2020 Census:
Enhanced Disclosure Avoidance

Ken Hodges

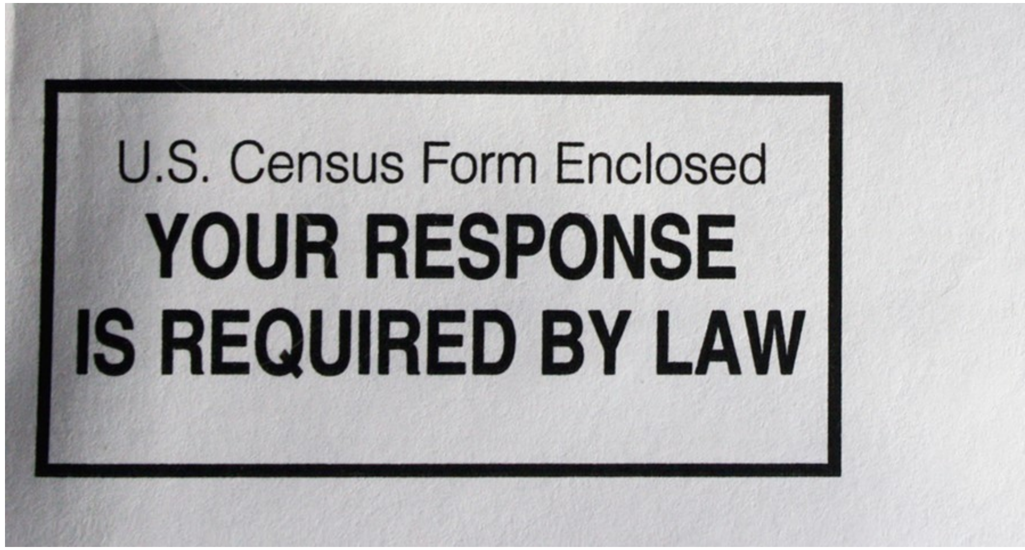
claritas

United States Census 2020

- Will ask us a few questions
- Data users ask Census . . . What data will we get?

Census Confidentiality Pledge

- The form says . . .



U.S. Census Form Enclosed

YOUR RESPONSE IS REQUIRED BY LAW

- The law that requires our response
- Also guarantees our response is confidential

Census Confidentiality Pledge

The Census publishes

- Data summarized for geographic areas
  - Counties, census tracts, block groups, blocks
- Microdata samples (PUMS)
  - Anonymous person and HH records
  - <u>Only</u> a sample of population
  - <u>Only</u> large areas (100,000+ population)
  - To protect confidentiality
- Businesses use more small area data

claritas

Protecting Confidentiality

- Census Bureau takes confidentiality pledge seriously
- Applies “disclosure avoidance”
- “Suppression” – One way to do this
  - Withhold data if numbers too small
  - Problem
    - Missing data
    - Blank cells
    - Small area totals do not sum to large area totals

.

Protecting Confidentiality

- CURRENT METHOD: “swapping”
  - Some HHs moved from one block to another
  - Don’t know which ones, or how many
  - Element of mystery
- Might think you identified a specific person
  - But can’t know. Was it swapped?
- Some impact on accuracy
  - But no blank cells, No missing data
- Users are accustomed to swapping

Protecting Confidentiality

- NOW: increased risk of disclosure
  - Powerful computers
  - Sophisticated math
- Can to take small area . . . (such as for tracts)
  - And reconstruct the HH/person records they are based on
  - Ability to identify individuals

.

Protecting Confidentiality

- Census Bureau has concluded:
  - Current disclosure avoidance methods are inadequate
- Publication of data for small areas
  - Like 2010 Census products
  - Constitutes a breach of confidentiality
- We are told:
  - There is no point in arguing about this

.

Differential Privacy

Therefore . . .

- For 2020: Census Bureau will apply “Differential Privacy”
- Applies “noise infusion”
  - Modifies tabulated statistics
  - Protects confidentiality
- Addresses the tradeoff between privacy and accuracy
  - The more privacy protection we need, the more accuracy we give up

.

Differential Privacy

The case for Differential Privacy . . .

