DEMOCRATIZING INNOVATION MICRODATA: HOW SYNTHETIC PUBLIC USE FILES CAN ADDRESS THE CREDIBILITY CRISIS IN APPLIED ECONOMICS

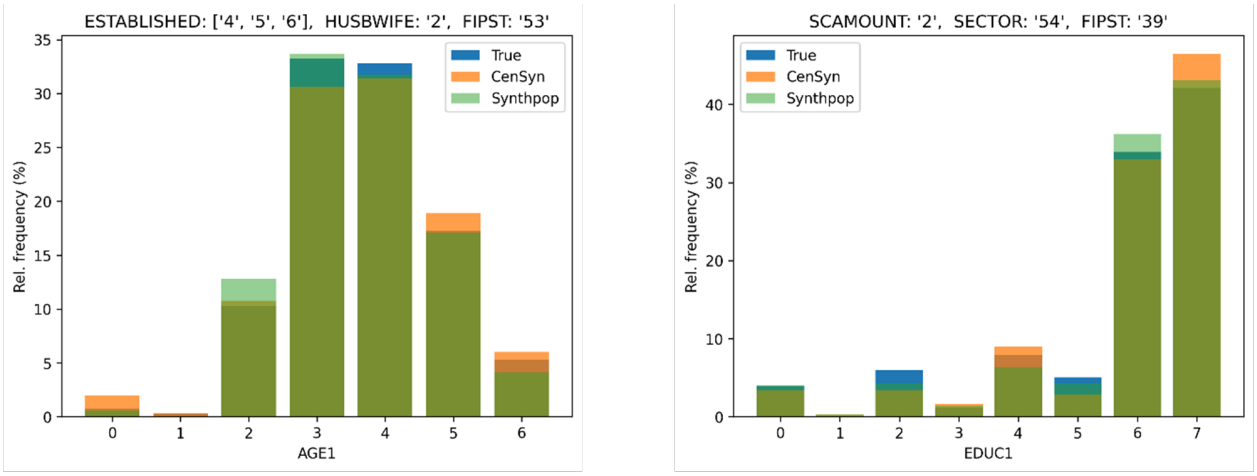
Audrey Kindlon[1], *Jorge Cisneros*[2], *Timothy Wojan*[3], *Matt Williams*[4], *Jennifer Ozawa*[4], *Robert Chew*[4], *Kimberly Janda*[4], *Timothy Navarro*[4], *Michael Floyd*[5], *DJ Streat*[5], *Heather Madray*[1]

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

Advances in data science have greatly increased disclosure risk for public use microdata samples (PUMS) that are a staple of applied economic analysis. The problem is most severe for releases of establishment microdata, as numerous public and proprietary datasets with identifying information increase the re-identification risk. Synthetic data generation processes applied to confidential data can produce accurate facsimiles that retain the moments of the original data but contain only synthetic firms. The objective of this research project is to produce synthetic public use files of the 2022 U.S. Census Annual Business Survey data using the CenSyn and R Synthpop synthetic data generation software. This will allow exploring phenomena related to innovation and R&D outside of the restricted Federal Statistical Research Data Center (FSRDC) system. In addition to increased data access for innovation researchers, synthetic data also has the potential to greatly increase the quality of research that uses FSRDC data. Using synthetic versions of ABS for exploratory specification testing preserves the true data for *de novo* confirmatory hypothesis testing, allowing researchers to avoid data dependent analysis where the same data are used for specification testing and hypothesis testing. This greatly reduces the probability of false discovery that is the main contributor to the credibility crisis in applied economics. In addition to providing rich information for a fully specified pre-analysis plan, synthetic data analysis also provides weakly informative priors needed for specifying Bayesian models in the confirmatory stage. The shrinkage in standard errors combined with the ability to directly estimate Pr(effect | sample) rather than the conventional frequentist test of Pr(sample | null) increases statistical power as well as provides much richer inferences. Use cases of various innovation phenomena demonstrate the increase in statistical power and the added value of probabilities assigned to coefficient estimate magnitudes.

As a precursor to this work, Cisneros et al. [1] present two synthetic PUMS developed for the 2007 Survey of Business Owners (SBO), similar to the ABS business data. Cisneros et al. [1] demonstrated close distributional alignment between the synthetic and original SBO public use file, matching the marginal distributions almost identically, with differences appearing for less frequent crossing of variables (Figure 1). In addition, the authors replicated results from "Transnational activities of immigrant-owned firms and their performances in the USA," a highly impactful article by Wang and Liu [3] using synthetic versions of the 2007 SBO PUMS. The central question investigated is whether firms owned by immigrants are more likely to be engaged in the transnational activities of operating a branch overseas, exporting, and foreign outsourcing. The economic impacts of transnational activities are also investigated with respect to sales per employee and payroll per employee. This provided several hard tests. First, the analysis focuses on relatively rare phenomena such as firms with operations overseas, firms that export, and firms that outsource. Second, the estimation includes industry fixed effects that expand the number of coefficient estimates that may differ in sign and significance from the true data. And third, the analysis includes both binary and continuous dependent variables estimated with logistic regression and ordinary least squares, respectively.

Similar to SBO synthesis by Cisneros et al. [1], we utilize CART-based synthesizers (the R *synthpop* library and CenSyn, developed by Knexus Research for and in collaboration with the US Census Bureau) to create two synthetic ABS 2022 micro-data files. In addition to the evaluation metrics built into the two libraries [2], we also perform a relevant econometric investigation. Unlike the SBO, the ABS now collects data on a variety of dimensions measuring innovation. The results we will present focus on questions such as (1) testing the construct validity of non-technological innovation (approach to design) and (2) associations of approaches to design with commercialization success/failure. Not only are we interested in the conclusions, but also in the resilience of the analysis to use of original vs. synthetic ABS as well as the role of each in the scientific analysis pipeline.



(a) (b)

Figure 1. Bar plots for inter-feature distributions: (a) the age of the first owner of firms in Washington that were established between 2000 – 2005 and that are jointly owned with a spouse, but primarily by the husband (i.e. the age of the husband), (b) the education level of the first owner of firms with startup capital between \$5,000 – \$10,000 involved in the "Professional, Scientific, & Technical Services" sector in Ohio. Note that the relative frequency is the percentage of the counts of occurrences of each response of the main feature under the imposed conditions, relative to all firms that meet the imposed conditions.

References

[1] Jorge Cisneros et al. "Developing synthetic microdata through machine learning for firm-level business surveys". In: *arXiv preprint arXiv:2512.05948* (2025).

[2] Aniruddha Sen et al. "Diverse community data for benchmarking data privacy algorithms". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 51409–51420.

[3] Qingfang Wang and Cathy Yang Liu. "Transnational activities of immigrant-owned firms and their performances in the USA". In: *Small Business Economics* 44.2 (2015), pp. 345–359.