

U.S. Census Confidentiality: Perception and Reality

Margo Anderson

*University of Wisconsin-Milwaukee, Department of History
Milwaukee, WI 53201, USA
margo@csd.uwm.edu*

Stephen E. Fienberg

*Carnegie Mellon University, Department of Statistics
Pittsburgh, PA 15213-3890, USA
fienberg@stat.cmu.edu*

1. Introduction

The 2000 decennial census of population and housing was perhaps the most controversial in U.S. history largely because of the longstanding political controversy over the sampling and adjustment (Anderson and Fienberg, 2001). Despite a growing U.S. concern over privacy during the 1990s, there was little discussion of the issue as the Bureau sought and received congressional approval for the specific questions to be included in census 2000. Nonetheless, as the mail-out mail-back phase of the census was in progress in the early spring of 2000, public outcries over the “intrusiveness” of the census long form (with 53 questions and sent to a sample of 1 household in 6) threatened to undercut the success of the Bureau’s campaign to increase the response rates over those from 1990. (see Anderson and Fienberg, 2001, Appendix J, and Robbin, 2001). The gap between the mail-out mail-back return rates for the short and long forms grew from under 6% in 1990 to over 12% in 2000. This decline in cooperation was paralleled by survey results showing that privacy concerns regarding the long form had increased from 10% at the beginning of the mail-out mail-back process to over 20% towards the end (Nie and Junn, 2000). Following the basic enumeration, the Census Bureau conducted the Accuracy and Coverage Evaluation Survey of over 300,000 households in 10,000 blocks across the country to produce sample-adjusted counts for all purposes other than Congressional apportionment.

Here, we review the Bureau’s approach to disclosure limitation for public use microdata files and the tables accessible online from its newly developed “American FactFinder” system. Then we summarize some threats to census confidentiality, perceived and real, articulated by the Bureau. We conclude with some observations on additional sources of protection against disclosure associated with census error and the possible uses of sample-based adjusted data.

2. The American FactFinder System

In 1963, the Bureau first released Public Use Microdata Samples (PUMS) containing individual long form records with unique identifiers removed to protect respondent confidentiality.

In Spring 2000, it announced plans to reduce the level of individual detail for the 2000 PUMS files, in part to assuage public concerns about confidentiality. The Bureau argued that advances in computer technology, data mining tools, and expanded access to census data demanded a higher level of confidentiality protection. While opposition from the professional users of PUMS data forced the Bureau to retreat from its initial proposals, it still refused to go back to the level of detail used in 1990 (see Robbin, 2001, and <http://ipums.mpc.umn.edu>).

Changes to data releases went well beyond the PUMS. Beginning in December 2000 with the state population totals for reapportionment and in March 2001 with the detailed population totals (to the block level) for redistricting, the Bureau shifted the release of its full range of Census 2000 tabulations to a new Internet data-delivery system known as “American FactFinder” (AFF) (see <http://factfinder.census.gov>). Halawa (2001) describes the different levels of data available from AFF: (a) *Tier1 Data* required for release by law; (b) *Tier2 Data* consisting of Bureau-defined tables approved by its Disclosure Review Board; (c) *Tier3 Data* which allows for user defined tables but only if they pass disclosure limitation rules. Both Tier2 and Tier3 releases are based on “microdata” subjected to the following disclosure limitation techniques:

- *Data Recoding*. Categorical variables such as detailed race, occupation, industry, Hispanic origin, group quarters are recoded into new variables that show less detail. Continuous variables, such as household/family income, individual income types, cost of electricity, property tax, mortgage payments, gross rent, are top-coded or bottom-coding.
- *Data Swapping*. Pairs of household records (either long form or short form) in different census blocks are match on a set of demographic characteristics and swapped. Records are selected for swapping with probability inversely proportional to block size, and those with unique race categories within blocks have an increased probability (for statistical details, see Fienberg, Steele, and Makov, 1996).
- *Query Filters*. Tier3 requests for cross-tabulations for specific geographic areas must pass a series of additional checks, including minimum population sizes for both geography level and data splits and the use of Bureau-defined cell measures. At most three variable are allowed in a cross-tabulation for a given geographic area. All requests that pass these checks go through a final *results filter* that reapplies the Bureau’s disclosure limitation rules to subtables, and imposes additional controls on recodes and table cell contents.

Sampling in a census context provides the additional uncertainty needed to protect many data releases when combined with these devices.

3. Threats to Census Confidentiality?

Among the new issues and perceived threats to the confidentiality of census 2000 data considered by the Bureau’s Disclosure Review Board were the following (e.g., see Gates, 2001):

- *Finding oneself*—Despite the use of disclosure limitation methods like swapping, respondents can more easily identify themselves because they know that they are in the database. Will those who identify themselves erroneously assume that others can also do so?