- Swapping was secretive
  - Not transparent
  - Impact on accuracy not known
- Differential Privacy measures impact on accuracy and privacy
  - Can address the question . . .
    - How much protection do we need?
    - How much accuracy are we willing to give up?
- Census Bureau seeking input from data users
  - Optimal balance between accuracy and privacy?
  - One-size-fits-all?

.

Business User Concerns

- Let’s agree:
  - Risk of disclosure has increased
  - Person data can be reconstructed from aggregations
- Let’s accept:
  - Differential Privacy is effective
  - Attractive and impressive features
- So what’s the big deal?
  - Concerns are less technical than practical
  - Census Bureau focus seems more technical than practical

.

Business User Concerns

First:

- Disclosure risk greater for smaller areas
  - More noise infused in small area data.
  - Less in larger areas
- Additive consistency lost (small areas don’t sum to large areas)
  - Tracts summed to county . . . One number
  - Published county data . . . A different number
- Swapping did not have this limitation
  - Some records moved, but everything still there
  - Additive consistency preserved
  - Practical matter: Data function as if not modified

Business User Concerns

- Why is this important?
  - Businesses aggregate to custom areas
  - Improves accuracy
  - With swapping . . .
    - For aggregate areas, fewer household swapped
    - Less impact on data
- Differential Privacy
  - Will not have this self-correcting feature
  - One level of infusion for tracts, another level for county
  - Don’t reduce “noise” by aggregating tracts
  - Go to next higher geographic level to get less noise infusion

Business User Concerns

- To those who advocate swapping . . .
  - Users assume it provides accurate data
  - Amount of error is unknown
  - Swapping is secretive -- not transparent
- My view:
  - We know swapping involves error
  - Not bothered that amount of error is unknown
    - Error reduced with aggregation
  - Swapped data function as if not modified
    - Swapping not transparent, but transparent to users
    - Differential Privacy not transparent to users

Business User Concerns

- Don’t worry too much about additive consistency
- Most businesses don’t use published census data
  - Use data from private suppliers
  - Estimates built from census data
- Suppliers will deal with additive inconsistency
  - Produce estimates that sum to larger areas
  - Accuracy will improve with aggregation
- However . . .
  - Inconsistency will remain in published 2020 data

Business User Concerns

Second:

- Disclosure risk increases as more tables published
  - More clues for reconstructing individual records
- Census Bureau will publish less data from 2020
  - Some tables eliminated
  - Others only for higher level geography
- To some extent: A return to data suppression
  - Differential Privacy – AND -- Suppression
- This is the bigger concern
  - More than additive consistency
  - More than accuracy measures

Business User Concerns

- Which tables are most likely to be cut?
- How many might be impacted?
  - Don’t know
- When will we find out
  - Nine months to a year
- Census Bureau seeking input from users
  - Which tables do we need most?
  - Which tables can we do without?
  - Good luck getting data users to agree

Business User Concerns

- Census Bureau sought input last year
- Massive spreadsheet
  - All 2010 census tables listed (334 from SF1 alone)
  - For each table, Users asked to report:
    - What data are used for
    - Lowest geographic level
    - All geographic levels used
    - If census did not provide, what would you do?

Business User Concerns

- Claritas responded from perspective of our clients
  - But we are not end user
  - Census Bureau needs to hear from end users
- Still much to be decided
- All users need to . . .
  - Stay tuned
  - Be engaged
- Mere squawking won’t save out data
- Let’s work <u>with</u> the Census Bureau on this
- Let them know what you need

Thank You

ken.hodges@claritas.com

claritas

Looking Long Term

Privacy discussion happening in interesting context

- Foundations of Evidence-Based Policymaking Act
  - Follow up to Commission on Evidence-Based Policymaking
  - Issued recommendations in 2017
- Ambitious and seemingly paradoxical goal
  - Improve access to federal data
  - Improve confidentiality protections
- National Secure Data Service
- Still in very early stages

Looking Long Term

- OMB developing a federal data strategy
  - Something it has lacked
- Regular updates from Chief Statistician of US
- Data strategy still a work in progress
  - But emphasis on data sharing and access
  - Identifying datasets of value to the private sector
  - How to make data available to private sector users

- Long Term: Encouraging words about access to data
- Short Term: Concern about losing what we have now