Capturing the Annual Business Survey in Synthetic Microdata: Construction and Use Cases of a Public Use File

Jorge Cisneros Paz[1], Audrey Kindlon[2], Timothy Wojan[1], Matt Williams[3], Jennifer Ozawa[3], Christine Task[4], DJ Streat[4]

[1] Oak Ridge Institute for Science and Education, [2] National Center for Science and Engineering Statistics, [3] RTI International, [4] Knexus Research Corporation

2024 FCSM Research and Policy Conference
College Park Marriott Hotel & Conference Center, Hyattsville, MD
22–24 October 2024

NATIONAL CENTER FOR SCIENCE AND ENGINEERING STATISTICS
NATIONAL SCIENCE FOUNDATION

Disclaimer

This presentation provides results of exploratory research sponsored by the National Center for Science and Engineering Statistics (NCSES) within the 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-DRN24-031.

Annual Business Survey

- The **Annual Business Survey** (ABS) collects data on innovation, globalization, and business owner characteristics for nonfarm, for-profit businesses in the United States and measures R&D expenditures for the subset of those businesses with one to nine employees
    - ABS is a mandatory survey conducted by NCSES in partnership with the Census Bureau

- ABS frame includes about 5 million businesses engaged in mining, utilities, construction, manufacturing, wholesale trade, retail trade, or services industries
    - ABS 2023 sample = **850,000** while other ABS years are approximately 300,000
    - Sample includes all companies classified in selected research intensive industries, such as scientific R&D services

- Currently ABS microdata is available only in **Federal Statistical Research Data Centers** (FSRDCs)

Public-Use Microdata Sample

Public-use microdata samples (PUMS) play a critical role in meeting demand for public access to government-funded data and are essential for exploratory research and decision-making

- Enable statistical organizations to **reach a wider public** and **support more use cases** in a timely manner
- Less reliance on FSRDCs for high-level analyses
- United Nations Economic Commission for Europe (UNECE) *Synthetic Data for Official Statistics (2022)*: **transparency** and **increased public access** to vital data
- Limited PUMS for business surveys
  - Survey of Business Owners (SBO) 2007 PUMS still requested: not private enough for today's standards and modern attacks
- **Challenge:** making a safer PUMS with analytically interesting data

Business **Public-Use Microdata Sample**

Challenges of developing and releasing PUMS file for business survey data

- Fewer firms than people, but differences across firms are much larger than across people (a zoo of different species versus a herd)
- Business data is **highly skewed** with a small share of firms accounting for a large share of employment, sales, etc.
- Collecting complete feature information for every firm rapidly becomes difficult to provide in a way that ensures good and equitable utility for all users
- Privacy can be breached when "secured" data can **be linked with publicly available data**
- Numerous **industry** and **geographic features** of interest

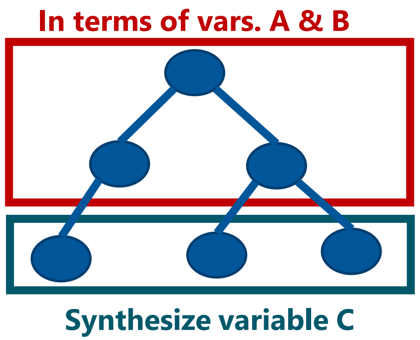
Synthetic Data

Synthetic data is artificially generated data that represents statistical properties of real data, while ensuring privacy and confidentiality of sensitive information

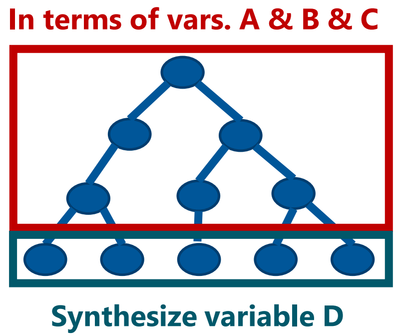
- **Preserves privacy** by releasing realistic data without exposing sensitive information, mitigating the risk of privacy breaches
- Maintains statistical relationships of original data, enabling **high utility** and meaningful analyses for preserved variables
- Privacy-preserving solution while maintaining data utility and integrity
  - US Census Bureau & IRS: Synthetic Longitudinal Business Database (2011, experimental)
- By learning and replicating patterns and structures within original data, AI-generated synthetic data ensures that key correlations and distributions are preserved