- *Cells with single observations*—Data swapping techniques still yield tables where some cells contain a single individual or household. Will respondents who don't understand the implications of swapping assume that such “uniques” reveal confidential information?
- *Multiple race reporting*—For the first time in 2000, the Bureau permitted people to report multiple races (5 categories plus *other* yielding 63 possible combinations compared with only 6 in 1990). Are swapping and rounding procedures sufficient to deal with the sparseness for some combinations at low levels of geography? How will the public perceive the reporting of small counts for racial categories?
- *Expanded access via Internet*—How do expanded Internet access to census data, improved general computer skills, and the wide availability of datamining tools affect disclosure risk?
- *Non-Census-Bureau uses of census data*—Some data vendors assign characteristics to people in non-census files based on geographic summaries (e.g., average income assigned to groups living in a certain ZIP code). How should the Bureau deal with concerns that these uses suggest that the individual census records were linked to external databases?
- *Public concerns about other data collectors*—Will concerns about private-sector privacy violations spill over in some way to the census?

4. Observations and Issues

The Census Bureau's confidentiality concerns relate to unknown but potentially unacceptable levels of risk and to how respondents might react to census content and disclosure policies. To date there have been no dramatic revelations of “intruders” identifying individuals in the data releases from past U.S. censuses although there have been claims that a small number of individuals have been able to identify themselves. While the demands for real-time online access to census data have grown dramatically in recent years, there is at best limited evidence that the risk of disclosure of census information has grown appreciably.

Many of the concerns raised deal not with disclosure harm but with the discredit harm that would accrue to the Census Bureau if someone were to claim to identify individuals in census releases, whether or not they had actually done so, or with perceptions of risk that do not fully reflect what the Bureau actually does to protect census data from disclosures.

Most assessments regarding disclosure risk assume that the data records for individuals and households have been measured without error. This is hardly the case for U.S. decennial census data. In 1990, the number of omissions plus the number of erroneous enumerations totaled approximately 25 million, or roughly 10% of the enumeration total. Early analyses of the results of the 2000 Accuracy and Coverage Evaluation (ACE) survey suggest a similar level of error in the 2000 census. Further, for households properly enumerated in the census many of the data elements are in error. Thus error in the census when combined with the regrettable non-response to the census long form adds substantial protection against disclosure risk.

Finally, we note that the Census Bureau was prevented from using sampling for non-response followup in the 2000 census as a result of a Supreme Court decision (see the discussion in Anderson and Fienberg, 2001), and the Bureau decided not to release adjusted counts at the block level (see <http://www.census.gov/dmd/www/EscapRep.html>). But had sampling been implemented for both of these components, and the results of the microdata files been adjusted accordingly, there would have been even greater protection from disclosure risk.

REFERENCES

Anderson, M. and Fienberg, S.E. (2001). *Who Counts? Census-Taking in Contemporary America*. Revised Paperback Edition. Russell Sage Foundation, New York.

Fienberg, S.E., Steele, R.J. and Makov, U.E. 1996). Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: Data swapping and loglinear models. *Proceedings of the U.S. Bureau of the Census Twelfth Annual Research Conference*, 87–105.

Gates, G. (2001). A holistic approach to confidentiality assurance in statistical data. Paper presented at *Joint ECE/Eurostat Working Session on Statistical Data Confidentiality*, Skopje, Macedonia, March 14-16, 2001.

Halawa, S. (2001). American FactFinder: U.S. Bureau of the Census works toward meeting the needs of users while protecting confidentiality. Paper presented at *Joint ECE/Eurostat Working Session on Statistical Data Confidentiality*, Skopje, Macedonia, March 14-16, 2001.

Nie, N.H. and Junn, J. (2000). America's experience with Census 2000: A preliminary report. http://www.intersurvey.com/about-intersurvey/press/0504200_census.html.

Robbin, A. (2001). Interpretations of privacy and confidentiality rules by government agencies. Paper presented at the Annual Meeting of the Population Association of America, March 29-31, 2001.

RESUME

The use of sampling for adjustment, the expanded multiple categories for racial groups, and the “intrusiveness” of the census long form were among the controversies surrounding the taking of the 2000 U.S. Census of Housing and Population. This paper discusses the privacy and confidentiality concerns considered by the Census Bureau and the methods it ultimately chose to limit disclosure risk. It also points out the links between the concerns for disclosure risk and the debate over sample-based adjustment of census counts.

Aux Etats-Unis, l'usage d'échantillon pour ajuster, l'évolution des multiples catégories de groupes raciaux, et l'indiscrétion du long processus de recensement étaient parmi les controverses animant les discussions sur le recensement américain de la population et de l'habitat. Ce papier développe les questions de confidentialité et d'intimité envisagées par le Bureau de Recensement et les méthodes qu'il a finalement choisies pour limiter le risque de divulgation. Il met également en évidence les liens entre les inquiétudes sur le risque de divulgation et le débat sur l'ajustement des échantillons de base pour le dénombrement du recensement.