Disclosure Avoidance and the American Community Survey

Rolando A. Rodríguez, U.S. Census Bureau
Presented to the COPAFS Quarterly Meeting
September 10, 2021

This presentation is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S Census Bureau.
Balancing data release under Title 13

- Census Bureau surveys have competing requirements
  - Release statistics
  - Protect privacy
- Quality of ACS releases are supported by numerous metrics
- Privacy of ACS releases is not currently quantified
Modern privacy theory makes clear that accuracy and privacy form a tradeoff.

Legacy privacy methods could not quantify privacy loss in a meaningful way.

Formal privacy methods can budget for privacy.

Survey designers and stewards are not accustomed to “budgeting” for privacy.
Privacy and sampling: It’s complicated

• Sampling has been called a privacy technique
• But those in the sample have an increased effect upon survey outputs
• Adjustments to weights can increase this individual impact
• Surveys also tend to collect more sensitive data than censuses
• Interaction between formal privacy methods and surveys is an active research area†‡

†https://www.census.gov/newsroom/blogs/research-matters/2020/09/three_cooperativeag.html
‡https://www.census.gov/newsroom/blogs/research-matters/2016/10/census_bureau_awards.html
When privacy needs lead privacy science

• Re-identification of respondent information is a Census Bureau issue†
• Respondents deserve better protection across products now
• Other paradigms, such as synthetic data, can help fill the gap between swapping and formal privacy

†https://www.census.gov/data/academy/webinars/2021/disclosure-avoidance-series/simulated-reconstruction-abetted-re-identification-attack-on-the-2010-census.html
Making public microdata synthetically

• Synthetic data methods are well-known to stats community
  • In use for certain surveys (SIPP†, ACS)
  • Software packages already available (e.g., synthpop in R)
• ACS will research a new fully-synthetic data product
• Data will be based on an Internal Reference File (IRF), with no other privacy methods applied

†https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html
Trust but verify

• Making synthetic data to satisfy all use cases is impossible
• Current research uses classification trees to create synthetic data
• Results of trees are assessed on a generalized marginal metric
• Users will be able to validate synthetic output against internal data
ACS Privacy Modernization Timeline

**Synthetic IRF and Validation**
- **2020**: Develop
- **2021**: Internal Review
- **2022**: Internal Test
- **2023**: External Test
- **2024**: Release
- **2025**:

**Formal Privacy**
- Research
- Outreach
- Collaboration
Data users will have numerous routes to obtain the ACS results they want and need.
Contact us

acsprivacy@census.gov