

Democratizing Innovation Microdata:

How Synthetic Public Use Files Can Address the Credibility Crisis in Applied Economics

Jorge Cisneros¹, Timothy Wojan², Matt Williams³, Jennifer Ozawa³, Robert Chew³, Kimberly Janda³, Timothy Navarro³, Michael Floyd⁴, DJ Streat⁴, Audrey Kindlon⁵, Heather Madray⁵

¹Oak Ridge Institute for Science and Education, ²Independent Consultant, ³RTI International, ⁴Knexus Research LLC, ⁵National Center for Science and Engineering Statistics, National Science Foundation

Federal Context for Public Data Use

The **U.S. Evidence Act (2018)** and the **CHIPS and Science Act (2022)** promote greater access to federal data while protecting sensitive information, including the **National Secure Data Service Demonstration (NSDS-D)** initiative.

Public-Use Microdata Samples (PUMS) broaden access, support more timely analysis, and reduce reliance on secure data centers. However, PUMS for business surveys are limited, and older releases such as the **2007 Survey of Business Owners (SBO) PUMS** no longer meet modern privacy standards.

How do we develop safer PUMS that remain analytically valuable?

Annual Business Survey (ABS)

- ABS** collects data on innovation, globalization, owner characteristics, and R&D for U.S. nonfarm, for-profit businesses
- Mandatory NCSES–Census Bureau** survey covering ~5 million firms; **2023 sample: 850K** (other years ~300K), with full coverage of research-intensive industries
- Microdata access limited to FSRDCs**

Synthetic Data

- Synthetic data are artificially generated to preserve the statistical properties of real data while protecting privacy.
- Sequential decision trees (e.g., CenSyn, R SynthPop)** are a popular and effective synthetic generation method for tabular data.
- More than synthesis:** evaluation, privacy checks, and consistency checks; handles weights, partial synthesis, etc.

Acknowledgements:

Funding for this work was provided through: NCSES: ABS Syn project through America’s Datahub Consortium (Knexus and RTI) ORISE: NCSES Research Ambassadors Program (Cisneros, Wojan) NSF/ITE: Designing Public Use Microdata Business Areas (PUMBAs) for Innovation # 2430732 (Williams, Ozawa)

This presentation provides results of exploratory research sponsored by the National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation (NSF). This information is being shared to inform interested parties of ongoing activities and to encourage further discussion. Any views expressed are those of the authors and not necessarily those of NCSES or NSF.

This product has been reviewed for unauthorized disclosure of confidential information under NCSES-DRN25-019T.

2022 ABS Demonstrations*

Commercialization

Variable	CenSyn	SynthPop	ABS 2022	Agree with truth?
Styling (2)	-0.36**	-0.18	-0.52**	Censyn
Functional (3)	0.19	0.31**	0.51**	SynthPop
Strategic (4)	0.07	-0.58**	0.49**	Neither
Random (State)	0.49**	0.61**	0.35**	Both

** indicates that less than 5% of the posterior distribution is opposite the estimated sign.

Geographic Variation

	Styling (1)			Functional (2)			Strategic (3,4)		
RUCC	CenSyn	synthpop	ABS	CenSyn	synthpop	ABS	CenSyn	synthpop	ABS
Large Metro	14.5	14.3	14.3	11.8	11.5	11.9	6.9	6.6	6.8
Medium Metro	14.4	15.6	14.1	11.1	13.1	11.0	6.6	7.0	6.4
Small Metro	13.3	14.7	13.3	10.0	12.0	9.9	6.0	6.1	5.7
Large Rural Adjacent	12.0	13.4	11.7	8.6	11.2	8.3	4.8	6.3	4.6
Large Rural Nonadjacent	12.8	14.3	11.5	8.5	11.1	7.8	4.9	5.6	4.4
Medium Rural Adj.	11.7	13.4	11.8	6.9	10.6	7.1	4.2	5.4	4.2
Medium Rural Nadj.	12.7	16.0	12.2	7.9	13.8	7.2	4.8	6.8	4.6
Completely Rural Adj.	11.7	14.8	10.8	6.4	9.8	6.4	2.8	5.2	3.5
Completely Rural Nadj.	9.2	12.0	9.2	6.0	9.4	6.5	3.4	5.1	2.9

Percentages are rounded to one decimal place.

