Moreau et al., 2023). However, acceptance in the economics community is not particularly widespread, so far. In the same 8,000 replication packages mentioned earlier, only 0.13% had used containers.

The use of containers in the context of synthetic data with validation is to provide users with access to data and coding resources such that their analysis is easily portable, and verifiably reproducible. Within containers, users can implement arbitrary methods of analysis in the statistical programming language of their choice, as most are compatible with containers, even those requiring licenses.[11] They will need to be aware of constraints imposed by the disclosure

[9] See Armour et al. (2016) for one example of such a project, affecting the widely-used CPS

[10] In the academic world, Singularity/Apptainer is another container technology typically used in high performance computing (HPC) environments.

[11] For a general example of containers for Stata, see Vilhuber (2024a). For a particularly complex example involving three different licensed software, see Vilhuber (2024b).

avoidance rules (Abowd and Schmutte, 2015), just as they would if accessing the confidential data directly, and as they should be if using public-use data. Crucially, containers can be checked for reproducibility before being forwarded to the confidential computing environment. Once determined to be reproducible, containers can then be extended to use confidential data, and enable a wide spectrum of plug-in disclosure avoidance measures. Crucially, all checks on reproducibility can be performed prior to validation using the confidential data, on open, possibly commercial platforms. Only once reproducibility is confirmed is the same analysis model ported to the confidential data.

Containers can be run both locally as well as on cloud infrastructure. The (potential) use of cloud providers removes the requirement for users of the synthetic data to install anything locally, and for statistical agencies to maintain such a public-facing infrastructure. Many academic environments provide some support for running containers, but crucially, academic support is not a requirement, potentially opening up the use of this mechanism to data journalists or citizen scientists. The open-source nature of the container technology allows users to do run containers themselves, when they want to, or when they have to. Thus, a container-based validation mechanism dramatically reduces the agency's cost of providing access to synthetic data. Containers not only allow for reproducible running of code, but are themselves reproducibly generated. This has favorable IT security implications, since no external software needs to be transferred, a regular problem point for security-conscious agencies. In fact, as I outline in VILHUBER-HDSR, the agency should be the entity defining the container image, exporting it to the public while maintaining a high security posture.

Once results have been generated, the usual disclosure avoidance workflow at the data provider is triggered. This might entail post-processing of the results, generation of additional supporting statistics (though these should generally be included in the processing), and finally, provision of the results to users. Scalability of a system as described here hinges critically on having streamlined output vetting. Ideally, this part must also be automated. At present, non-automation of output vetting is likely the single most important bottleneck of this system. However, the challenge of creating automated and reliable disclosure avoidance procedures is not unique to the validation process described here.

5 Conclusion

The use of containers ensures reproducibility, reliable portability, and enables scalability. The use of cloud-based commercial services requires no infrastructure or software maintenance by either data provider or users, but is not a necessary condition, as users can easily provide their own infrastructure. Crucially, this means that the cost to statistical agencies of providing users with compute resources is avoided. With very little effort, automation is possible (potentially through web forms), and the only likely constraint to full automation is the absence of automated output vetting algorithms. While containers are not the "normal" way of developing statistical models in economics, they are increasingly being used in statistics, and at the cutting edge of the social sciences.

Thus, containers satisfy most of the desiderata outlined earlier. They do need to rely on synthetic data with sufficient complexity, if not analytical validity, in order to allow for the interactive development of analyses, and a privacy-protection mechanism that can scale. If such a privacy-protection mechanism can be tuned to acceptable protection levels (on par with traditional mechanisms that are applied to unrestricted public-use products), then validation can be made highly automated, and the quality of the synthetic data itself can be decreased, while maintaining high levels of user acceptance due to a fast validation process.

Containers offer the promise of streamlining and improving indirect access to confidential data. As a currently under-utilized technology in the space of the federal statistical agencies, it may be a way to modernize and adapt the way that synthetic data and remote processing interact in a researcher-friendly way.

Disclosure Statement

The author have no conflicts of interest to declare. The mention of commercial entities is not meant to endorse any such providers, and the author holds no financial interest in any of the mentioned commercial entities.

Acknowledgments

I have benefited from discussions with many folks, including Gary Benedetto, John Abowd, Rob Sienkiewicz, and from feedback following presentations to the National Academies, Census Bureau, and at the NBER conference on "Data Privacy Protection and the Conduct of Applied Research." The original development of the idea was partially funded by Alfred P. Sloan Foundation Grant G-2015-13903. The Synthetic Data Server project received funding from NSF grant SES-1042181, NSF Grant BCS-0941226 and from the Alfred P. Sloan Foundation as well as the Edmund Ezra Day chair at Cornell University.

Contributions

LV conceived the topic, wrote the text, and prepared the examples.

