Research And Methodology Directorate

A History of the American Housing Survey and Disclosure Avoidance

By Laura McKenna

Issued April 2019

United States™ Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

INTRODUCTION[1]

The U.S. Census Bureau conducts the American Housing Survey (AHS) for the Department of Housing and Urban Development (HUD) under Title 13, U.S. Code, Section 9 mandate to not "use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or make any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or permit anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual reports (13 U.S.C. § 9 (2007))." The Census Bureau applies Disclosure Avoidance (DA) techniques to its publicly released statistical products in order to protect the confidentiality of its respondents and their data. None of the information in this paper is confidential.

HISTORY OF THE AHS

Historical housing data can be found from the last 60 years of Census Bureau decennial censuses. Construction history tables are also available from the Census Bureau's Web site including building permits and housing starts, as well as vacancies and homeownership data. Patterns illustrate characteristics of neighborhoods and housing tenure. HUD uses Census Bureau data to create thematic maps of household vacancies, income, and building permits. Fannie Mae's research includes information on housing and mortgage industries, consumers, demographics, and the economic and mortgage markets, <www.census.gov/history/www/reference/publications/historic_housing_data.html>.

HUD wanted a specific database that would assess the quality of housing, housing characteristics, what Americans were paying for housing, and how things change over time. The Census Bureau worked with HUD to develop a survey to measure these things, and in 1973, the Census Bureau launched the first Annual Housing Survey, <www.huduser.gov/portal/sites/default/files/pdf/Introducing-the-2015-American-Housing-Survey-Slides-and-Notes.pdf>. It was a national survey, which added metropolitan surveys 1 year later. There were 60 metropolitan areas, with 21 of them being surveyed at 6-year intervals, seven each survey year, and each survey included 3,000 to 5,000 sample housing units. For many years, the national survey was conducted annually. Since 1983, it has been conducted every 2 years and the name was changed to the American Housing Survey, <https://en.wikipedia.org/wiki/American_Housing_Survey>.

There have been three different national samples, two from the 1970 Census and one from the 1980 Census. With each new survey, the sample is updated with additions from new construction or other sources. The survey has changed over time to include new variables such as mortgage information, lines of credit, and householder's country of birth. Race was changed to include mixed race categories, and questions on income have been redefined. The fact that housing units remain in sample over time makes this a longitudinal survey, but note that the people living in a given housing unit are not followed when they move. Only the housing unit itself is of interest.

DISCLOSURE AVOIDANCE FROM THE 1990S TO THE PRESENT

Tabular Data

Due to the small sample size with large weights, AHS tabular data is published for very large areas for data quality reasons. No additional DA techniques were deemed necessary to protect the data.

Microdata

Removal of Direct Identifiers

The Census Bureau removes direct identifiers from the file such as name, address, phone number, etc.

Geographic Threshold

All geographic areas identified must have a population of 100,000 or more. When calculating this population, all geography-related variables on the file are cross-tabulated to obtain the final population count of an area that can be identified as a piece of geography.

Topcoding and Bottom-Coding

The Census Bureau uses topcoding and bottom-coding to eliminate outliers in a file for continuous variables such as value of housing unit and mortgage amount. A topcode (cutoff) is in place for 0.5 percent of all values or 3 percent of all nonzero values, whichever is the larger of the two. Originally, all topcoded values were replaced with the topcode cutoff itself. Later in 1996, the topcoded values were replaced with the mean of the topcoded values. At least three values must be topcoded or the topcode is lowered to meet this requirement. Bottom codes are the same except on the other side of the distribution. A bottom code might be

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

applied to year built of the housing unit. For variables that are part of a sum, the individual summands are topcoded prior to their summation.

Rounding/Recoding

Each category of a categorical variable must contain at least 10,000 weighted people or households (depending on the universe of the variable) for that particular variable nationwide. If a category does not meet this threshold, it must be combined with other categories until it does.

Dollar amounts must follow one of two rounding/recoding schemes.

Round to two significant digits, or use this recoding scheme:

- Zero rounds to zero.
- 1 to 7 rounds to 4.
- 8 to 999 rounds to the nearest multiple of 10.
- 1,000 to 49,999 rounds to the nearest multiple of 100.
- 50,000 and greater rounds to the nearest multiple of 1,000.

Any totals or other derivations are calculated using the rounded numbers.