CenSyn Synthetic Data Generator

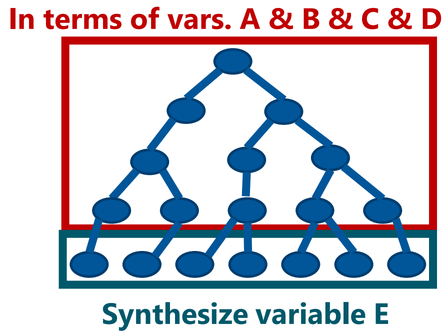
- **Classification and Regression Trees (CART)** is a straightforward data modeling method that predicts value of variable based on values of other variables using decision trees



- Different groups in data are channeled down different paths in a tree, as the **tree learns how best to partition population** into low-entropy and self-similar groups of firms with respect to target variable
  - To synthesize a single variable, we could use one decision tree



- To synthesize many variables, we use sequence of many decision trees
  - Algorithm that underlies **CenSyn Synthetic Data Generator** developed by Knexus



CenSyn Synthetic Data Generator

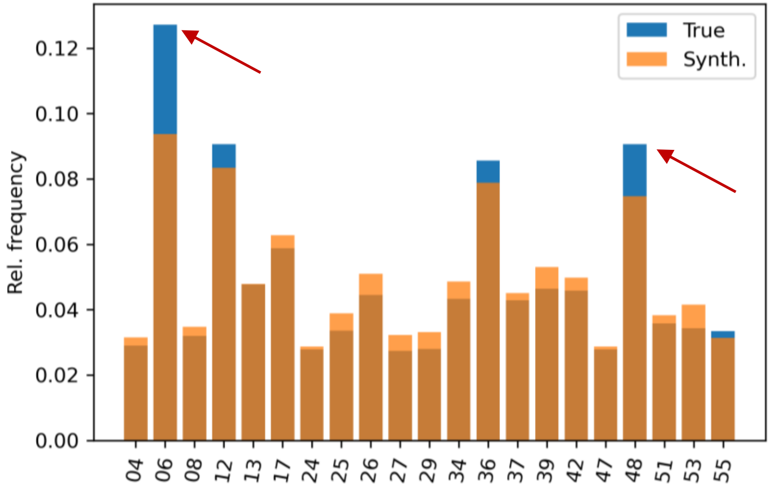
- Several other tree-based synthesizers exist, also built by or for statistical agencies:
  - **R synthpop library** for Scottish Longitudinal Survey
  - R tidysynthesis by Urban Institute for US IRS

- CenSyn was built for US Census Bureau projects
  - Configured for multiple groups and data products
  - Capabilities to deal with complexities that arise at scale of Census Bureau

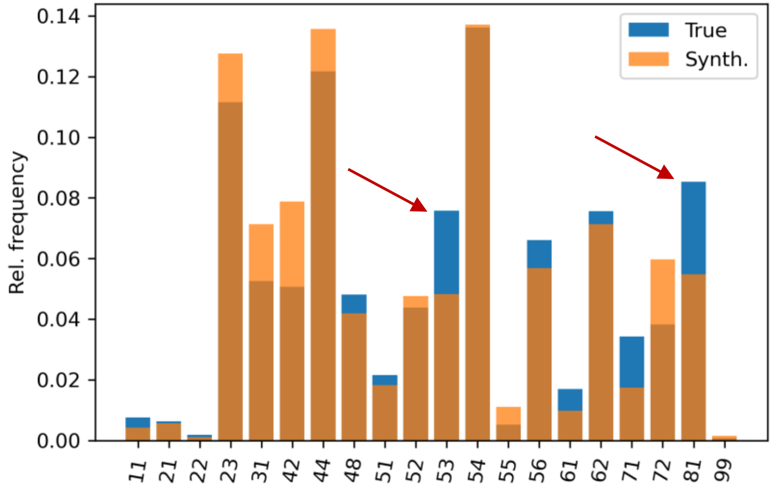
- Efficiently performs synthesis, evaluation, privacy checks, and consistency checks; handles weights, partial synthesis, etc.
  - Preserves distribution of data, in all its diversity, while also ensuring any real record is difficult to find in released product
  - Iterative nature with stakeholders and survey managers

