Risk assessment procedures for the 2020 U.S. Census

Steven Ruggles
Lara Cleveland
David Van Riper

Institute for Social Research and Data Innovation
University of Minnesota

UNECE/Eurostat Expert Meeting on Statistical Data Confidentiality (1 December 2021)
Problem statement

• US Census Bureau justified its transition to a formally private disclosure avoidance system based on the results of a database reconstruction and re-identification experiment.
Problem statement

• US Census Bureau justified its transition to a formally private disclosure avoidance system based on the results of a database reconstruction and re-identification experiment.

• How does the Bureau’s results compare with a similar (reconstruction) experiment based on chance?
Reconstruction Results

• 46% of Bureau’s reconstructed cases match someone in confidential data on census block, age, sex, race, Hispanic
• 41% of our random individuals match someone in random block on age, sex, race, Hispanic ethnicity
Gurobi

46% of reconstructed cases match someone on the same block in the original microdata on age, sex, race and whether Hispanic.
How many matches would be expected by chance?

1. National age/sex distribution & block size distribution
2. Random number generator
3. Blocks populated with random age/sex combos
4. 10,000 random individuals

53% of random individuals match someone in random block on age & sex.
Adding Race and Ethnicity

53% of random individuals match someone in random block on age & sex

Assign modal race and ethnicity to all persons in each block

41% of random individuals match someone in random block on age, sex, race and whether Hispanic
Database reconstruction experiment

• The Census Bureau “Reconstruction” matched on block, age, sex, and race/ethnicity in 46% of cases

• A random number generator combined with a simple assignment rule for race and ethnicity yields a comparable match rate of 41%

• Despite the Census Bureau’s massive investment of resources and computing power, their database reconstruction does not perform much better than a roll of the dice

• This is analogous to a clinical trial in which the treatment and the placebo produce virtually the same outcome.
Re-identification Results

• 45% of Bureau’s reconstructed cases match someone in commercial data on census block, age, and sex
• 38% of names harvested from commercial database match a name on same block in confidential data
• 17% (38% of 45%) of Census respondents “re-identified”
45% of reconstructed cases match someone in commercial data on Age, sex & block (race/ethnicity not used)

(One would expect 53% by chance)
The re-identification experiment

• Within each block the Census Bureau searched a commercial database for people who matched the age and sex of each case of their reconstructed database.

• In just 45% of cases was there at least one person in the commercial data who matched the age, sex and block number of a row of the hypothetical database.

• This is a lower match rate than the 53% one would expect by chance, probably because of the poor quality of the commercial data.
45% of reconstructed cases match someone in commercial data on Age, sex & block (race/ethnicity not used)

38% of names harvested from commercial database match a name on the same block in original microdata

Census claims 17% “reidentified” (38% of 45%)

(One would expect 53% by chance)
The re-identification experiment

- The Census Bureau then harvested the names of the 45% of cases in the commercial database that matched on age and sex and searched for them in the original census data.
- They found that 38% of the 45% were actually present on the block.
- They claimed to have “reidentified” 16.85% of the population (38% of 45%).
Census Bureau re-identification experiment

Among the 45% of cases where the reconstructed data match someone in the commercial data on age, sex, and block, 38% match a name on the same block in original microdata.

- There is no null comparison
- One would expect that people residing on a particular block in the commercial data would also be enumerated in the census.
- Is 38% high or low?
Evaluating the re-identification experiment

• The Census Bureau could easily evaluate the re-identification by matching the names of people randomly selected from the commercial database to persons in the 2010 census living on the same census block, without any reference to the Census Bureau’s database reconstruction.

• If the 38% match rate on names for the reconstructed population is no higher than the match rate for a randomly selected subset of the commercial data, it would mean the database reconstruction had no effect on re-identification risk.

• Because they made no null comparison, the re-identification results cannot be interpreted.
Summary: Database reconstruction experiment

• “Correctly” guessed age, sex, race and Hispanic ethnicity for 46% of persons on each block
• Great majority of those correct guesses would be expected by chance
• An outside attacker would have no means of determining which “reconstructed” records were true
Summary: Re-identification experiment

- The re-identification experiment shows only that some people found in a commercial database are also listed on the same block in the census.

- It fails to show that using the published tabulations had any impact whatsoever on the match rate between the commercial database and the original enumeration.
Consequences of disclosure panic

The disclosure avoidance system introduces unacceptable levels of error for many applications of the census.
The use of differential privacy for census data and its impact on redistricting: The case of the 2020 U.S. Census

Christopher T. Kenny, Shiho Kurikawa, Cory McCartan, Evan T. R. Rosenman, Tyler Simko, Kosuke Imai

Census statistics play a key role in public policy, revealing individual information, many of which add noise to census statistics must be processed. The Census Bureau’s latest disclosure avoidance strategy, called the Differential Analysis System (DAS), is based on the idea of adding noise to the data to protect the privacy of individuals. This approach is designed to prevent disclosure of sensitive information while still providing useful statistics.

INTRODUCTION

In preparation for the official release of U.S. Census Bureau data, a new system (DAS) has been developed to protect individual responses to the 2020 Census. The DAS is based on differential privacy, which adds noise to the data to protect the privacy of individuals. This approach is designed to prevent disclosure of sensitive information while still providing useful statistics.

Differential Privacy and the Accuracy of County-Level Net Migration Estimates

Richelle L. Winkler, Jaclyn David Egan-Robertson

Received: 5 March 2021 / Accepted: © The Author(s) 2021

Abstract

Each decade since the 1950s, census data has been used to estimate migration estimates by age, sex, and race at the county level. However, these estimates are susceptible to noise and disclosure avoidance. The 2020 Census will allow for more accurate estimates by employing a new system (DAS) that provides differential privacy for the data.

Differential Privacy in the 2020 Census Will Distort COVID-19 Rates

Mathew E. Hauer

Abstract

Scholars rely on accurate COVID-19 pandemic, with deaths at older ages. Populations subject to noise inferences on differential privacy. Use of differential privacy can introduce substantial distortions, reducing ability to detect very small differences.

How differential privacy will affect our understanding of health disparities in the United States

Alexis R. Santos-Lozada, Jeffrey T. Howard, and Ashton M. Verduzco

Abstract

The application of a currently proposed differential privacy algorithm to the 2020 United States Census data and additional data products may affect the usefulness of these data, the accuracy of estimates and rates derived from them, and critical knowledge about social phenomena such as health disparities. More is known about the impact of DP on important uses of census data, particularly in light of concerns discussed at the workshop on December 2019.

Why are there concerns about privacy? Title 13 of the United States Code imposes heavy obligations on the U.S. Census Bureau...
Findings of recent publications include:

- Systematically undercounts the population in mixed-race and mixed-partisan precincts.
- Introduces substantial distortion in COVID-19 mortality rates, sometimes causing mortality rates to exceed 100 percent.
- Net migration estimates by five-year age groups would only be accurate enough for use in about half of counties.
Conclusions

• The Census Bureau is now in the midst of a radical escalation of disclosure control justified by an unsuccessful database reconstruction experiment.

• The actual threat posed by the Census Bureau’s reconstruction is similar to the threat posed by randomly guessing respondent characteristics.
Conclusions

• The database reconstruction experiment failed to demonstrate a disclosure threat and cannot justify degradation of the nation’s statistical infrastructure.

• Reducing data quality is likely to backfire, as it will erode public trust and weaken the core justification for census data collection.
Thank you.

Questions?