Commercialization

Test for association between venture capital funding and integration of design for R&D performing micro-businesses

Geographic Variation

Assessing the synthetic distribution of design across the gradient of urban/rural county status

Construct Validity

Ordered logit corresponding to the styling/functional/strategic rungs of the design ladder

Construct Validity

Variable	CenSyn	SynthPop	ABS 2022	Agree?
Resources for Design	-	0.52	0.56	Yes
Design Complexity IP	0.63	0.46	0.71	Yes
Design Patent IP	0.51	0.15	0.26	Yes
Designers	0.54	0.44	0.48	Yes
Company Leadership	0.55	0.42	0.58	Yes
Other Occupations	-0.62	-0.51	-0.80	Yes
Work Independently	-	-0.32	-0.55	Yes

All frequentist estimates significant at the 0.01 level.

* Census Disclosure Review Board Approval Number CBDRB-FY26-007.

Replicability

A key debate in economics is whether specifications be **finalized prior to estimation** or should one learn from the data first—**let the data speak?** (Miguel et al. 2014). Hybrid approaches use holdout samples and newer methods treat exploratory results as weakly informative priors. An open question is whether **synthetic microdata** can serve as a **holdout** sample (Dwork et al., 2015) .

Our approach supports new hypothesis testing using synthetic data by providing:

- Informed specifications for pre-analysis plans,
- Weakly informative priors for modeling

Summary

The **NSDS Demonstration Project** aims to expand access to federal data to improve research and policy. The **Synthetic ABS project** engages users to understand how synthetic public microdata can complement other access tiers, while balancing privacy and usability.

This work focuses on:

- increasing the number and quality of ABS projects approved through the **FSRDC Standard Application Process**, and
- demonstrating how a fully specified **pre-analysis plan** can be incorporated using a dual-method approach.

References:

Anderson, M. L., & Magruder, J. (2017). Split-sample strategies for avoiding false discoveries (No. w23544). National Bureau of Economic Research.
Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. Chapman and Hall/CRC. <https://doi.org/10.1201/9781315139470>
Cisneros, J., Wojan, T., Williams, M., Ozawa, J., Chew, R., Janda, K., Navarro, T., Floyd, M., Task, C. & Streat, D. (2025). Developing synthetic microdata through machine learning for firm-level business surveys. arXiv preprint arXiv:2512.05948.
Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. Science, 349(6248), 636-638.
Galindo-Rueda, F. and V. Millot (2015), "Measuring Design and its Role in Innovation", OECD Science, Technology and Industry Working Papers, No. 2015/01, OECD Publishing, Paris, <https://doi.org/10.1787/5js7p6lj6zq6-en>.
Kinney, S. K., Reiter, J. P., Reznick, A. P., Miranda, J., Jarmin, R. S., & Abowd, J. M. (2011). Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. International Statistical Review, 79(3), 362–384.
Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., ... & Van der Laan, M. (2014). Promoting transparency in social science research. Science, 343(6166), 30-31.
National Center for Science and Engineering Statistics (NCSES). 2025. The Different Approaches to Design by U.S. Businesses. InfoBrief NCSES 25-330. Alexandria, VA: National Science Foundation. Available at <https://www.nsf.gov/statistics/2025/ncses25330/>
Rep. Ryan, P. D. [R-W.-1. (2019, January 14). H.R.4174 - 115th Congress (2017-2018): Foundations for Evidence-Based Policymaking Act of 2018 (2017-10-31) [Legislation]. <https://www.congress.gov/bills/115th-congress/house-bill/4174>
Rep. Ryan, T. [D-O.-13. (2022, August 9). H.R.4346 - 117th Congress (2021-2022): Chips and Science Act (2021-07-01) [Legislation]. <https://www.congress.gov/bills/117th-congress/house-bill/4346>
Wojan, T. R., & Lambert, D. M. (2025). A novel framework for increasing research transparency: Exploring the connection between diversity and innovation. PloS one, 20(1), e0313826.