References

Abowd, John M and Ian M Schmutte (2015) "Economic Analysis and Statistical Disclosure Limitation," *Brookings Papers on Economic Activity*, 50 (1), 221–267, 10.1353/eca.2016.0004.

Abowd, John M., Martha Stinson, and Gary Benedetto (2006) "Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project,"Technical report, U.S. Census Bureau, `http://hdl.handle.net/1813/43929`.

Abraham, Katharine G., John C. Haltiwanger, Claire Hou, Kristin Sandusky, and James R. Spletzer (2020) "Reconciling Survey and Administrative Measures of Self-Employment," *Journal of Labor Economics*, 10.1086/712187, Publisher: The University of Chicago Press.

Abraham, Katharine G., John Haltiwanger, Kristin Sandusky, and James R. Spletzer (2013) "Exploring Differences in Employment between Household and Establishment Data," *Journal of Labor Economics*, 31 (S1), S129–S172, 10.1086/669062.

Alexander, J. Trent, Michael Davern, and Betsey Stevenson (2010) "Inaccurate Age and Sex Data in the Census Pums Files: Evidence and Implications," *Public Opinion Quarterly*, 74 (3), 551–569, 10.1093/poq/nfq033.

American Economic Association (2016) "Allied Social Science Associations Program," Program 2016, American Economic Association, San Francisco, `https://assets.aeaweb.org/asset-server/files/815.pdf`.

Anderson, Margo J. (2015) *The American census: a social history*, New Haven: Yale University Press, second edition edition, OCLC: ocn904081530.

Armour, Philip, Richard V. Burkhauser, and Jeff Larrimore (2016) "USING THE PARETO DISTRIBUTION TO IMPROVE ESTIMATES OF TOPCODED EARNINGS," *Economic Inquiry*, 54 (2), 1263–1273, 10.1111/ecin.12299.

Barrientos, Andrés F., Alexander Bolton, Tom Balmat et al. (2018) "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*, `http://arxiv.org/abs/1705.07872`, arXiv: 1705.07872.

Benedetto, Gary, Martha Stinson, and John M. Abowd (2013) "The creation and use of the SIPP Synthetic Beta,"Technical report, US Census Bureau, `http://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/SSBdescribe_nontechnical.pdf`, tex.timestamp: 2015.02.11.

Boettiger, Carl (2015) "An introduction to Docker for reproducible research," *ACM SIGOPS Operating Systems Review*, 49 (1), 71–79, 10.1145/2723872.2723882.

Card, David E., Raj Chetty, Martin S. Feldstein, and Emmanuel Saez (2010) "Expanding Access to Administrative Data for Research in the United States," *SSRN Electronic Journal*, 10.2139/ssrn.1888586.

Carr, Michael D., Emily E. Wiemers, and Robert A. Moffitt (2023) "Using Synthetic Data to Estimate Earnings Dynamics: Evidence from the SIPP GSF and SIPP SSB," presentation, NBER.

Cole, Shawn, Iqbal Dhaliwal, Anja Sautmann, and Lars Vilhuber (2021) *Handbook on Using Administrative Data for Research and Evidence-based Policy*: Abdul Latif Jameel Poverty Action Lab, 10.31485/admindatahandbook.1.0, tex.ids= cole2020.

Guimarães, Paulo (2023) "Reproducibility With Confidential Data: The Experience of BPLIM," *Harvard Data Science Review*, 5 (3), 10.1162/99608f92.54a00239.

Karr, A. F., C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil (2006) "A framework for evaluating the utility of data altered to protect confidentiality," *The American Statistician*, 60 (3), 1–9, 10.1198/000313006X124640.

Kinney, Satkartar K., Jerome P. Reiter, and Javier Miranda (2014) "Improving The Synthetic Longitudinal Business Database," *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics*, 30 (2), 10.3233/SJI-140808.

Kinney, Satkartar K., Jerome P. Reiter, Arnold P. Reznek, Javier Miranda, Ron S. Jarmin, and John M. Abowd (2011) "Towards unrestricted public use business microdata: The synthetic longitudinal business database," *International Statistical Review*, 79 (3), 362–384, 10.1111/j.1751-5823.2011.00153.x, tex.owner: vilhuber tex.publisher: Blackwell Publishing Ltd tex.timestamp: 2012.09.04.

Larrimore, Jeff, Richard V. Burkhauser, Shuaizhang Feng, and Laura Zayatz (2008) "Consistent cell means for topcoded incomes in the public use march CPS (1976–2007)," *Journal of Economic and Social Measurement*, 33 (2-3), 89–128, 10.3233/JEM-2008-0299, Publisher: IOS Press.