Noise Infusion

Noise infusion is used to hide very unusual characteristics of a person or household at a given point in time. For example, consider a woman with sextuplets or a 10-year-old in college or a household with 13 people in it. Such unusual circumstances are often well known and sometimes in the news. Census Bureau editing procedures capture and alter many, but not all, of these types of unusual circumstances.

In addition, AHS is longitudinal and changes in personal or household characteristics can often be found in public records. For example, a birth, death, marriage, divorce, change in occupants, or sale of a housing unit would be reflected in the AHS, while a given housing unit is in sample.

The Census Bureau does not publicly release the details of how noise is added to protect these types of data that pose a disclosure risk.

A REIDENTIFICATION STUDY

In 2013, an external repackager of census data brought to the Census Bureau's attention that he reidentified one housing unit record in the New York City Housing and Vacancy Survey (NYCHVS) Public Use File (PUF). The Census Bureau's Center for Disclosure Avoidance Research examined the suspected reidentification, but the Census Bureau did not inform the repackager whether it was correct. The NYCHVS has many of the same variables as the AHS, but the AHS is a much larger survey. This incident led to a reidentification study on the AHS, using CoreLogic data as an attacker file. After DA techniques are employed, it can be useful to conduct a motivated intruder reidentification study to assess the disclosure risk of microdata and tabular data products before they are made publicly available. For microdata, such reidentification studies are performed by looking for unique combinations of variables in the microdata that are thought to be identifying, looking for externally available datasets that contain the same variables, and then linking data records in the two datasets using the linkage variables. Finally, it is necessary to verify the proposed matches by comparing the suppressed identities in the microdata with the identities in the external data set to see if the matches are true matches or false matches. This last comparison step is vital because, often, survey records are unique within the sample but not in the population. Census Bureau staff members conducting this study (Aref Dajani, Tamara Cole, and their staffs) worked with Dr. Shawn Bucholtz of HUD. The group had to do a great deal of work to get to the point where they could try linking an AHS file with CoreLogic data due to differences in definitions and categories of variables.

Once both data sets were ready, linkage attempts were made using three different metrics. The three metrics were called unicity, taxicab (L1 norm), and Euclidean (L2 norm), with a focus on the taxicab metric. The unicity metric bins continuous variables and matches records in the attacker and defender file if they have frequency '1' with respect to any cross-tabulation of variables. The other metrics bin variables and use the similarities in binned values to create a distance between any attacker and defender record, which is then compared against a cutoff to

determine whether any suspected linkages exist. The researchers were interested in three different percentages: the percentage of records that were suspected linkages, the percentage of records that were confirmed linkages, and the conditional rate, which is the percentage of suspected linkages that were confirmed. Since sometimes a single record in the AHS would match to many records on the attacker dataset, research focused on records in the defender dataset that matched to at most five records in the attacker dataset.

The reidentification study was very helpful to the Census Bureau. Both the correct linkage rate and the ratio of correct to suspected rate were very small. An in-depth analysis of the results helped the Census Bureau to determine which variables could cause correct reidentifications. As a result, a few variables were recoded or removed from AHS PUFs. The analysis of the study's results is confidential to the Census Bureau.

THE FUTURE

Recently, the Census Bureau has embarked on an aggressive effort to replace its legacy DA methods with modern DA techniques based on formal privacy methods, <https://privacytools.seas.harvard.edu/formal-privacy-models-and-title-13>. Current methods will gradually change with the introduction of formal privacy (Nissim et al., 2018). Most of the current Census Bureau's DA research is focused on formal privacy for all types of data (Nissim et al., 2007).

An algorithm operating on a private database of records satisfies formal privacy if its outputs are insensitive to the presence or absence of any single record in the input (Dwork, 2006). The Census Bureau's Disclosure Review Board[2] is quickly learning about formal privacy and how it protects Census Bureau data products.

REFERENCES

C. Dwork, "Differential Privacy," International Colloquium on Automata, Languages, and Programming (ICALP), 2006, pp. 1–12.

K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth Sensitivity and Sampling in Private Data Analysis," Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing, 2007, pp. 75–84.

K. Nissim, T. Steinke, A. Wood, M. Altman, A. Bembenek, M. Bun, M. Gaboardi, D. O'Brien, and S. Vadhan, "Differential Privacy: A Primer for a Non-technical Audience (Preliminary Version), Harvard University Privacy Tools for Sharing Research Data, 2018, <http://privacytools.seas.harvard.edu>.

[2] All Census Bureau data products must be approved before dissemination by the Disclosure Review Board.