Comparisons on SBO 2007 PUMS

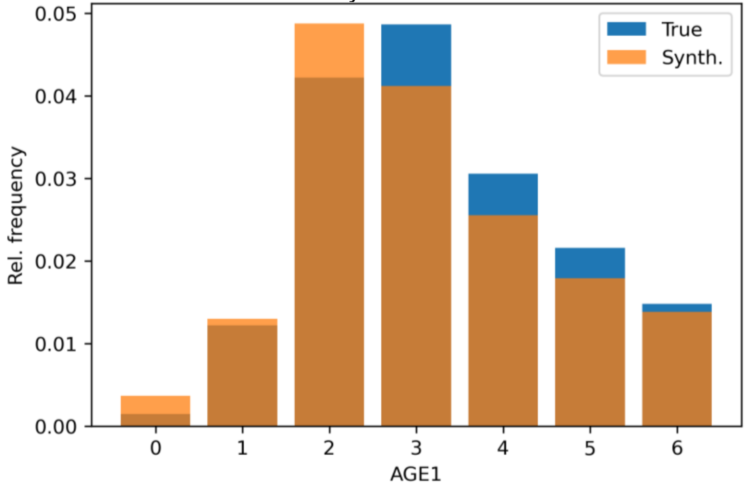
US States



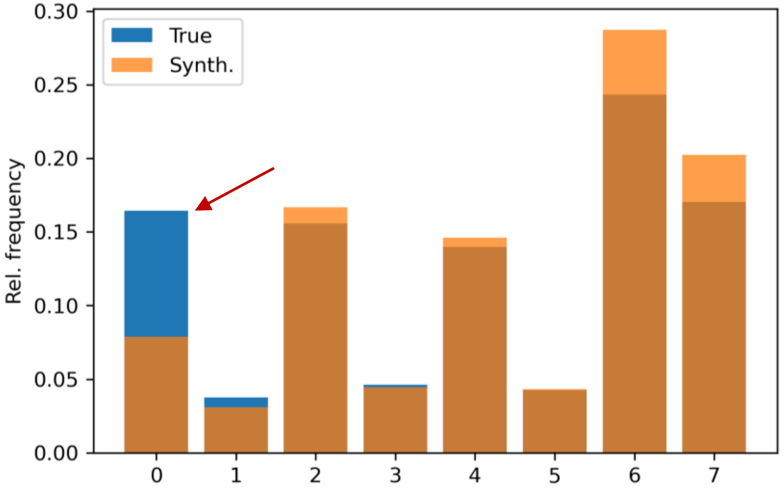
Sector (2-digit NAICS)



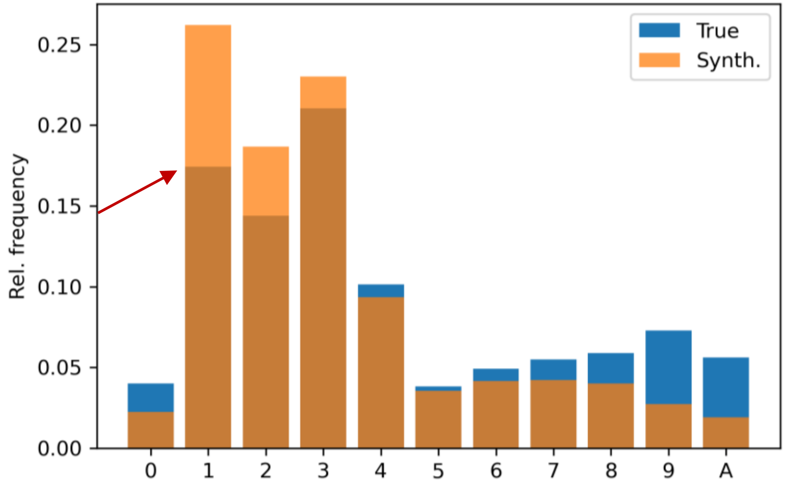
WA, 2000-05, jointly owned with spouse and primarily owned by husband



