ATTACHMENT I – PROJECT TOPIC

Creation of Synthetic Data for the Annual Business Survey (ABS) and Use of Verification Metrics

Key Objective

The objective of this project is to test and compare methods for creating synthetic data to support a tiered access model; explore the use of synthetic data for evidence-building; and test the use of verification metrics in validating estimates produced from synthetic data. The Annual Business Survey (ABS) is an annual survey conducted by the U.S. Census Bureau on behalf of multiple sponsors, one of whom is the National Center for Science and Engineering Statistics. One objective in disseminating ABS data is the production of a public use microdata file that could be used by researchers, policy makers, and public-data users. To address disclosure concerns, a synthetic datafile option is being explored. This project would build on this work to inform the National Secure Data Service Demonstration Project (NSDS-D) and support tiered access through the development of a publicly available dataset that could be used without the barriers involved in accessing the restricted-use data.

Background

America's DataHub Consortium brings together capabilities and infrastructure to securely fill information gaps and to take on key analytic questions and evidence building challenges. As demand for access to confidential federal data assets increases alongside novel analytical approaches, privacy protections must be in place to ensure the protection of privacy and the confidentiality of the data. Use of synthetic data can reduce disclosure risk while allowing data users to access microdata for research and other statistical purposes. A synthetic dataset does not contain the exact records of the original dataset, but instead retains the statistical properties of the original dataset, preserving information useful to researchers and their queries. The anonymity of the original dataset is not compromised since synthetic records do not correspond to real ones. The production and use of synthetic data for the ABS is being explored to increase the utility of the data while ensuring strong privacy protections.

The ABS includes, but is not limited to, data on business owners, including demographic information, research and development (R&D), innovation, and technology. These data are highly valuable in a wide range of areas including, but not limited to, informing the evaluation of programs that target businesses; assessing minority-owned businesses by industry and area and to educate industry associations, corporations, and government entities; analyzing business operations in comparison to similar firms; comparing R&D costs across industries; and determining where R&D activity is conducted geographically and in which types of businesses.

Currently, a restricted-use version of the ABS microdata is available for researchers who are approved for use by the Census Bureau, the Internal Revenue Service, and NCSES. In addition to the initial approvals, researchers must obtain Special Sworn Status, take annual data stewardship training, and use the data in a Federal Statistical Research Data Center (FSRDC). The administrative costs for government in managing projects and researchers that use this data is also significant, requiring proposal review, IT resources, and administrative resources to process Special Sworn Status applications for researchers.

The availability of a synthetic dataset would provide a tiered access option that would allow researchers to use microdata without the requirements for secure access. This option could prove invaluable to researchers in making maximum use of these data while preserving privacy. The availability of a synthetic dataset could prove invaluable to researchers in making maximum use of these data while preserving privacy. The production of synthetic data, though, can prove challenging and resource intensive.

This project will explore the following:

1. A user workshop would be conducted to identify variables of interest for a synthetic public use file. This workshop will target both current data users given their experience and knowledge of the data as well as potential data users who could have interest in synthetic publicly available data. A report would be produced from this workshop that highlights user feedback and a potential initial list of variables or variable groupings that are of interest to ABS users.
2. Two methods of producing ABS synthetic data would be utilized to produce two separate synthetic ABS microdata datasets. One method, using proprietary CenSyn software, has been partially implemented and would require completion. The second method would use R packages to create a second synthetic version of the ABS public-use microdata file.
3. Once produced, the two datasets and the process for creating them would be compared to determine any differences in data quality and validity between the two products and the cost and ease of production given the two methods.
4. The Census Bureau Disclosure Review Board (DRB) in conjunction with the Internal Revenue Service, Statistics of Income Division (SOI), would review the chosen synthetic file for disclosure concerns. Any publicly released dataset would require both Census Bureau and IRS-SOI approval prior to release which may require multiple iterations with some risk of rejection.
5. The public-use synthetic dataset(s) would be utilized for an evidence-building project in the NCSES-managed compute environment, with the use of verification metrics to test this method of verifying the quality of estimates produced from the synthetic data. Use of the data in this environment will assist in testing a new secure compute environment. One example of a verification metric is a flag indicating whether the synthetic estimate fell within the confidence interval of the true data. Additional verification metrics may be explored. One option to produce the metrics, would be to send a query to Census to run on the restricted-use data. The metrics would undergo Census Bureau disclosure review and once cleared, be transferred to the NCSES-managed compute environment to inform the analysis based on the synthetic data.

Information Gaps

This project will identify:

- What methods can be used to create synthetic datasets, focusing on ease of creation, data quality, and cost.
- How verification metrics can be used to inform the quality of synthetic data in evidence-building.
- A path to produce verification metrics for an agency's data and transfer those metrics to a centralized shared server.

- Use cases for a synthetic version of a high-value economic dataset.
- A potential path for collaborative opportunities for federal agencies in creation and assessment of synthetic datasets.

Key Evidence Building Considerations

- Key focus questions (address one or more) to assess innovation in the following areas: data acquisition, data security, data linking, privacy, and engagement:
  - What are the lessons learned to using the presumption of accessibility in the Evidence Act?
  - Can synthetic data be used to support a tiered access model?
  - Which novel techniques for data, privacy, and confidentiality protections can be used?
  - Are the resulting data and models fit to inform policy discussions and to make data available more equitability?
  - What types of collaboration and stakeholder engagement are needed to help inform these questions, data, and analysis?
  - What mechanisms are needed to access the resulting data that uphold privacy requirements?