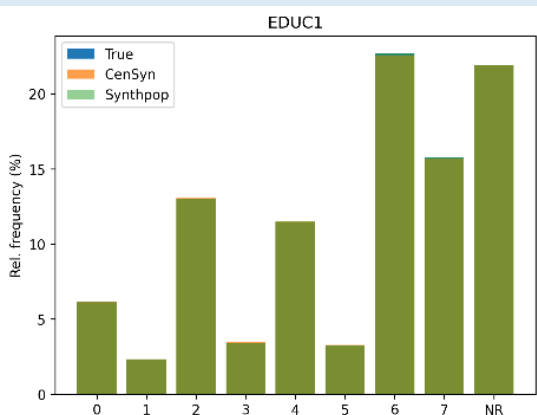
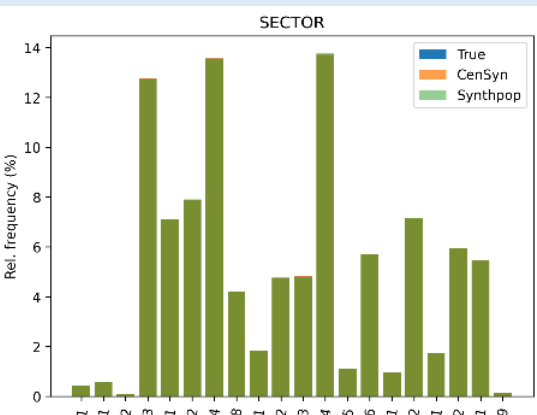
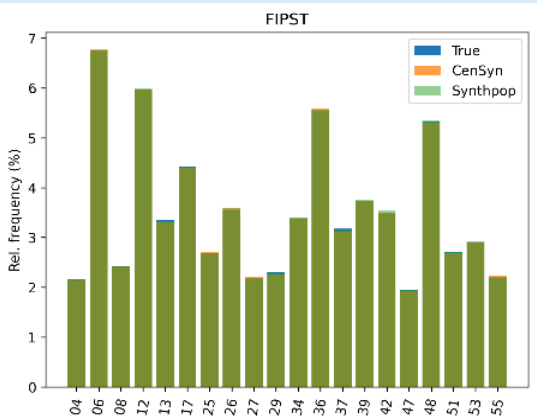
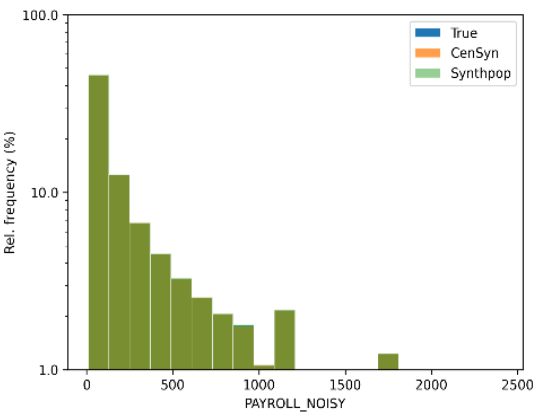
2007 SBO Demonstrations

Cisneros et. al (2025) demonstrated close alignment between the synthetic (CenSyn and Synthpop) and original SBO public data

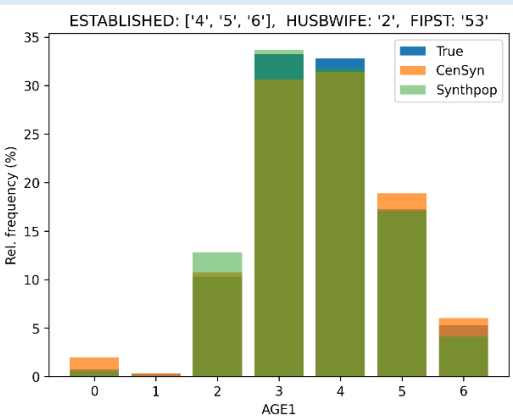
- Marginal distributions match almost identically (left)
- Larger differences appear for less frequent crossing of variables (right)

Cisneros et. al (2025) also replicated results from Wang and Liu (2015) using synthetic versions of the 2007 SBO PUMS.

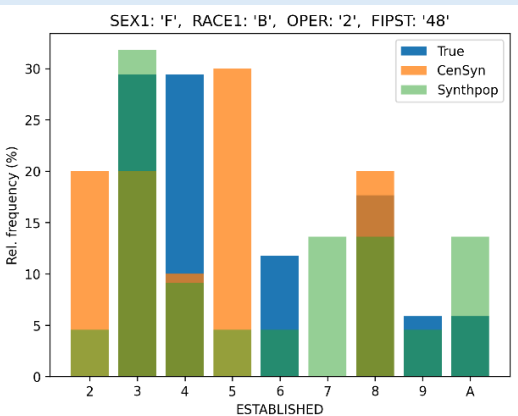
- Whether firms owned by immigrants are more likely to be engaged in transnational activities of operating a branch overseas, exporting, and foreign outsourcing.



Utility degrades as more variables are conditioned, improving privacy of rare subgroups.



Age of first owner in Washington established between 2000 – 2005, jointly owned with a spouse (**three variables**)



Texas businesses closed under female-black ownership (**four variables**)