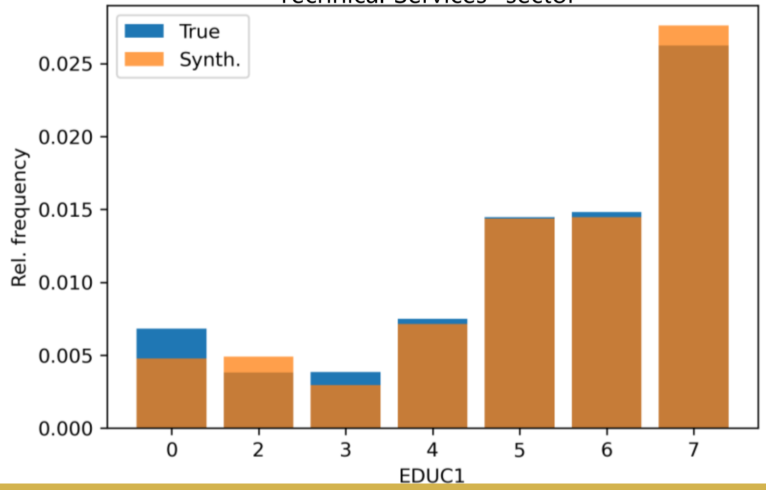
1st Owner Education



Year Business Established



OH, startup capital $5,000-10,000, “Professional, Scientific, & Technical Services" sector



ABS User Workshop

Specific objectives

- Develop a prioritized list of variables that users and potential users find important in academia, industry, government agencies, etc.
- Explore how users would use a PUMS and determine if a synthetic PUMS could meet needs

Target audiences

- Group 1 (*current* ABS users): Have completed the FSRDC process and currently have access to ABS restricted-use data
- Group 2 (*potential* ABS users): Researchers who do not have access to ABS restricted-use data but have expressed interest or who are current users of existing ABS public-use tables

Findings

- Both groups emphasized importance of preserving key features, such as business owner demographics, geographic information, and innovation activity
- Consensus on the utility of synthetic PUMS for generating descriptive statistics, conducting preliminary analyses, and assessing research feasibility

Conclusion

- Demonstrated difficulties that arise in preserving **privacy, equity, and utility** when working with complex feature sets
- Introduced idea of using models to better capture feature correlations
  - **CenSyn** operates well with large, complex national statistical data
  - Models tuned to data have potential to collect "right" information from different groups to preserve full, diverse data distribution (without releasing real establishment records)
- Having synthetic PUMS does not solve all problems
  - **Geography + industry** exposes firms: membership and attribute disclosure concerns
  - Industry + salient place characteristics (rural, affluent, disadvantaged areas) may work
- ABS User Workshop input and feedback
- **Next steps:** longitudinal data, quality metrics, clustering industry and geography variables

Thank you! Questions?

