Preparing Data Users for Differential Privacy in the 2020 Census

FESAC December 14, 2018

Erica L. Groshen, Cornell University—ILR School

Cornell University
ILR School

Agenda

- Recap of motivation
- Census 2020 disclosure plan
- How to prepare?
  - Census
  - Users
  - Other stats agencies

2020 Census is key data infrastructure

- Policy
- Research
- Private sector, including users of other “big data”
- Next decade and after

“Sometimes, the riskiest thing to do is nothing.”

Timothy Geithner, President of NY Fed, mid-2007

(as remembered by me)

Cornell University
ILR School

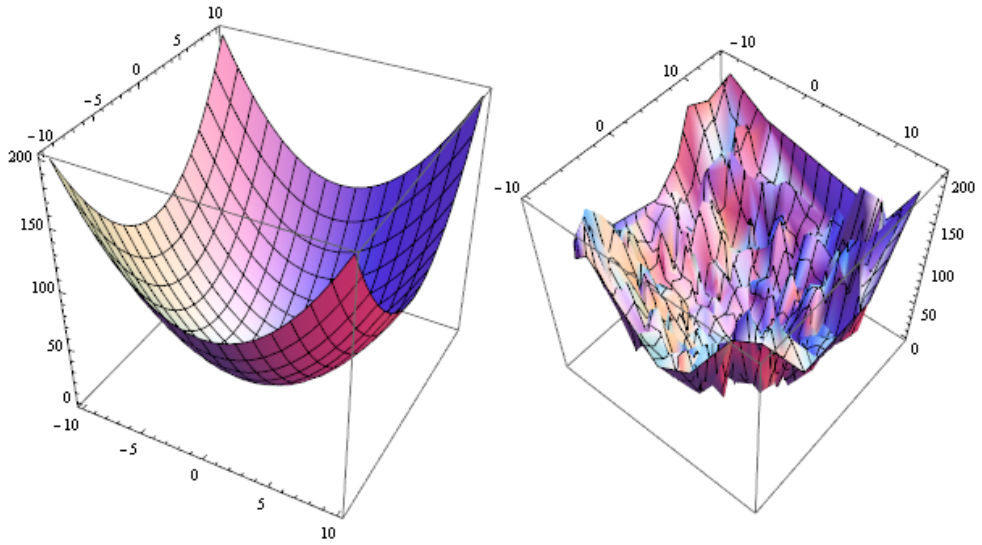
Motivation for confidentiality protection

- Unchanged mission
  - Statutory (Title 13, CIPSEA and predecessors)
  - Data quality (respondent trust and cooperation)
  - User trust
- Proliferating
  - External micro data
  - AI and computational capacity
  - Variety of uses, users and products
  - Demands for transparency and assurance of protection

2020 Census disclosure plan

2020 Census disclosure plan

- Top-down, centralized Differential Privacy budget
- Aggregate tables for redistricting to look same as before
  - Highly aggregated statistics and “invariants” unaffected
  - As population shrinks in unit of aggregation, effect of infused noise more evident
  - “Adding up” preserved
  - Properties of added noise public, except actual “key”



2020 Census disclosure plan, continued

- Accuracy of other tabulations
  - Informed by use-cases provided in response to 2 Census FRNs
  - Importance of needs used to allocate fixed privacy loss budget
- Public use file (iPUMS): no decision yet
  - Micro-data file that's input to official tabulations (MDF) can be released with no additional privacy loss
  - Fully reproduces tabulations exactly for all published detail
- FSRDC projects using 2020 Census confidential files
  - Capacity won't expand much beyond current plans (+3 new sites/year)
  - Proposals and released products restricted, subject to as-yet-undetermined explicit privacy loss budgets
- Implementation details will evolve…

Any upside for data users?

- More transparency, scientific honesty
- Data quality
- Opportunity for input
- Time to prepare
- Risk management

Risks needing management

- Damage to respondents who are identified
- Hackers sell reconstructed/identified microfiles
- Suboptimal/reactionary external requirements
  - Congress
  - Commerce Dept./White House
  - Courts
- People fear answering Census
  - Less accurate responses
  - Lower overall or item response rates

Preparation

“By failing to prepare, you are preparing to fail.”

Benjamin Franklin

Preparation: Census

- Prioritize uses for privacy budget
  - Top: Redistricting
  - Next: Allocating funds and uses noted in Federal Register Notices
- Provide public-use MDF file as iPUMS
  - Release exact micro-data input for official tabulations
  - **Implement replication process to check MDF results' consistency with findings using confidential data**
    - **Build confidence (hopefully!) and transparency**
    - **Inform future applications of DP**
    - **Inform methodologies that use MDF**

Preparation: Census , continued

- Inform and educate users
  - Predict impacts for products and users
  - Advise of updates and changes
- Work with statisticians, econometricians, programmers, tech firms to adapt methodologies
- Develop communication materials
  - For respondents, field staff, journalists, policymakers
  - About privacy protection and rationale for “fuzzing” information products

Preparation: Users

- Researchers
- Methodologists
- Tech companies
- Analytical journalists
- Policy makers
- Policy analysts and advocates

Preparation: Users, continued

- Communicate needs, concerns, priorities to Census
  - Develop relationships, information channels
  - Respond to Federal Register Notices, requests for input
- Plan for changes
  - Estimators, standard errors, etc.
  - Alternative access and/or products
- Monitor for updates and changes in plans
- Sponsor or attend educational events for colleagues and students
- Promote participation in the Census

Preparation: Other statistics agencies

- Assess implications for products that rely on Census data
  - Sample frames
  - Aggregate statistics
  - Research programs
- Assess current disclosure limitation processes in current context
  - If or when to adopt DP?

Conclusion: Work ahead for many stakeholders

- Gold-standard 2020 Census is key infrastructure for evidence-based decisions
- Trust in stats agencies is mission-critical
  - Current context makes accuracy and privacy more costly
  - Differential Privacy aims to minimize losses, but cannot eliminate consequences
- Census and stakeholders must all prepare
  - Prioritizing, planning, communicating, education, methodology…
- Uncertainties remain

Erica L. Groshen
Visiting Senior Scholar, Cornell-ILR
erica.groshen@gmail.com

Cornell University
ILR School

Federal Register questions

1. How are the data from each individual table and data product used? Include any specific legal, statutory, or programmatic uses. Please cite any supporting federal laws or regulations.
2. Why are decennial census statistics used for this purpose? Please provide a clear justification.
3. Without decennial census data, how would this activity be accomplished (*e.g.,* other data sources)?
4. Who are the users of the specific table or data product?
5. Who is affected by the use of the data in this specific table or data product?
6. How much funding is distributed based on these data?
7. What is the lowest level of geography (*e.g.,* county, census block, etc.) at which data need to be published for each specific table? Please explain why data are needed at this level of geography.
8. In what additional levels of geography (*e.g.,* county subdivision, school district, etc.) or geographic components (*e.g.,* urban, rural, etc.) do data need to be published for each specific table?
9. What programmatic, statutory, or legal uses are there for decennial census data that are not being met by the current suite of decennial census products?