Moreau, David, Kristina Wiebels, and Carl Boettiger (2023) "Containers for computational reproducibility," *Nature Reviews Methods Primers*, 3 (1), 1–16, 10.1038/s43586-023-00236-9, Number: 1 Publisher: Nature Publishing Group OA version: https://minio.carlboettiger.info/shared-data/Containers.pdf.

Raghunathan, Trivellore and Bradford Chaney eds. (2023) *A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation*, Washington, D.C.: National Academies Press, 10.17226/27169, Pages: 27169.

Raghunathan, Trivellore E. (2021) "Synthetic Data," *Annual Review of Statistics and Its Application*, 8 (1), null, 10.1146/annurev-statistics-040720-031848.

Raghunathan, Trivellore and V. Joseph Hotz (2024) "A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation SIPP," presentation, NBER.

Reeder, Lori B., Jordan C. Stanley, and Lars Vilhuber (2018) "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, `http://www2.ncrn.cornell.edu/ced2ar-web/codebooks/ssb/v/v7`.

Reiter, Jerome P, Anna Oganian, and Alan F Karr (2009) "Verification servers: Enabling analysts to assess the quality of inferences from public use data," *Computational statistics & data analysis*, 53 (4), 1475–1482, 10.1016/j.csda.2008.10.006.

Reiter, Jerome P. and Jennifer Park (2024a) "Toward a 21st Century National Data Infrastructure: Managing Privacy and Confidentiality Risks with Blended Data," presentation, NBER.

——— eds. (2024b) *Toward a 21st Century National Data Infrastructure: Managing Privacy and Confidentiality Risks with Blended Data*, Washington, D.C.: National Academies Press, 10.17226/27335, Pages: 27335.

Rubin, Donald B (1993) "Discussion: Statistical disclosure limitation," *Journal of Official Statistics*, 9 (2), 461–468, `https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/`

discussion-statistical-disclosure-limitation2.pdf, https://ecommons.cornell.edu/bitstreams/dd0b63ff-4494-4491-96ba-a69811563dee/download.

Schouten, Barry and Marc Cigrang (2003) "Remote access systems for statistical analysis of microdata,"Technical Report 03004, Statistics Netherlands, https://www.oecd.org/std/37502934.pdf.

Stanley, Jordan C. and Evan S. Totty (2023) "A Penny Synthesized is a Penny Earned? An Exploratory Analysis of Accuracy in the SIPP Synthetic Beta," presentation, NBER.

United Nations (2007) "Managing statistical confidentiality and microdata access - Principles and Guidelines of Good Practice,"Technical report, United Nations Economic Commission for Europe - Conference of European Statisticians, https://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf.

U.S. Census Bureau (2011) "Synthetic LBD Beta Version 2.0," [Computer file], Cornell University, Synthetic Data Server [distributor], Washington,DC and Ithaca, NY, USA, http://www2.vrdc.cornell.edu/news/data/lbd-synthetic-data/, Published: Computer file.

——— (2015) "SIPP Synthetic Beta Version 7.0," [Computer file], Cornell University, Synthetic Data Server [distributor], Washington,DC and Ithaca, NY, USA, http://www2.vrdc.cornell.edu/news/data/sipp-synthetic-beta-file/, Published: Computer file.

Vilhuber, Lars (2013) "Codebook for the Synthetic LBD Version 2.0 [Codebook file]," DDI-C document, Comprehensive Extensible Data Documentation and Access Repository (CED2AR), Cornell Institute for Social and Economic Research and Labor Dynamics Institute [distributor]. Cornell University, Ithaca, NY, USA, http://www2.ncrn.cornell.edu/ced2ar-web/codebooks/synlbd/v/v2.

——— (2024a) "AEADataEditor/docker-stata," repository, Github, `https://github.com/AEADataEditor/docker-stata`.

——— (2024b) "AEADataEditor/docker-aer-2023-0700: Docker image for software stack used in AER-2023-0700," April, 10.5281/ZENODO.11080718.

Vilhuber, Lars and John M. Abowd (2022) "End of life for the Cornell Synthetic Data Server September 30, 2022," blog post, Cornell University, `https://web.archive.org/web/20230602202220/https://web.archive.org/web/20221130032540/https://www2.vrdc.cornell.edu/news/`.

Vilhuber, Lars, John M. Abowd, and Jerome P. Reiter (2016) "Synthetic establishment microdata around the world," *Statistical Journal of the IAOS*, 32 (1), 65–68, 10.3233/SJI-160964, tex.owner: vilhuber tex.publisher: IOS Press tex.timestamp: 2016.09.30.

Weinberg, Daniel H, John M Abowd, Philip M Steel, Laura Zayatz, and Sandra K Rowland (2007) "Access Methods for United States Microdata,"Technical Report 07-25, Center for Economic Studies, U.S. Census Bureau, 10.2139/ssrn.1015374.