Jorge Cisneros Paz
jcisnero@associates.nsf.gov

Audrey Kindlon
akindlon@nsf.gov

Timothy Wojan
twojan@nsf.gov

Heather Madray
hmadray@nsf.gov

Matthew Williams
mrwilliams@rti.org

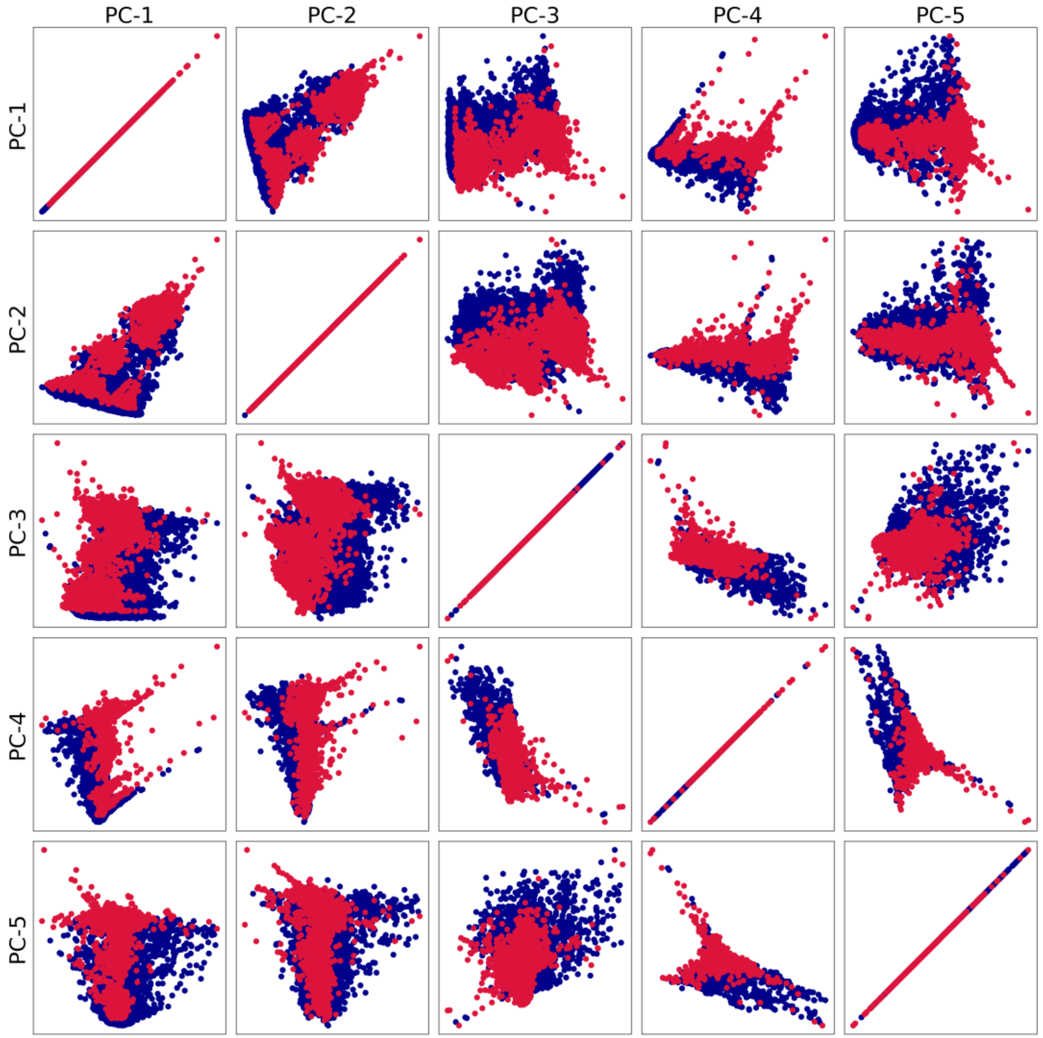
Jennifer Ozawa
jozawa@rti.org

Christine Task
christine.task@knexusresearch.com

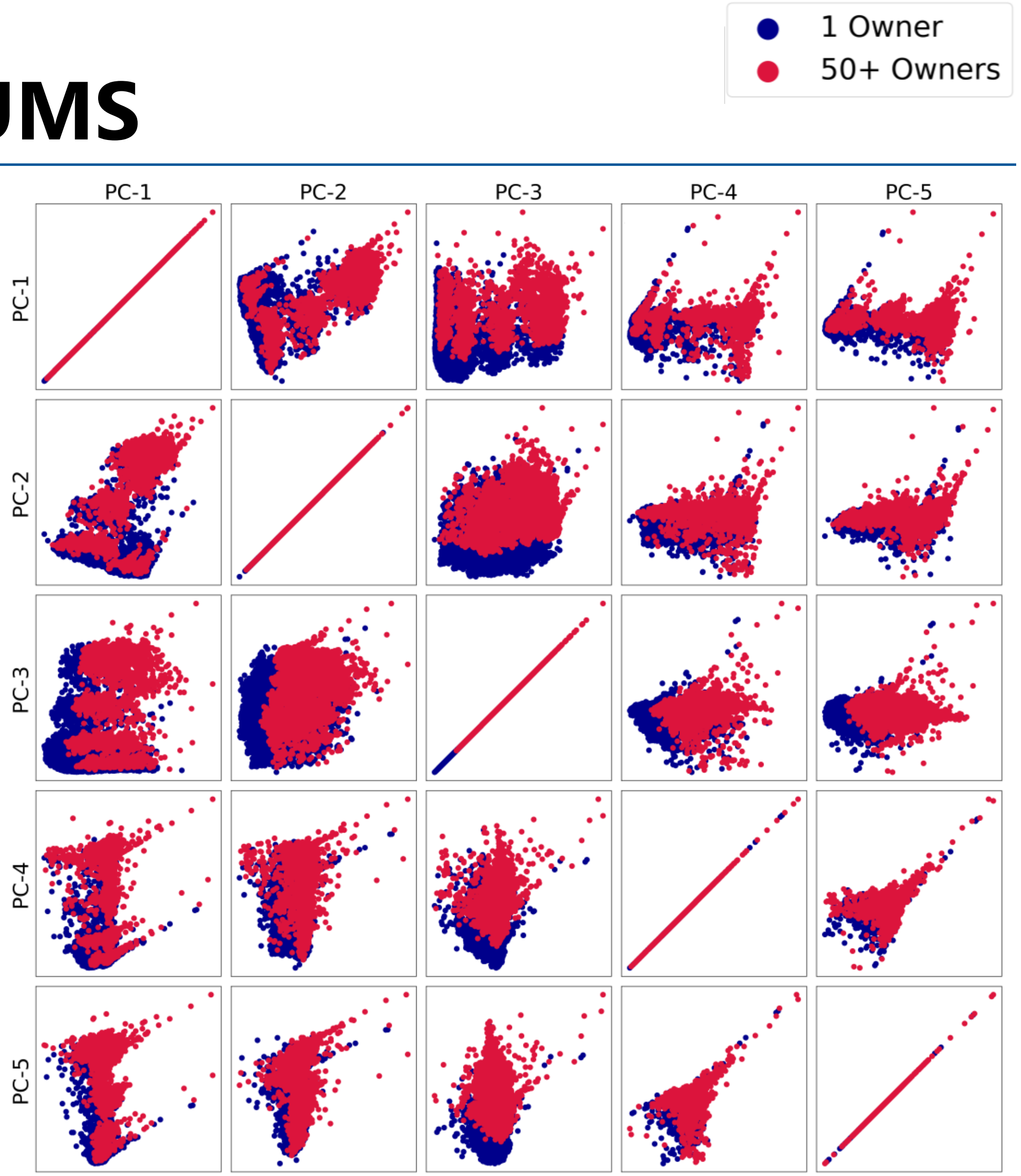
DJ Streat
dj.streat@knexusresearch.com

Additional Slides

Comparisons on SBO 2007 PUMS



Original data distribution shape



De-identified data distribution shape

Group 1 (Current ABS Users) Findings

- **Utility prior to FSRDC access:** Help researchers understand how the data are organized, run pilot analyses, publish more descriptive statistics; many researchers link ABS to Longitudinal Business Database, Business Enterprise Research & Development Survey, and SEC data, so synthetic data would need to include these linkages

- **Variables of interest:** All technology modules, innovation incidence, business owner demographics, startup capital

- **Concerns about using a synthetic PUMS:** Mixed; some researchers said they were wary of publishing based on synthetic data, while others did not think it would be a problem; would want to know what characteristics went into the synthetic data and validation tests

Group 2 (potential ABS users) Findings

- **Variables of interest:** Analyzing innovation outputs by geography (urban, suburban, rural), by industry, by business owner demographics

- **Use of ABS PUMS:** Make tabulations, develop descriptive statistics, and develop estimates to respond to policymaker questions, linkage with other data sets

- **Concerns about using a synthetic PUMS:** Few reports; most trust statistical methods used by NCSES and the Census Bureau